The mutational landscape of metastatic cutaneous squamous cell carcinoma

Bruce Graham Ashford
University of Wollongong

Follow this and additional works at: https://ro.uow.edu.au/theses1

Recommended Citation

Research Online is the open access institutional repository for the University of Wollongong. For further information contact the UOW Library: research-pubs@uow.edu.au
The mutational landscape of metastatic cutaneous squamous cell carcinoma

Bruce Graham ASHFORD

Supervisors:
Professor Marie Ranson
Professor Jonathan Clark AM
Professor N. Gopalakrishna Iyer

This thesis is presented as part of the requirement for the conferral of the degree:
Doctor of Philosophy

The University of Wollongong
School of Chemistry and Molecular Bioscience

March 2019
Abstract

Cutaneous squamous cell carcinoma (cSCC) is a common skin malignancy. Rates of cSCC in Australia are the highest in the world, approaching 800 cases/100 000 people per annum in North Queensland. Between 2-5% of cSCC will metastasise to regional lymph nodes, representing a higher clinical stage and requiring more aggressive and more morbid treatment, principally surgery and radiotherapy. Little is known of the molecular mechanisms of metastasis in cSCC, which makes a stratified method of appropriate surveillance challenging.

Patients with biopsy proven metastatic cSCC to lymph nodes of the head and neck were identified and recruited for fresh tumour and whole blood harvest. DNA extracted from 19 nodal metastases of cSCC was sequenced using whole genome sequencing. Following quality control, in particular verification of tumour cellularity, this number was reduced to 15 for the purposes of the study. Tumour DNA was compared to DNA from whole blood (germline) to establish the pattern of somatic variation.

Approaching 98% of the somatic short variants observed were in the noncoding regions of the genome in all samples. A mutational burden (207/Mb) greater than any other malignancy previously described underpinned the mutational landscape, characterised by UV implicated C>T single nucleotide variation in known oncogenes and tumour suppressor genes, a highly amplified genome and significant structural variation, particularly involving TTC28.

Recurrent high impact short variants were seen in known cancer associated genes including TP53, CDKN2A and NOTCH1, but also in less well described but emerging genes of interest including MECOM, PTPRD, PLCB4, PCLO, CSMD3 and FAT4. Non-coding variants were particularly evident in the TERT promoter region identifying a variant pattern not previously found in cSCC. Significant amplification of microRNA miR-3147 was seen in all samples.
Prominent amplification of cancer associated long non-coding RNAs not previously identified in cSCC was observed. These included \textit{PVT1}, \textit{MALAT1}, \textit{HULC} and \textit{NORAD}.

Expression changes resulting from somatic mutation were explored using co-extracted RNA on the NanoString platform. Prominent overexpression of key genes in cancer progression and metastasis was observed such as \textit{NDRG1}, \textit{PIK3CA} and \textit{SOX2}.

This work has catalogued the extent and pattern of somatic variation in metastatic cSCC and has provided potential new targets that can now be investigated for their utility as biomarkers of progression and metastasis.
Acknowledgments

My Principal Supervisor, Professor Marie Ranson has believed in this project and me since I first came to her as a confused clinician. She had no reason to invest her precious time and effort, but she did. And her organisational and scientific prowess has overcome many of the deficiencies that I brought to the table. Without the support of a Cancer Scientist of her calibre, this idea and this project would never have progressed.

I have been incredibly fortunate to have found myself working with some of the great head & neck surgeons of our time. In particular, Professor Jonathan Clark has been a mentor and confidante across the full extent of my practice and these early ventures into research. This project and my involvement in it would simply not have occurred without his influence and guidance. Similarly, Professor N Gopalakrishna Iyer provided the methodological intelligence for this work in the research he had been undertaking in mucosal SCC. His knowledge in the field of cancer genomics in head and neck cancer has been a touchstone for many of us who are late to the field and slow to learn. Professor Ruta Gupta has been masterful in her patience and always willing to guide me to better understand this pathology and to ask the difficult questions.

I have been assisted by my practice staff who have collated patient information and consents, by my fellow surgeons in collecting specimens, by the nursing staff lead by Sr Jenny Spillane in helping curate and care for each precious tumour and by my Research Assistant Elahe Minaei and fellow PhD student Jay Perry in teaching me the secrets of working and surviving in a wet lab. It is not possible for me to overstate the assistance and advice given by the team of Cancer Bioinformaticians Dr Mark Cowley, Dr Vel Gayevskiy and Dr Maely Gauthier. They have forgotten more than I will ever know about bioinformatics and I am in their debt.

In particular, I wish to acknowledge the sophisticated figures in this thesis as largely the work of Dr Gauthier.
My family have always supported my work and have given me the time and space to think and worry. It has been a hectic 5 years that we will likely not forget quickly. Their love and advice do not go unnoticed and are always cherished.

To both the Illawarra Cancer Carers and Sydney Head & Neck Cancer Institute go thanks for supporting the funding required for a project of this type.

And finally to the patients who have willingly supported our research by literally donating themselves to the progress of this project. This work would not have been necessary but for them and would not have been possible without them. It is my hope that this work will allow us to more thoroughly understand this debilitating disease and to avoid the morbidity and death that it so regularly delivers.
Certification

I, Bruce Graham Ashford, declare that this thesis submitted in fulfilment of the requirements for the conferral of the degree Doctor of Philosophy, from the University of Wollongong, is wholly my own work unless otherwise referenced or acknowledged. This document has not been submitted for qualifications at any other academic institution.

Bruce Graham Ashford

23rd March 2019
List of Abbreviations

AJCC – American Joint Committee on Cancer
AK – Actinic keratosis
BAF – Beta allele frequency
BCC – Basal Cell Carcinoma
CLND – Complete/Completion Lymph Node Dissection
CNV – Copy Number Variation
cSCC – Cutaneous Squamous Cell Carcinoma
EMT – Epithelial Mesenchymal Transition
HREC – Human Research Ethics Committee
SCC – Squamous Cell Carcinoma (in head and neck cancer refers to mucosal disease unless otherwise stated)
SNB – Sentinel Node Biopsy
SNP – Single Nucleotide Polymorphism
SNV – Single Nucleotide Variation/Variant
SV – Structural Variation/Variant
VAF – Variant allele frequency
WES – Whole Exome Sequencing
WGS – Whole Genome Sequencing
Table of Contents

ABSTRACT ........................................................................................................................... I

ACKNOWLEDGMENTS.......................................................................................................... III

CERTIFICATION ................................................................................................................. V

LIST OF ABBREVIATIONS ................................................................................................... VI

TABLE OF CONTENTS ....................................................................................................... VII

TABLE OF FIGURES ........................................................................................................... XI

PAPERS INCLUDED IN THIS THESIS: ............................................................................... XIV

OTHER RELEVANT PUBLICATIONS DURING CANDIDATURE ........................................ XVI

1 INTRODUCTION ............................................................................................................... 1

1.1 CLINICAL CONTEXT ....................................................................................................... 1

1.2 AETIOLOGY OF cSCC .................................................................................................... 5

1.3 STAGING OF cSCC ....................................................................................................... 6

1.4 CONVENTIONAL METHODS FOR ASSESSMENT OF METASTATIC POTENTIAL OF cSCC .................................................................................................................... 7

1.5 GENOMIC OBSERVATIONS ........................................................................................... 15

1.5.1 TP53 family ............................................................................................................... 17

1.5.2 NOTCH family ......................................................................................................... 18

1.5.3 RAS Family ............................................................................................................. 19

1.5.4 CDKN2A ................................................................................................................. 19

1.5.5 Protein tyrosine phosphatase receptors .................................................................... 20

1.5.6 More recent genomic observations for cSCC ............................................................. 21

1.6 NON-CODING RNA – MICRORNA AND LncRNA ....................................................... 23

1.7 EPIGENETIC CHANGES IN cSCC .................................................................................. 25

1.8 STROMAL INFLUENCES AND EMT IN THE TUMOUR MICROENVIRONMENT ............. 26

1.9 DIFFERENTIAL EXPRESSION IN cSCC ......................................................................... 28

1.10 PROJECT DESIGN, RATIONALE AND AIMS ................................................................. 31
1.10.1 Aims of this project ........................................................................................................... 32

2 METHODS ................................................................................................................................. 33

2.1 PATIENT RECRUITMENT ......................................................................................................... 33

2.2 TISSUE HANDLING .................................................................................................................. 37

2.3 NUCLEIC ACID EXTRACTION ............................................................................................... 38

2.3.1 Germline (blood) nucleic acid extraction ............................................................................. 39

2.4 SEQUENCING .......................................................................................................................... 39

2.5 BIOINFORMATIC WORKFLOW ............................................................................................... 40

2.5.1 Sequencing outputs .............................................................................................................. 41

2.6 SCORING VARIANT EFFECTS ................................................................................................. 43

2.6.1 SIFT (Sorting Intolerant From Tolerant) .............................................................................. 43

2.6.2 PolyPhen ............................................................................................................................... 44

2.6.3 CADD (Combined Annotation Dependent Depletion) .......................................................... 44

2.7 GENE EXPRESSION ANALYSES ........................................................................................... 45

3 CLINICOPATHOLOGICAL CHARACTERISTICS AND BROAD GENOMIC CHARACTERISATION ............................... 46

3.1 TUMOR CELLULARITY AND CLINICOPATHOLOGICAL PARAMETERS .................................. 46

3.2 SEQUENCING COVERAGE ..................................................................................................... 49

3.3 MUTATIONAL BURDEN AND SIGNATURES ........................................................................... 49

3.4 DISCUSSION ............................................................................................................................. 53

4 MAJOR STRUCTURAL VARIATION INCLUDING COPY NUMBER VARIATION ................................. 56

4.1 OVERVIEW ............................................................................................................................... 56

4.2 MAJOR STRUCTURAL VARIATION .......................................................................................... 56

4.2.1 Results .................................................................................................................................. 56

4.2.2 Discussion ............................................................................................................................. 73

4.3 GENE FUSIONS ......................................................................................................................... 74

4.3.1 Results .................................................................................................................................. 74

4.3.2 Discussion ............................................................................................................................. 76

viii
Table of Figures

Figure 1-1 Synchronous lip cSCC with ipsilateral nodal metastases..............................2
Figure 1-2 Lymph node levels of the neck.................................................................3
Figure 1-3 Patterns of metastasis of primary cSCC.......................................................4
Figure 1-4 Kaplan-Meier survival analysis Sentinel node study.................................12
Figure 1-5 Expanded survival data for patients undergoing sentinel node biopsy.........15
Figure 2-1 Haematoxylin and eosin histopathological assessment of cellularity...........35
Table 2-1 Clinicopathologic data for all 19 samples ................................................36
Table 3-1 Sample cellularity variance......................................................................47
Table 3-2 Clinicopathologic data for 15 cases passed quality control with mutational burden estimates..................................................................................48
Table 3-3 Sequencing coverage for each sample tumour and blood(germline)..............49
Figure 3-1 Mutational burden of variety of malignancies............................................50
Figure 3-2 Mutational density and signatures ............................................................52
Figure 4-1 Circos plot sample 4699 ...........................................................................57
Figure 4-2 Circos plot sample 9120...........................................................................59
Figure 4-3 Circos plot sample 33432........................................................................60
Figure 4-4 Circos plot sample 34366........................................................................61
Figure 4-5 Circos plot sample 34394........................................................................62
Figure 4-6 Circos plot sample 35562........................................................................63
Figure 4-7 Circos plot sample 35649........................................................................64
Figure 4-8 Circos plot sample 35818........................................................................65
Figure 4-9 Circos plot sample 38532........................................................................66
Figure 4-10 Circos plot sample 48585.......................................................................67
Figure 4-11 Circos plot sample 183410......................................................................68
Figure 8-4 Expression of EMT Genes ................................................................. 139

Figure 8-5 Expression of genes with breakpoints ............................................. 140

Figure 9-1 Publication rates per year for the last 10 years comparing cSCC, Skin SCC and Melanoma ................................................................. 144
Papers included in this thesis:


*This manuscript was used as the basis for the introduction chapter with updates and expansion as required.*

**BA contribution:**
BA – concept development, data collection and analysis, interpretation of results, manuscript draft and revisions


*Text of this paper and the manuscript for a larger follow up study are included in the Introduction.*

**BA contribution:**
BA – patient recruitment, therapeutic interventions and data analysis

Mooney CP, R Martin, R Dirven, **B Ashford**, K Shannon, CE Palme, Q Ngo, J Wykes, S Davies, K Gao, S Ch’ng, TH Low, R Gupta, JR Clark Sentinel node biopsy in 105 high risk cutaneous SCC of the head and neck: results of a multicentre prospective study. (Under review) *Text of this manuscript is included in the introduction regarding an expanded data set in our prospective sentinel node study in cSCC.*

**BA contribution:**
BA – patient recruitment, therapeutic interventions and data analysis

Sections on non-coding variation are included from this manuscript.

BA contribution:
Study design, data collection, manuscript


BA contribution:
– Study design, data collection and analysis, manuscript review
Other relevant publications during candidature


1 Introduction

The perfect storm for cutaneous malignancy would be for a fair-skinned people to occupy a land with extreme ultraviolet radiation for most months of the year, to serially avoid sensible preventive strategies to avoid excessive solar skin damage, and to have an excellent health system which means the majority of the inhabitants live beyond 80 years of age. This perfect storm is the public health and clinical scenario that has given rise to the tsunami of cutaneous Squamous Cell Carcinoma (cSCC) in Australia, New Zealand and the US.

For at least 40 years, Australia has been at the forefront of public health campaigns to promote preventive sun exposure strategies. It is likely these efforts will eventually lead to a decrease in the incidence of cSCC. But aside from a few examples, our research has centred on elegantly describing what we see in the clinic and observe during follow up. Very little is known about the molecular basis of advanced cutaneous squamous cell carcinoma. The era of massive parallel sequencing has allowed for a deeper understanding of the molecular biology of many cancers. The accessibility of such processing lead to our employing whole genome sequencing to further explore the mutational profile of metastatic cSCC. Our challenge, and responsibility, was to use the opportunity afforded us by our clinical workload to progress the understanding of the molecular and genetic basis of metastatic cSCC.

1.1 Clinical context

Cutaneous squamous cell carcinoma (cSCC) is the second most common malignancy, after basal cell carcinoma (BCC), affecting up to 500 000 people in the United States (US) annually (American Cancer Society 2018). The non-melanoma skin cancers (NMSC) are more common than all other cancers combined (Fransen, Karahalios et al 2012). The burden of NMSC is so great that central cancer registries in both the US and Australia specifically exclude the collection of data on their incidence. Though absolute incidence of NMSC and cSCC in Australia is simply unknown, current estimates are >176,000 cases per year (Perera
et al 2015). In a 2017 review of Australian Health Insurance Commission billing items and tissue subtypes (Pandeya et al 2017), the incidence of primary cSCC was estimated at 270 per 100 000 person-years.

![Image of a 29 year old male with left lower lip cSCC with metastasis to ipsilateral upper cervical lymph nodes. Treatment entailed radical excision of lower lip cSCC and selective neck dissection followed by adjuvant external beam radiotherapy (Ashford patient).](image)

The highest incidence is observed in Queensland, Australia where the annual age adjusted incidence of cSCC was determined to be 573/100 000 in males (Pandeya et al 2017).

Lymph node metastases occur in approximately 2-5% of cSCC (Venebles, Autier et al 2018, Forest, Clark 2010). The majority of cSCC arising in the scalp and face generally show a predictable pattern of spread predominantly to intraparotid, level II (upper jugular) and perifacial lymph nodes (Forest, Clark et al. 2010) see also Figure 1-1 and Figure 1-2.
The pattern of lymph node metastasis from varying regions of the head and neck was first described in a landmark survey of 209 patients with metastatic cSCC from known primary locations (Vauterin, Veness et al 2006). Of particular note was the importance of the nodal basins of the parotid gland and the drainage along the external jugular vein lymphatics (Figure 1-3). This was the first detailed description of this pattern of metastasis in cSCC in an Australian cohort.
Limited progress has been made in the management of regionally advanced disease over the last 15 years. Most patients are managed with parotidectomy, neck dissection and adjuvant external beam radiotherapy depending on the site and stage at diagnosis (Veness 2005). Treatment failures are usually locoregional. Durable salvage following locoregional failure is rarely achievable as there are no effective second line therapies. Options for retreatment in this setting principally include radical surgical resections and are highly morbid, with profound physical, functional and psychosocial effects that greatly affect the quality of life for both patients and their carers. The ageing Australian population also means that there will be an increase in the incidence of cSCC, principally neglected head and neck cutaneous primary disease and metastases to the parotid and the regional lymph nodes.
The Sydney Head and Neck Cancer Institute manages advanced/metastatic cSCC in a multidisciplinary setting and treats in excess of 80 cases per annum and has published extensively on the staging, clinicopathological features and outcomes of metastatic cSCC (Ch’ng 2013, Forrest 2010, Gore 2016). My clinical practice in Wollongong included more than 30 cases of metastatic cSCC/year over the years 2013-16.

Unfortunately, despite high-level prospective data, the best predictors of nodal metastases such as tumour thickness and perineural invasion are not sufficiently discriminatory to change clinical practice because the majority (>85%) of ‘high risk’ tumours will not develop metastases. In addition, the ‘high risk’ predictors for nodal metastases do not address the critical group with medium to low-risk local disease where early intervention for nodal metastases will impact on survival (Ch’ng, Clark et al 2013). This was clearly demonstrated in the largest series of sentinel node biopsies for ‘high risk’ cSCC where mortality due to local recurrence out-weighed any beneficial effect (see Gore et al 2016 below).

1.2 Aetiology of cSCC

Ultraviolet (UV) radiation exposure and immunosuppression are the key factors in the development of cSCC (Alam, Ratner 2001). UVB (280-315nm wavelength) is the dominant environmental risk factor for cSCC (Armstrong, Kricker 2001). Long term sun exposure implies an increased risk, particularly in individuals with skin photosensitivity and lower melanin content as determined by the Fitzpatrick classification of skin phenotypes. This is more frequent in males and increases with age. UVB induced p53 loss of function is the index event in cSCC carcinogenesis.

Immunosuppression is a critical risk factor for both the development and metastasis of cSCC(Euvrard, Kanitakis et al. 2003). Nearly 82% of Australian transplant recipients will
develop skin cancers 20 years post transplant (Veness, Quinn 1999). The immunosuppressed population shows higher recurrence rates of cSCC. The rates of metastasis are ten times greater than immunocompetent individuals. The degree of immunosuppression also confers greater risk for development of cSCC following transplant. For example, heart transplant recipients and those who suffer from rejection events within the first year generally require greater levels of immunosuppression and have higher rates of cSCC than the wider transplant population (Veness, Quinn 1999).

Rare inherited genetic conditions such as Xeroderma pigmentosa, an autosomal recessive disorder of DNA repair following UV induced injury, lead to cutaneous malignancy and death due to melanoma and cSCC (Soufir 2000).

1.3 Staging of cSCC

Staging of solid organ malignancies is according to the American Joint Committee on Cancer (AJCC) Cancer Staging Manual, currently in 8th edition. This Manual stratifies disease in clinical and pathological stages in terms of primary Tumour, Lymph Nodes and distant Metastasis (TNM). The current version for SCC of the head and neck is presented in Appendix 1. Nodal metastasis confers a five year survival ranging from 55 to 75% (Brunner, Veness 2013).

It is worth noting that there is no discrimination between mucosal and cutaneous disease, which has led to criticism as to the utilities of the staging system to act as a prognostic tool (Liu, Ebrahimi 2017). The limitations of the 8th edition of the AJCC Cancer Staging Manual with respect to cutaneous head & neck SCC are indicative of our relative lack of understanding of this disease.
1.4 Conventional methods for assessment of metastatic potential of cSCC

Traditional means of assessing risk of metastasis in primary cSCC include clinical features such as location of the primary (lip and ear) and immunosuppression, and histopathological features such as tumour thickness, poor differentiation, lymphovascular invasion and perineural infiltration (Peat, Insull et al. 2012). These semiqualitative factors are highly unreliable predictors of metastatic potential.

Immunohistochemistry (IHC) is a cost effective method of detecting protein expression that can be easily integrated into the routine clinical diagnostic workup of patients. Differential protein expression secondary to mutations or post translational changes could be explored in patients with cSCC.

Currently, there are no reliable immunohistochemical stains that can be used for evaluation of metastatic potential in routine diagnostic work up of any organ system. Various experimental stains including focal adhesion kinase (FAK) are being evaluated in gynaecologic and urologic malignancies with promising results in estimation of metastatic spread (Aust, Auer 2014). Other experimental immunostains such as wingless type receptor homolog 6 (FZD6), pleiotropin (PTN), cathepsin and matrix metalloproteinases (1,10,13) have been evaluated in research settings but are not used in clinical practice.

In a well-designed prospective study (Brantsch, Meisner et al. 2008), thickness, immunosuppression, location on the ear and maximum tumour dimension were identified as significant predictors of metastasis. Factors predicting for nodal metastases are often interchangeable with prognostic factors. In a review of 239 patients with established nodal metastases from cSCC (Ch’ng, Clark 2013), multivariate analysis of the primary tumour
including tumour differentiation, margin status, tumour size and thickness found that only poor tumour differentiation of the primary was associated with shorter disease free survival. Similarly, Brantsch et al identified thickness >6.0mm and poor differentiation as significant factors for metastasis. Our study outlined below (Gore, Shaw et al 2016) reporting a large series of sentinel node biopsies for high risk cutaneous SCC supports the importance of tumour thickness. Nodal metastases only occurred in tumours more than 4mm thick and the number of high risk factors (4 or more) was the best predictor of nodal metastases, rather than any individual clinicopathological variable.

We are not the first group to attempt to establish a link between types of primary tumours and the likelihood of metastatic disease. Sentinel node biopsy (SNB) is used in numerous malignancies including breast and penile carcinoma, as well as in melanoma to stage nodal disease in higher risk primary tumours (Giuliano, Kirgan et al 1994, Kelley, Ollila et al 1998). Sentinel node biopsy allows for the detection of low volume nodal metastasis in the most likely first echelon node immediately downstream of the index lesion. The facility of detailed examination of a single lymph node is to allow for multiple and serial sectioning of a single node not possible in the multiple lymph nodes seen in a larger sample harvested in a formal neck dissection. The examination of sentinel nodes in high risk cSCC might allow for the alignment of risk factors beyond those previously described for metastasis.

Our group has published the largest prospective series of sentinel node examination in high risk cSCC (Gore, Shaw et al 2016). Patients with high risk primary cSCC were prospectively enrolled to undergo sentinel node biopsy. Full clinicopathological status was documented and the patients were followed for recurrence and also emergence of occult nodal disease despite negative sentinel node status (false negative biopsy).
Inclusion criteria comprised at least one of these characteristics in a patient with cSCC:

1. Tumour size > 2cm
2. Invasion into subcutaneous fat or tumour thickness >5 mm
3. Poorly differentiated tumour
4. Perineural invasion
5. Lymphovascular invasion
6. Local Recurrence in the setting of adequate prior resection margins
7. Ear or lip location
8. Immunocompromise (post-organ transplant, chemotherapy)
9. Carcinoma in pre-existing scar

Enrolled patients received treatment based on the nature of their presentation. Patients who presented with a high-risk primary or recurrent cSCC were offered wide excision of the tumour and concurrent SNB. Those who presented following excision of a cSCC that was subsequently confirmed (by pathological or clinical criteria) to be high-risk were offered secondary SNB, either alone or along with a wider excision if that were deemed appropriate.

Identification of the likely sentinel node was achieved using both radio-isotope and intraoperative mapping. Pre-operative lymphoscintigraphy was performed according to local nuclear medicine department protocols. Intra-operatively patent blue dye or isosulfan blue was injected intradermally at four points around the middle of the scar, tumour or the edge of a small skin graft or flap.

SNB was typically performed prior to primary tumour excision, unless the location of the primary tumour hampered the detection of the sentinel node field in which case the primary tumour site was resected first. Incisions for SNB were made in appropriately planned sites for inclusion in potential completion lymph node dissections. Sentinel nodes were identified by
the combination of the pre-operative lymphoscintigram, the visually identified “blue” node and the use of a hand-held gamma probe.

Primary tumour resection margins were individualized according to clinicopathological criteria, in all cases aiming for macroscopic tumour clearance and histologically clear margins. Wound closure was at the discretion of the operating surgeon.

All Sentinel nodes were cut along their longitudinal axis in 3 mm thick slices and embedded entirely in paraffin blocks following tissue processing. Four sequential 5μm tissue sections were cut from each block and stained with haematoxylin and eosin (sections 1 and 4) and cytokeratins for immunohistochemistry (sections 2 and 3). Each section was examined microscopically for the presence of metastatic tumour cells by an experienced histopathologist.

SNB was undertaken in 45 patients (79%) at the time of cSCC resection, of which 17 (30%) were recurrent tumours. In 12 cases (21%) pathological analysis of the primary tumour following excision prompted SNB and further wide local excision was performed to achieve adequate margins.

The mean tumour diameter was 25mm (range 6 - 65mm) and mean depth of invasion was 9.2mm (range 1.0 – 22mm). In 44 cases the tumour was over 5mm thick or was at least invading to Clark level 4. In 22 patients previous surgery had been performed in the region of the tumour and 31 patients had been treated for prior non-melanoma skin cancer. The mean number of sentinel nodes identified on pre-operative lymphoscintigraphy was 2.2 (range 0-6). Two cases underwent immediate selective neck dissection because of the finding of
macroscopic tumour whilst the remaining 55 cases had only sentinel nodes resected. In the 55 SNB cases the number of resected nodes was slightly higher than those identified on lymphoscintigraphy (mean 2.7, range 0-7). In four cases no nodal tissue was identified on histopathology despite confirmation with blue dye and intra-operative gamma probe. This occurred exclusively in patients with prior radiotherapy and scalp primaries with drainage to retroauricular or occipital nodal basins.

In total, seven patients (12.3%) had subclinical nodal metastasis detected at the time of planned SNB. In five cases (8.8%) micro-metastatic SCC was detected on pathological examination and in two cases (3.5%) macroscopic tumour was discovered at the time of sentinel node biopsy exploration. All seven patients proceeded to therapeutic lymphadenectomy; the five SNB cases were performed as completion lymphadenectomies following histopathology results whereas the two patients with macroscopic disease proceeded to immediate selective neck dissection. Of five patients who had nodal micro-metastasis and proceeded to CLND, two had further metastatic disease identified in the neck dissection specimen (one further positive node in each of these cases).

Median follow up was 19.4 months (range 2.4 - 41 months). At the time of analysis nine (15.8%) patients had developed recurrence, of which three had subclinical metastasis detected by SNB. There were six (10.5%) patients who had died of cSCC, of which two had subclinical metastasis detected, giving a three-year disease specific survival rate of 82%. There were eight local failures (SNB +ve n=2), two regional failures (SNB +ve n=1) and three distant failures (SNB +ve n=2) One patient developed regional recurrence after failed sentinel node biopsy increasing the true number of patients with subclinical nodal metastasis to eight (14%). This patient underwent preoperative lymphoscintigraphy but although nodal tissue was removed this was not confirmed to be the sentinel node by standard criteria (blue
The patient developed recurrence at the site of exploration. In addition, one patient developed in-field regional recurrence after positive sentinel node biopsy, bilateral neck dissection and postoperative radiotherapy. This patient subsequently died of disease and was the only patient to die with established regional recurrence following positive SNB. The other five patients who died of disease during the follow-up period suffered local recurrence (n=4) and/or distant disease (n=2). No episodes of distant recurrence in the absence of local or regional recurrence have been noted. Patients with confirmed subclinical metastatic disease had a significantly higher mortality rate than those whose SNBs were negative for disease (p=0.0082).

In this study the factors associated with a high risk (>20%) of nodal metastasis were depth of invasion ≥10mm, lymphovascular invasion, perineural invasion, poorly differentiated tumours, location on the lip, nose or ear, and four (or more) factors combined. On multivariable analysis the strongest predictors of metastasis were the number of high risk tumour factors present, the presence of perineural invasion and also the presence of lymphovascular invasion. Depth of invasion remains an important consideration as all patients with metastasis had primary tumours more than 5mm thick.
Figure 1-4. Kaplan-Meier survival analysis of patients with (‘Nodal_Metastases = Y’) and without (‘Nodal_Metastases = N’) nodal metastasis. Patients with metastasis had significantly worse survival than those without (p=0.0082).

Our group has recently updated the data on this study to include a total of 105 SNB procedures were performed on 104 patients, with one patient undergoing two SNB’s for metachronous lesions 32 months apart (currently under consideration for publication). The average age at time of SNB was 65 years with a strong male preponderance (male:female = 90:14). The SNB was performed at the time of initial lesion excision on 41 occasions (39%), after pathological examination of the lesion, with or without further excision to achieve adequate margins, on 31 occasions (30%) or at the time of recurrent lesion resection on 34 occasions (32%).

In total, 15 patients (14.3%) had subclinical nodal metastases, including 10 patients with a positive sentinel node (9.5%) and an additional five patients (4.8%) who developed nodal recurrence on follow up. Macroscopic disease was identified at the time of SNB in four cases and CLND was undertaken at the time of SNB. Microscopic disease was identified on pathological examination of the sentinel nodes in six cases with three having CLND as per protocol. The remaining two patients underwent post-operative radiotherapy and one declined any further intervention. This patient developed recurrence 11 months after SNB and subsequently underwent a level II-V neck dissection with 3 of 26 nodes positive. Of the seven patients who underwent either immediate or staged neck dissection as per protocol, the median number of nodes removed was 35 (range 6-60) and the median number of involved nodes was 2 (range 1-14), including nodes removed with the SNB. The negative predictive value of SNB was 94.7%. Overall sensitivity for SNB was 66.7% with a 100% specificity.
Median follow up was 26.2 months (range 0.3-78 months). At the time of analysis, there were 13 local recurrences (four SNB positive), eight regional recurrences (three SNB positive) and four distant metastases (three SNB positive). One patient developed nodal metastases more than five years after SNB in the context of multiple subsequent cutaneous SCC excisions and this was not believed to be related to the index lesion for which the SNB was performed. Another patient developed distant disease (L2 vertebra deposit) in the absence of local or regional recurrence. In total, 10 patients died from cSCC during follow up with a 5-year disease-specific survival rate of 83.1%. Patients with subclinical nodal metastases had a significantly higher mortality compared to those without nodal metastases (p<0.0001). Of the deceased, death was due local failure in three, regional failure in two, distant failure in two, local and regional failure in one, local and distant failure in one, and one patient died with local, regional, and distant failure.

On univariable analysis, the only significant predictor of subclinical nodal metastases was depth of invasion. The rate of nodal metastases in patients with DOI \( \geq 5 \text{mm} \) was 19.7% compared to 0% in patients with DOI < 5mm (p=0.01). The rate of nodal metastases in patients with DOI \( \geq 10 \text{mm} \) was 25% compared to 0% in patients with DOI < 5mm (p=0.001). However, in patients with both a depth of invasion \( \geq 5 \text{mm} \) and perineural invasion, the rate of nodal metastases was 28% compared to 8.2% in patients who did not fulfil both criteria (p=0.02). The median number of inclusion criteria for patients with subclinical nodal metastases was four as compared to three for patients without nodal metastases (p=0.036).

No individual high-risk feature (i.e. inclusion criteria) was significant in predicting subclinical nodal metastasis on multivariable analysis using logistic regression. Using a backward elimination method, only the number of inclusion criteria was statistically
significant (p = 0.035; OR 3.3 for 4 or more compared to 3 or less). When the number of inclusion criteria was excluded from the model and the depth criteria was set to greater than 10mm (as all positive SNB lesions were greater than 5mm), then the only significant predictor of subclinical nodal metastases was depth greater than 10mm (p = 0.043; OR 3.2). Once again the survival data underlines the significant negative prognostic impact of nodal metastasis.

![Graph showing survival data](image)

**Figure 1-5.** Expanded survival data for patients undergoing sentinel node biopsy. Nodal metastasis (N=15), no nodal metastasis (N=89). Data taken from Mooney, Martin manuscript (listed in publications cited earlier)

### 1.5 Genomic observations

As discussed above, the conventional clinicopathologic prognostic markers in cSCC have been shown to be highly unreliable predictors of metastasis in our own prospective study of SNB. Metastasis driver mutations may exist in cSCC that could be used to more reliably predict metastasis risk in primary high risk cSCC at the time of initial presentation. However,
relatively little is understood of the molecular and genetic basis of metastasis in cSCC. Clinically relevant and reproducible molecular signatures that can predict metastatic potential in the primary cSCC at initial presentation have not been described. An understanding of the molecular and genetic pathways of metastasis in cSCC has the potential to positively influence countless lives affected by cSCC in countries with high solar exposure by allowing for better stratification of risk, more efficient surveillance and the delivery of targeted therapies that may have lower toxicity and increase potency in patients with recurrent disease.

Within the genomic examination of cSCC have been papers utilising targeted next generation sequencing (NGS) and whole exome sequencing (WES) (discussed further below). There are currently no reports of whole genome sequencing (WGS) being employed in cSCC (primary or metastatic) and there is no WGS data for cSCC in The Cancer Genome Atlas (TCGA) (https://portal.gdc.cancer.gov). Reference will be made throughout this introduction to the techniques used as applicable.

Skin is our first line of defence against environmental assault. The established model for carcinogenesis (Hanahan, Wienberg 2011) involves genomic instability, cell cycle dysregulation, induction of telomere maintenance mechanism and an angiogenic switch. Normal, sun-exposed skin harbours many of the key driver mutations that are seen in cSCC (Martincorena 2015). This study used targeted NGS and highlighted the breadth and frequency of UV associated mutations in macroscopically normal skin excised in the course of blepharoplasty. Common gene mutations included those within the \textit{NOTCH} family and \textit{TP53}, at a frequency about 10\% of that seen in cSCC. This underlines the continuum
between early genotypic changes tolerated in “normal” skin and the phenotypic expression of atypia, dysplasia and ultimately cSCC.

One of the few publications to cite genetic abnormalities in cSCC highlights the density of mutations (Durinck, Ho et al. 2011). These authors performed WES of eight primary cSCC and found approximately 1300 somatic single-nucleotide variations per cSCC exome (1/30000 base pairs). This is amongst the highest degree of mutation observation for any cancer described. The authors hypothesize that the constant reinjury to the basal epithelial cells by UV radiation is at the root of the multiple mutational events in cSCC, including significant TP53 (100% of samples) and NOTCH1,2 mutations. Variation in established cancer associated genes are outlined below for cSCC. These include TP53, NOTCH, RAS, CDKN2A and PTPR.

1.5.1 TP53 family

In cSCC, UVb exposure is implicated in gatekeeper TP53 mutation. This hypothesized initiating mechanism is seen in both pre-neoplastic actinic keratosis as well as in invasive SCC in UV exposed skin changes (Benjamin and Ananthaswamy 2006). TP53 (chr17p13.1) encodes the protein p53 which belongs to a family of transcription factors comprising p53, p63 and p73. p53 impacts transcription regulation via effects on cyclin dependant kinase activity. UV-induced mutation of TP53 has long been thought to be critical in loss of programmed cell death in abnormal squamous cells (Armstrong, Cricker 2001) and inactivating TP53 mutation has been understood since the 1990s as a cause of cutaneous malignancy (Basset-Seguin, Moles et al 1994).
TP53 family mutations can give rise to measurable decreases in expression of other putative tumour suppressor factors such as IFR6, a member of the Interferon Regulatory function transcription factors. IFR6 is involved in skin development, due in part to its interaction with p63. A study comparing cSCC with normal matched skin (Botti, Spallone et al. 2011) found subsequent downregulation of IFR6 in 71% of cSCC (cell line A431). This analysis correlated downregulation of IFR6 expression with poor differentiation and increased invasive potential. Altered expression probably also infers loss of intercellular adhesion, and perhaps, metastatic potential. Mutation of another tumour suppressor, DICER, has been suggested to augment the deleterious effect of TP53 mutation (Su, Chakrvarti 2010 and Lyle, Hoover et al 2014).

1.5.2 NOTCH family

The NOTCH signalling pathway is crucial to epidermal development and maturation. (Moriyama, Durham et al. 2008). There are 4 families of NOTCH receptors (1-4). The genes are located at NOTCH1 (chr9q34.3), NOTCH2 (chr1p12), NOTCH3 (chr19p13.12) and NOTCH4 (chr6p21.32). NOTCH is a transcellular membrane molecule that once activated transfers its intracellular domain into the nucleus. Intranuclear activation of various genes are reliant on NOTCH.

NOTCH receptor activation or signalling, depending on the cellular context, can have either an oncogenic or tumour suppressing role. Whole exome sequencing of head and neck mucosal SCC demonstrated that up to 15% of tumours show mutation in NOTCH1 receptors (Agarwal 2011) and loss of NOTCH1 is associated with disease progression. Thus whether it is inactivation of oncogene action, or loss of tumour suppression function, changes in NOTCH activity has implications in epithelial maturation and differentiation. NOTCH expression and activity can be manipulated by commensal human papilloma virus (HPV).
1.5.3 **RAS Family**

The RAS gene family comprise 3 RAS genes (H-, K- N-RAS) which encode a family of small GTPases (Marshall 1996). These genes map to chromosome 11p (HRAS), 12p (KRAS) and 1p(NRAS). The active form exists with bound GTP (rather than GDP in the inactive state). RAS activation leads to downstream effects on cellular regulating molecules such as RAF, MEK and MAPK. Some mutations of RAS can lead to prolonged activation due to insensitivity to processes that dephosphorylate GTP. Activators of RAS, eg RASGRP1, a guanine nucleotide exchange factor, have been shown in mouse models to promote spontaneous cutaneous tumours when overexpressed (Sharma, Fonseca et al. 2014). Knockout mice also resist tumour formation.

Activating RAS mutations are expected in up to 9% of cSCC. HRAS mutation is more commonly associated with cSCC than NRAS and KRAS. In a recent German study assessing focused sequencing analysis of FFPE extracted DNA, only 1/31 cSCC showed a RAS mutation (Mauerer, Herschberger et al. 2011). Again, the pattern of tumour suppression loss rather than oncogene activation that typifies cSCC was observed.

In a study of targeted sequencing of DNA from cSCC in 21 patients on the various vemurafenib trials (Su, Viros, 2012), activating RAS mutations were identified in 60% of the cancers indicating that the MPAK altering repercussions of BRAF inhibition seem to have a potent oncogenic influence. In this study, HRAS mutations were most prominent.

1.5.4 **CDKN2A**

CDKN2A maps to chromosome 9 and encodes for two cell cycle regulatory proteins p16 and p14, which act via retinoblastoma and p53 pathways respectively. Mutations of CDKN2A lead to loss of function, and subsequent loss of expression.
Primary and metastatic cSCC were compared in a study investigating the prognostic significance of $TP53$, CDKN2A and HPV status in metastatic cSCC (Kusters-Vandevelde, Van Leeuwen et al. 2010). This study examined formalin-fixed, paraffin embedded specimens from both primary and metastatic cSCC, and normal tissue from 35 patients. Normal tissue was included for exclusion of germline mutations. Highly targeted sequencing was followed by PCR amplification from extracted DNA to assess $TP53$ and CDKN2A (both p16 and 14) mutations. They observed an increased rate of CDKN2A mutation (31%) in the metastatic tumours when compared to sporadic primary cSCC (Soufir, Daya-Grosjean 2000), but at similar levels to Xeroderma pigmentosa primary cSCC.

### 1.5.5 Protein tyrosine phosphatase receptors

PTPRs are cell surface, transmembrane receptors involved in cell signalling via tyrosine phosphorylation. There are 21 PTPR. Some members of the family of PTPRs function as tumour suppressors. Loss of function due to mutation may have flow on effects on downstream signalling. PTPR mutations have been observed in mucosal SCC, and a recent publication (Lui, Peyser et al. 2014) hypothesizes a role for loss of function of PTPRT leading to elevated levels of phosphorylated (activated) STAT3(oncogene) in head and neck (mucosal) SCC. There is no data to support or refute the role of this group in cSCC. This early exploration of genomic mutations associated with cSCC development barely begins to explain the complex pathway interaction dysfunction in cSCC. Some of these mutant pathway modifications will no doubt be responsible for not just cutaneous carcinogenesis, but will be implicated in the metastatic process, by altered expression of factors that enhance proliferation, extracellular stromal interaction and ultimately escape from the primary site.
1.5.6 More recent genomic observations for cSCC

Pickering et al (Pickering, Zhou et al. 2014) have recently progressed the survey of the genetic landscape of high risk cSCC. Their group performed WES of DNA extracted from high risk head and neck cSCC fresh tissue. Their aim was to establish the mutation pattern and frequency, and to identify driver mutations. They identified a high rate of mutations with a common pattern of UVB induction (C>T transition), and a strong preponderance toward inactivating mutations of tumour suppressor genes. Predictable loss of function mutations in tumour suppressor genes including TP53, NOTCH1 and NOTCH2 were observed. New driver gene mutations were observed in RIPK4 and RASA1. However, the identification of actionable mutations in cSCC was thwarted by the predominance of tumour suppressor gene inactivation as opposed to oncogene activation. Detailed differential mutational patterns in primary and metastatic cSCC were not explored.

Li, Hanna et al. (2015) examined single nucleotide variation and copy number variation in 29 formalin fixed, paraffin embedded lymph node metastases of cSCC using a 504 cancer gene panel and categorized mutations as belonging to any of 4 categories: the RAS/RTK/PI3K pathway, cell cycle pathway (TP53, CDKN2A), squamous differentiation pathway (TP63, NOTCH) or epigenetic (chromatin remodelling) genes. This is the first genomic analysis published addressing nodal metastases and offers valuable information. The most recurrently altered genes seen were TP53 (79% of cases), CDKN2A (48%) and NOTCH 1/2/4 (69%). They observed activating alterations in the RAS/RTK/PI3K pathways in 45% of samples. An association was observed between this activation and shorter progression free survival in their cohort of 29 cases. Other activating mutations in various tyrosine kinase / kinase pathways, upstream of known significant cell survival augmenting mediators such as MEK
and mTOR were observed. These included *KIT, KRAS* and *BRAF*. All specimens were judged as HPV negative by p16 IHC and analysis of HPV E6/7 genes. Interestingly, these authors reported no adverse effects from the challenges of nucleic acid extraction, processing and interpretation in the setting of formalin fixed specimens.

The substantial mutation rate in the *PIK3/Akt/mTOR* pathway in the study by Li et al (2015) suggests a role for this central cascade in the metastatic spread of cSCC. This has been previously described as a point of difference between gastric cancers with and without peritoneal metastatic spread (Liu 2010). Using RT-qPCR on extracted RNA they demonstrated a 5x and 2x higher rate of mutation of the *PIK3CA* (encoding the catalytic subunit of p110alpha) in normal gastric and primary gastric cancers, respectively. This disordered regulation of the *PIK3/Akt/mTOR* pathway is a potential site of action on mutations enhancing metastasis.

The identification of deleterious mutations likely to be drivers of carcinogenesis, and potentially metastasis, is hampered by the high level of background mutation in sun exposed skin as mentioned above (Martincorena 2015). A prospective study of sentinel node analysis on high risk cSCC (Gore, Shaw et al. 2016) confirmed non-metastatic cSCC with proven negative sentinel nodes and long clinical follow-up. In a related collaborative study using DNA from FFPE primary cSCC using a 48 cancer gene panel a surprising number of likely deleterious mutations in key driver genes (eg *PIK3CA, NRAS, APC*) were found when compared to uninvolved lymph node for germ line control (Zilberg, Lee et al, 2018).

In an analysis of the mutational landscape of metastatic tumour deposits (total n=10000) from multiple malignancies, Zehir et al included 27 cases of metastatic nodal cSCC all with DNA
extracted from FFPE samples and subjected to targeted NGS (Zehir, Benayed et al. 2017). The custom panel of genes targeted included 410 genes that included known oncogenes and tumour suppressor genes, as well as key non coding regulatory areas like the TERT promoter. Overall, only 6% of their samples had insufficient DNA yield to be used, and their average coverage for the panel was 718X. They found the most frequently observed variant in metastatic cSCC (indeed in all tumours analysed) was TP53. They did note 32% of patients harboured TERT promoter variants, albeit in a pattern different from that described by Huang et al (Huang, Hodis et al. 2013) in melanoma.

A recent reanalysis of 40 cSCC (Inman, Wang et al. 2018), predominantly in immunocompromised patients (mostly solid organ transplant recipients) found a new suite of significantly mutated genes. These included the often reported TP53, NOTCH 1 and 2 and CDKN2A, but also identified HRAS, MAP3K9, PTEN, SF3B1, VPS41 and WHSC1 as well as deletion of HRAS. The key finding of this paper was the description of a new mutational signature (based on the Catalogue of Somatic Mutations in Cancer (COSMIC) signature patterns) within cSCC related to therapy with the immunosuppressant azathioprine.

1.6 Non-coding RNA – microRNA and IncRNA

There has been emerging interest in the role of both micro RNA and other non-coding but regulatory RNA elements in cancer. The research interest stems from not just their potential role in carcinogenesis and progression of malignancy, but also their potential for use as biomarkers. MicroRNAs are small (22 nucleotide) RNAs transcribed in the nucleus in a primer form to be activated by Dicer after transport into the cytoplasm (Reinhart, Slack et al. 2000). There they act on mRNA to block translation (Almeada, Reis et al. 2011).
In cSCC, a number of studies have examined the role of microRNA (miR). Perhaps the first of these was an examination of miR in cSCC cell lines, normal skin, actinic keratosis and frank cSCC (Xu, Zhang et al 2012). By RT pCR and immune fluorescence hybridization, the authors reported up-regulation of 4 miR (31, 135b, 21 and 223) as well as down-regulation of 54 miR, including 125b, in both cSCC and cell lines when compared to both actinic keratosis and normal skin which they found by microarray analysis targeted \textit{MMP13}. They identified that miR-125b suppressed proliferation and colony formation as well as the migration and invasive capacity of cSCC cells and that down regulation reduced the deactivation of \textit{MMP13}.

miR-203 was found to be inversely correlated with tumour differentiation in cSCC samples and cell lines (UT-SCC-7 (metastatic) and A341), and was active against \textit{c-MYC}, effectively having a tumour suppressor effect (Lohcharoenkal, Harada et al 2016).

Using qPCR on RNA from in situ and invasive cSCC, miR-21, miR-103a, miR- 186, miR-200b, miR-203, and miR-205 expression levels were compared (Stojadinovic, Ramirez et al 2016). They found between invasive and in situ disease, both miR-21 and 205 were significantly upregulated, proposing a role in a more advanced state, by action of down-regulation of genes \textit{MEIS1}, \textit{KAT2B}, and \textit{BLMH}.

There is evidence for upregulation of miR-31 in invasive cSCC compared to both actinic keratosis and normal healthy skin (Wang, Landen et al 2014). miR-31 is thought to oppose \textit{ITGA5}, \textit{RDX} and \textit{WAVE3} and \textit{RhoTBT1} (Lin, Zhou et al 2017). With the exception of the inclusion of the (incompletely characterised) metastatic cell line UT-SCC-7, none of the aforementioned studies of microRNA have examined the metastatic stage of cSCC.
There is limited evidence for the role of long non-coding RNA (lncRNA) in metastasis of cSCC. Pipponen et al (2016) analysed the role of the long intergenic ncRNA LINC00162 in both cSCC and normal skin cell lines. They identified is overexpression in tumour cells (by RNA in situ hybridization) and not in normal skin cells. Knockdown resulted in suppression of growth of the tumour cells. They theorised LINC00162, which they renamed PICSAR (p38 inhibited cutaneous squamous cell carcinoma associated lincRNA), to act by regulation of ERK1/2 via DUSP6 downregulation.

1.7 Epigenetic changes in cSCC

Epigenetic changes reflect alterations in the histone supports of the nucleic acids. A recent report identified methylation profile differences in key CpG promoter sites between metastatic and non metastatic SCC and BCC (Darr, Colacino et al. 2014). DNA extraction and methylation analysis using formalin fixed and paraffin embedded specimens of 37 primary cSCC (and 5 BCC) showed hypermethylation at CpG sites and thus silencing of FRZB, TFAP2C and ASCL2 (Goldengate Cancer Panel) in cSCC that developed metastases as opposed to the cSCC that did not. Pickering et al (Pickering, Zhou et al. 2014) showed an inactivating mutation of KMT2C in 39 cSCC, which encodes a histone methylation complex to alter transcriptional regulation. This mutation has been identified in other malignancies and was associated with increased incidence of bone invasion and a shorter time to recurrence in cSCC.
Chromatin accessibility as a result of genomic variants is an important area of interplay between genomic and epigenomic analysis. An example of this is in the promoter region of *TERT* where alteration in ETS binding site motifs with resultant increased affinity for the transcription factor as a result of somatic variants leads to an increase in histone methylation and chromatin opening (Liu, Yuan et al. 2016).

### 1.8 Stromal influences and EMT in the tumour microenvironment

Tumour-stromal interaction is key to the metastatic process. Early invasion in cSCC is characterized by a desmoplastic stromal reaction. Fibroblasts from the stroma associated with cSCC have architectural and behavioural differences when compared to normal dermal fibroblasts (Commandeur, Ho et al. 2011). Increased expression of vascular endothelial growth factor C has been shown in supraglottic (mucosal) SCC to be predictive of lymph node metastases (Baek et al 2009). Similarly altered expression of VEGF is seen in the stromal microenvironment in cSCC (Moussai, Mitsui et al. 2011).

Phenotypic changes may result from pluripotent subpopulations of cells that behave as cancer stem cells. These cells can exhibit both epithelial-mesenchymal transition (EMT) and mesenchymal-epithelial transition (MET). The transition between epithelial and mesenchymal phenotypes may be partial and may also be reversible (Lamouille et al 2014). The normal junctional integrity and polarity of epithelial cells in skin can, when exhibiting EMT, be characterised by a loss of E-cadherin expression and an abundance of expression of N(neural)-cadherin. Even this simple observation and explanation allows an appreciation of the altered function mediated by transcription factors that are either under or overproduced in the tumour microenvironment. Transcriptional repressors of E-cadherins include the mesenchymal markers Snail, Slug and Twist (Thiery et al 2009). Expression of these
repressors has been shown to be increased in poorly differentiated cSCC when compared to benign or well differentiated tumours. (Chen, Takahara et al. 2013). Overexpression of the mesenchymal proteins Twist, Zeb1, Vimentin, beta-catenin and Podoplanin was observed in metastatic when compared to non-metastatic cSCC (Moussai, Mitsui et al. 2011). TGF-β through SMAD activation and intranuclear transport can upregulate EMT by activating transcription factors and inhibiting repressors, either directly, or through SNAIL to effect E-cadherin expression (Lamouille et al 2014).

Mesenchymal stem cells (MSCs) probably constitute a percentage of tumour associated cell population but may be attracted by tumour cells into the tumour-stromal interface (Chamberlain, Fox 2007). Karnoub, Dash et al (2007) described paracrine inducement of MSCs in the presence of breast tumour stroma, with increased metastasis. The urokinase-type plasminogen activation (uPA) system is a key component of extracellular matrix and basement membrane degradation and is overexpressed in more aggressive malignancies (Ranson, Andronicos 2003). The activation of this system allows for fibrinolysis via the conversion of plasminogen to plasmin and involves genes including \textit{PLAU}, \textit{PLAUR}, \textit{SERPINE2} and \textit{SERPINB2}.

Recently, Laurenzana, Biagioni et al (2015) showed the effect of TGF-β (canonical SMAD activation) on uPAR mediated mesenchymal expression within melanoma cells incubated with MSCs in vitro, as an upregulation of N-cadherin, a-SMA and vimentin, a decrease in E-cadherin and enhanced expression of the E-cadherin transcriptional repressors SNAIL1/2. In vivo, when bone marrow stem cells were co-injected with melanoma cells in a mouse model, tumour progression was rapid, however, animals were euthanized due to tumour burden prior to the exhibition of established metastatic disease. Such growth factor induced altered
expression with upregulation of pro-EMT mechanisms will be a focus of future research in metastasis, particularly with the emergence of organotypic culture models.

We have previously published (Morosin et al 2016) a pilot study investigating the presence of circulating tumour cells (CTC) in the blood of patients with known metastatic cSCC. In this study, peripheral blood was drawn from patients, with resectable lymph node disease peripheral circulation prior to any manipulation of the tumour at the time of lymphadenectomy. Ep-CAM and cytokeratin markers were used to positively identify CTC after exclusion of circulating cells expressing CD45 (lymphocytes). CTC were identified in 8/10 patients with metastatic cSCC, with tumour microemboli found in 3/10 samples. Notwithstanding technical considerations around cell surface expression in different EMT states, this study identified a potential role for CTC analysis in surveillance of patients post therapy to detect early recurrence.

1.9 Differential expression in cSCC

In a study comparing 2 cell lines of cSCC with a Bowen’s disease (carcinoma-in-situ) cell line, disease stage dependency dictated the expression of 1895 genes by using organotypic cultures and IHC, ISH and microarray (Serewko, Popa et al 2002). They found overexpression of EGFR, but reduced expression of FRA-1, MAPK and MAPKK.

More recently, the predictive utility of inositol polyphosphate 5-phosphatase (INPP5A) was analysed and measured against biological behaviour in cSCC (Cumsky, Costello et al 2019). This was built upon earlier work using gene array, FISH and IHC showing loss of INPP5A may occur early in cSCC evolution (Sekulic, Kim et al 2010). They theorised that deletion of the short arm of chr10 contributed to the loss. Using IHC, Cumsky et al showed that reduced
INPP5A expression was also more consistent in more aggressive disease, including in worse
differentiation, LVI and those that metastasised.

The landmark study comparing both transplant requiring immunosuppression induced cSCC
and actinic keratosis with that of normal skin in both the transplanted and other non-
transplanted patients (Nindl, Dang et al 2006) compared 22283 genes by microarray and
identified 9 genes over expressed in the tumours. These genes were CDH1, MAP4K4, IL-
1RN, IL-4R, NMI, GRN, RAB31, TNC, and MMP1. There were 4 genes that were under
expressed including ERCC1, APR-3, CGI-39 and NKEFB.

Garcia-Diez et al (2019) identified underexpression of NEK10 and overexpression of both
FOSL1 and BNC1 when comparing cSCC, actinic keratosis and normal skin in 10 matched
pairs of immunocompetent patients.

Expression of cyclin D1, a member of a family of cell cycle regulatory proteins, increases
proliferation and disorganization of epithelium and is increased in both pre-neoplastic and
malignant skin changes (Burnworth, Arendt et al. 2007). These authors describe a difference
in p16 expression levels between pre-neoplastic and invasive carcinoma in their study cohort,
hypothesizing that decreased p16 was an essential trigger for these preliminary changes to
allow for the invasive progression. They also identified increased telomerase reverse
transcriptase (hTERT) expression in both AK and SCC.

The search for biomarkers of progression in cSCC has thus far borne limited fruit. In a
review of analysis of their own cell lines and FFPE specimens, Kivisaari and Kahari (2013)
highlight extracellular proteases and inhibitors including MMP-7 and SerpinA1 as potential
candidates. Using genome wide expression profiling comparing cSCC cell lines with normal keratinocytes they were able to identify expression differences between malignant and benign conditions (Farshchian, Kivisaari et al 2011, Kivisaari, Kallajoki et al 2008). They then utilised rtPCR to validate the expression observations and ultimately used Western blot to confirm the translated protein effects of these. Primary cSCC FFPE samples were used for both IHC and also for tissue microarray construction.

Erythropoietin-producing hepatocellular Type B2 (EphB2) is a ligand for a class of receptor tyrosine kinase. It has been shown to have altered expression in metastasis in colorectal cancer (Guo, Zhang et al 2006), wherein loss of EphB2 is associated with a more aggressive phenotype using tissue microarrays from adenomas, colon cancers and their metastases. Farshchian, Nissinen et al (2015) conversely showed that EphB2 expression promotes carcinogenesis, invasion and migration in cSCC cell lines and xenografts suggesting it could be used as a biomarker. This study did not make specific observations of the metastatic context.

Using two datasets from the Gene Expression Omnibus (https://www.ncbi.nlm.nih.gov/geo/) comprising normal skin/cell lines and primary cSCC, Wei, Chen et al (2018) determined that only EGR3 was consistently and significantly differentially expressed. EGR3 encodes a transcription factor of the EGR family and impacts up to 330 genes some of which may lead to progression of malignancy through inflammatory mediators including IL-6 and 8 (Baron, Pio et al 2015). Once again, the datasets used by Wei, Chen et al did not include any metastatic tumours or cell lines.
1.10 Project design, rationale and aims

The design of this project was in response to a clinical question. An ongoing effort within our clinical research group is to identify the key determinants of metastasis and recurrence in cSCC. Annually we see more than 100 patients with metastatic cSCC. These patients had all had primary cSCC, often treated absolutely as per protocol, and yet they had recurred, commonly presenting with multiple lymph node metastases to the neck. The required treatment would usually include both surgery and adjuvant external beam radiotherapy, at significant quality of life cost to the patient, and significant financial cost to the community. As outlined above, our group had previously published a prospective study investigating the role of sentinel node biopsy in high risk cSCC with an interest in identifying which types of primary lesions were most likely to result in metastasis (Gore, Shaw et al 2016) Depth of invasion and perineural invasion were identified as key predictors of risk of metastasis, although the actual numbers of metastatic cases were low (15/104 patients enrolled). Nonetheless, the potential of interrogating the primary lesion by established clinical and pathological means to determine risk had been surveyed.

In addition to our focus on the primary lesion in a bid to understand the metastatic process in cSCC, it became an emerging question as to whether looking at the metastatic deposit for clues would shed further light on a means to establish risk. The theory behind this was that the clone responsible for metastasis must in part persist as one of the dominant clones in the metastatic deposit. Sequencing options available for this project included targeted NGS, WES and WGS. Each modality has advantages and disadvantages including technical and logistic demands, depth and breadth of coverage and cost of both sequencing and bioinformatic analysis. Given the discovery nature of the project and the availability of fresh tissue for sequencing, a determination was made to proceed with WGS. The benefits that were
anticipated by this choice included the coverage of non-coding and regulatory regions of the genome and the ability to make genome wide assessments of mutational burden. Ultimately this assessment proved valid. To understand the genomic profile of the metastasis, the aim was to analyse fresh, viable tumour DNA by WGS, as this had never been reported in metastatic cSCC. We felt that to use this investigative process in a disease which dominated our clinical practices constituted sound clinical research and offered us an opportunity to describe for the first time the mutational landscape of this disease.

1.10.1 Aims of this project

**Overall aim** – to describe the overall pattern of somatic variation in metastatic cutaneous squamous cell carcinoma and to define particular regions of the genome, patterns of variation and genes for further analysis in order to facilitate identification of potential biomarkers of risk for metastasis.

**Specific aims**

1. Collection of matched samples of blood (germline) and nodal metastatic cSCC (tumour) and extraction of nucleic acids for downstream analysis

2. Undertake WGS of quality controlled DNA to greater than 60X coverage for tumour and greater than 30X for germline

3. Subject raw data to manipulation and then bioinformatics analysis to assess overall somatic mutational burden, structural variation, copy number variation and to detect short variants across both coding and non-coding regions of the genome.

4. Analyse genes or regions that are either amplified or deleted, or genes that exhibit high impact or recurrent short variants for altered expression.
2 Methods

2.1 Patient Recruitment

Patients were recruited according to the protocol contained in the approved ethics application 14/397 UOW/ISLHD HREC and LH 15.047 RPA/LH HREC. The ethics process involved about 4 months of work and 2 revisions to enable the approval to be granted and the study to commence.

Recruitment commenced in February 2015. Patients with biopsy proven metastatic cutaneous SCC scheduled for surgery with a curative intent were eligible for inclusion of tissue for this study. All patients were initially seen in the consulting rooms of the surgeon and were then discussed at the local Head & Neck Cancer Multidisciplinary Team (MDT) meeting. Informed consent for both the procedure and for the collection of blood and tumour samples was obtained separately and, as per ethics approval, no patients had any parts of their treatment altered if they were unwilling to be involved in the study.

Recruiting patients was not an issue for the same reason the project was conceived; the burden of disease is nearly overwhelming. So it was that the flow of patients willing to contribute to the research effort was plentiful. In the first 18 months of collection of samples, the patient accrual target of 60 cases had been achieved.

Two centres were involved in sample collection. The Chris O’Brien Lifehouse at RPA (Camperdown) was the secondary site for sample collection and has had a tumour bank in place for some time. This is a formal process with a separate consent that allows for a standardized collection of key tissues, including demographics and sample handling. This is the ideal arrangement for a study such as this. The primary site for sample collection was
Wollongong Hospital. This was on the basis of proximity to our wet laboratory facilities and was particularly useful for the purposes of establishing a cell culture from freshly resected metastatic cSCC (J. Perry PhD Candidate UOW (Ranson)). At the commencement of the study, no formal tissue banking arrangement was in place in the Illawarra and Shoalhaven Local Health District (ISLHD). This meant that the quality control, clerking and storage of most of the tissue samples for the study fell to us and required a significant amount of time and technique critical detail. This requirement has now been usurped with the involvement of the Wollongong Hospital in the CONCERT (Centre for Oncology Education Research Translation) Biobank (http://concert.org.au/research/research-capabilities/concert-biobank). This facility is based at the Ingham Institute, but now runs a full service at the Wollongong Hospital.

Regular reports to the HREC were delivered and there was no departure from the promulgated protocol. No patients withdrew consent from the study and there were no complaints identified by our research team pertinent to the research protocol.

Patients were identified for inclusion in the study if they had biopsy proven metastatic cSCC. Most of these patients had a historical index lesion that had been excised or biopsied. All of our patient’s index lesions were from head and neck sites.

Resources allowed for the whole genome sequencing of 20 matched whole blood: tumour pairs to the 30X and 60-90X respectively. Prior to nucleic acid extraction, a snap frozen tumour sample was subjected to Pathologist review to determine tumour cellularity. Samples deemed > 30% tumour cellularity were then further processed. Clinicopathological data for
the patients with metastatic disease from which DNA was used for sequencing is presented in Table 2.1.

Examples of haematoxylin and eosin histopathological assessment of cellularity are presented in Figure 2.1.

Figure 2.1. Pictures clockwise from upper left (i) 2 x magnification of 35% cellularity sample moderately differentiated cSCC revealing islands of tumour separated by non-tumour fibrous tissue, (ii) Same tumour at 20 x magnification showing more cytological detail, (iii) 2 x magnification of 70% cellularity sample with less fibrous tissue, (iv) 20 x magnification of same tumour as (iii) showing cytological detail including higher grade (poorly differentiated) features.
Table 2.1 Clinicopathologic data for samples originally included for WGS analysis. Staging is according to AJCC 8th edition. LNY: Lymph node yield – total number of nodes resected. LNR: Lymph node ratio – number of positive nodes /total number of nodes. ECS: presence of extracapsular (nodal) spread. R Status: R0 – microscopically clear margin, R1 – Microscopically involved margin, R2 – macroscopically involved margin. Other: CLL- chronic lymphocytic leukaemia. Liver Txplant- liver transplant recipient on immunosuppression. RA Immomodulation- rheumatoid arthritis on immunomodulation therapy.

<table>
<thead>
<tr>
<th>ID</th>
<th>Sex</th>
<th>Age</th>
<th>Node site</th>
<th>Stage</th>
<th>Differentiation</th>
<th>LNY</th>
<th>LNR</th>
<th>ECS</th>
<th>Known primary</th>
<th>R status</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>184136</td>
<td>M</td>
<td>68</td>
<td>L parotid</td>
<td>N3b</td>
<td>Poorly diff</td>
<td>56</td>
<td>2/56</td>
<td>Yes</td>
<td>Ipsilateral temple poorly diff</td>
<td>R1</td>
<td></td>
</tr>
<tr>
<td>321773</td>
<td>M</td>
<td>78</td>
<td>L parotid</td>
<td>N2a</td>
<td>Poorly</td>
<td></td>
<td></td>
<td></td>
<td>Ipsilateral forehead PNI</td>
<td>R0</td>
<td></td>
</tr>
<tr>
<td>193958</td>
<td>M</td>
<td>74</td>
<td>R parotid</td>
<td>N3b</td>
<td>Poorly</td>
<td>42</td>
<td>2/42</td>
<td>Yes</td>
<td></td>
<td>R0</td>
<td></td>
</tr>
<tr>
<td>48585</td>
<td>F</td>
<td>79</td>
<td>L perifacial</td>
<td>N3b</td>
<td>Poorly</td>
<td></td>
<td></td>
<td>Yes</td>
<td>Left cheek</td>
<td>R1</td>
<td></td>
</tr>
<tr>
<td>184577</td>
<td>M</td>
<td>78</td>
<td>R parotid</td>
<td>N3b</td>
<td>Mod/Poorly</td>
<td>52</td>
<td>2/52</td>
<td>Yes</td>
<td>Ipsilateral ear</td>
<td>R1</td>
<td></td>
</tr>
<tr>
<td>4447</td>
<td>M</td>
<td>72</td>
<td>L parotid</td>
<td>N3b</td>
<td>Mod</td>
<td>62</td>
<td>2/62</td>
<td>Yes</td>
<td>Ipsilateral preauricular</td>
<td>R0</td>
<td>CLL</td>
</tr>
<tr>
<td>4699</td>
<td>M</td>
<td>77</td>
<td>R parotid</td>
<td>N3b</td>
<td>Mod</td>
<td>64</td>
<td>4/64</td>
<td>Yes</td>
<td>Ipsilateral pinna ear</td>
<td>R0</td>
<td></td>
</tr>
<tr>
<td>9120</td>
<td>M</td>
<td>66</td>
<td>L scalp</td>
<td>N2b</td>
<td>Mod/Well</td>
<td>2</td>
<td>2/2</td>
<td>No</td>
<td>Scalp</td>
<td>R0</td>
<td></td>
</tr>
<tr>
<td>200971</td>
<td>M</td>
<td>64</td>
<td>Bilateral</td>
<td>N2c</td>
<td>Mod</td>
<td>55</td>
<td>3/55</td>
<td>No</td>
<td>Lip</td>
<td>R0</td>
<td></td>
</tr>
<tr>
<td>183410</td>
<td>M</td>
<td>30</td>
<td>L SMG</td>
<td>N3b</td>
<td>Well</td>
<td>27</td>
<td>3/27</td>
<td>Yes</td>
<td>Ipsilateral lower lip</td>
<td>R0</td>
<td></td>
</tr>
<tr>
<td>34934</td>
<td>M</td>
<td>87</td>
<td>L parotid</td>
<td>N3b</td>
<td>Mod</td>
<td>16</td>
<td>1/16</td>
<td>Yes</td>
<td></td>
<td>R1</td>
<td></td>
</tr>
<tr>
<td>35649</td>
<td>M</td>
<td>64</td>
<td>L parotid</td>
<td>N3b</td>
<td>Poorly</td>
<td>109</td>
<td>3/109</td>
<td>Yes</td>
<td>Scalp</td>
<td>R0</td>
<td>CLL</td>
</tr>
<tr>
<td>35818</td>
<td>M</td>
<td>69</td>
<td>R parotid</td>
<td>N3b</td>
<td>Poorly</td>
<td>11</td>
<td>2/11</td>
<td>Yes</td>
<td>Ipsilateral face</td>
<td>R1</td>
<td></td>
</tr>
<tr>
<td>37130</td>
<td>M</td>
<td>74</td>
<td>L parotid</td>
<td>N1</td>
<td>Mod</td>
<td>70</td>
<td>1/70</td>
<td>No</td>
<td>Forehead</td>
<td>R0</td>
<td></td>
</tr>
<tr>
<td>38532</td>
<td>M</td>
<td>77</td>
<td>R SMG</td>
<td>N3b</td>
<td>Mod</td>
<td>108</td>
<td>3/108</td>
<td>Yes</td>
<td>Ipsilateral nose</td>
<td>R0</td>
<td>CLL</td>
</tr>
<tr>
<td>35562</td>
<td>M</td>
<td>65</td>
<td>R parotid</td>
<td>N3b</td>
<td>Mod</td>
<td>29</td>
<td>2/29</td>
<td>Yes</td>
<td>Forehead</td>
<td>R0</td>
<td>Liv Tx</td>
</tr>
<tr>
<td>658492</td>
<td>M</td>
<td>86</td>
<td>R parotid</td>
<td>N1</td>
<td>Mod</td>
<td>20</td>
<td>1/20</td>
<td>No</td>
<td>Ipsilateral face</td>
<td>R0</td>
<td>CLL</td>
</tr>
<tr>
<td>33432</td>
<td>M</td>
<td>70</td>
<td>Left neck</td>
<td>N3b</td>
<td>Poorly</td>
<td>4</td>
<td>4/4</td>
<td>Yes</td>
<td>Ipsilateral face</td>
<td>R1</td>
<td>RA Imm</td>
</tr>
<tr>
<td>34366</td>
<td>M</td>
<td>86</td>
<td>R parotid</td>
<td>N2b</td>
<td>Poorly</td>
<td>43</td>
<td>2/42</td>
<td>Yes</td>
<td></td>
<td>R0</td>
<td></td>
</tr>
</tbody>
</table>
Neck dissection for metastatic cSCC involves not simply the removal of the diseased node/s, but rather clearance of anatomical regions (levels) of lymph nodes allowing for the capture or harvest of occult metastases in more distant echelons. This allows for a lymph node yield that maximizes the accuracy of pathologic staging and is likely to deliver the best curative effort. Following surgery and pathological reporting, all patients were then again considered by the local MDT. Herein recommendations were made to the patient about adjuvant treatment which in most cases consisted of radiotherapy. Following definitive treatment, patients had regular surveillance in the rooms of the treating doctors.

There was no departure from established protocols for managing metastatic cSCC. Preoperative blood was drawn at the time of intravenous cannulation, usually by the attending Anaesthetist. At least 2ml of whole blood was collected in EDTA tubes that were then transferred to the laboratory for storage at -80°C.

### 2.2 Tissue Handling

On the day of surgery, a sample of the metastatic deposit, approximately 10mm³ was preserved for storage and downstream analysis. It was critical that the specimen retrieval did not compromise the work of the Pathologist. The aim was to sample a section of the metastatic deposit from an area free of macroscopic necrosis and at the leading edge of the tumour:stromal interface. The specimen was then restored in overall arrangement by suturing and submitted to normal histopathological processing initially involving fixation in buffered formalin.
The harvested fresh sample was stored in cold buffered PBS for transport to the wet laboratory. Once received at the laboratory, the sample was divided into 30mg cubes and snap frozen in liquid nitrogen to be stored at -80°C.

Prior to any nucleic acid extraction, samples were formally assessed by a Consultant Pathologist experienced in cutaneous malignancy. One of the 30mg blocks was delivered to the pathology laboratory on dry ice. Once received, the sample was treated as a normal frozen section; embedded in medium and sectioned on a cryotome. An estimate of the cellularity of the specimen (that component comprising viable tumour cells) was made. An arbitrary cut-off of 35% cellularity was used to determine is a sample was to be submitted for further processing. Almost half of our cases satisfied this cellularity threshold.

### 2.3 Nucleic Acid Extraction

Samples identified as having adequate cellularity were identified for nucleic acid extraction. Preparatory experiments for the nucleic acid extraction included comparing automated techniques of tissue homogenization. Following the homogenization and nucleic acid quality control application, the Miltenyi Gentle MACSTM system using the RNA01 program was selected for use with frozen tissue samples. No specific homogenization program for DNA extraction is supplied on this platform. Tissue was processed in M tubes (a proprietary canister with an internal blade that is driven by the housing on the Gentle MACS processor), after the application of 600µl of RLT buffer. The post processing emulsion was then centrifuged at 4000 rpm for 5 minutes at 4°C.

The resultant solution was then processed to extract both DNA and RNA using the AllPrep DNA/RNA-Protein Mini Kit (80004, Qiagen), according to the manufacturer’s instructions.
All resultant DNA and RNA samples were quantified using the NanoDrop (ND1000, Thermoscientific). The aimed absorbance levels for DNA and RNA were 260:280 ratios of 1.8 (+/−0.1) and 2.0 (+/−0.1), respectively. The results of nucleic acid extraction are presented in Appendix 2. DNA samples were further analysed using 1% TAE agarose gels. Results of gel electrophoresis are included in Appendix 3.

2.3.1 Germline (blood) nucleic acid extraction

Whole blood was used for germline DNA. Some samples acquired initially as serum delivered insufficient DNA even using a kit specifically design for same. It was decided that whole blood (as per the initial protocol) was the preferred germline source and the PureLink® Genomic DNA Mini Kit was employed. Samples were quality controlled quantitatively with Nanodrop and gel electrophoresis.

2.4 Sequencing

The staging of funding allocation necessitated the sequencing of DNA through 2 laboratories. Both the Macrogen service outsourced through the Australian Phenomics Facility at the Australian National University and the inhouse sequencing provided by the Kinghorn Centre for Clinical Genomics (KCCG, Garvan Institute of Medical Research) utilised an Illumina HiSeq X platform.

Specifically, the process at Australian Phenomics Facility (ANU) included:

1. Sample initial QC using nanodrop, agarose gel electrophoresis and Picogreen assay and sample final QC check using SNP arrays.

2. WGS reads were generated on HiSeq X (Illumina Inc, San Diego, CA) as follows:

   a. Library Kit Type: Truseq Nano DNA kit 350bp insert
b. 150bp Paired End; 60X mappable and 45X mappable (sequencing depth)
c. Expected Output: \( \geq 100\text{Gb} \) per lane at raw data level

3. Raw data was delivered in FASTQ, although both BAM and VCF were available.
The KCCG service only differed in the sequencing depth, this being 90X for tumour DNA and 30X for germline DNA.

### 2.5 Bioinformatic Workflow

The process of turning WGS raw data into interpretable events is always complex. This is made moreso by an increased mutational burden. Prior to this study, melanoma had been identified as having the highest mutational burden, averaging in the order of 49 mutations per megabase for cutaneous melanoma (much greater than for variants including uveal and acral lentiginous) (Hayward, Wilmott et al. 2017).

The process from sequencing to files able to be interpreted by a non-bioinformatician is long and expensive. For a single matched (tumour and germline) DNA sample undergoing WGS, the cost from eluted DNA to output data in 2018 was in excess of \$10000. And beyond output files, significant bioinformatic interpretation and formatting for presentation is required.

The pipeline consists broadly of variant calling from aligned reads, establishing the effect of variants, determining copy number (ploidy) of individual spans of reads, and interrogating break points to determine structural variation.
Seave, a bespoke Gemini-based program for searching and collating genomic variation across all forms had been previously used by co-workers to investigate the mutational landscape of lung cancer (Gayevskiy, Roscioli et al. 2019). This set of applications runs in-series analysis across multiple bioinformatic file types and allows the user to dictate not just what type of variant is being sought, but also the parameters of confidence and incidence required for reporting.

2.5.1 Sequencing outputs

Sequencing provides short reads (up to 200 bases long) as an output FASTA file (Pearson, Lipman 1988). When the FASTA also includes quality data in the form of a Phred score, it is termed FASTQ. This is the usual output form for subsequent alignment to the reference genome.

Following alignment to the reference genome (in our case GRCh37/hg19) files are in 2 forms, SAM (Sequence Alignment Map) (Li, Handsaker et al. 2009) and BAM (the binary form of SAM). Both file types include sequencing data aligned to the genome from which can be made an index file to tag the relative positions of sequence within the entire genome. BAM is the input used for the Integrated Genome Viewer (IGV) (Broad Institute). IGV is the basis for confirming a mutation or anomaly seen on some other platform. It allows for assessment of the entire genome and can also enable very localized assessment and can review reads/point mutations and even breakpoint areas. To navigate around IGV using BAM files, the Index file (.bai) must be loaded concurrent with the .bam. IGV is not a discovery tool, but rather a means by which outputs from bioinformatics analysis can be checked for veracity.
Sequenza (Favero, Joshi et al. 2015) was employed to calculate cellularity and copy number variation, a process dependant on adequate read depth in the region of interest. Less confidence in the output, as always, is the result of analysis of regions with poor coverage.

*Manta* was the principle program utilized for structural variant analysis. Like Sequenza, it uses BAM files to look for breakpoints, and identifies insertions, deletions, inversions and translocations. SV analysis was further interrogated by using Mobile Element Locator Tool (MELT) (Gardner, Lam et al 2017). This tool enables the identification of somatic breakpoints built upon local sequence identifiers and the degree of confidence of the call for Mobile Element Insertions (MEI).

Maftools (Mayakonda, Koeffler 2016) was used to derive Mutation annotated format (maf) files from Variant Calling Format files (see below). Thereafter, visual representation of short variants, and copy number variants were delivered in oncoplot form.

Circos plots were generated to display an overall genomic variant profile, including structural variants using Purple. PURPLE is a purity ploidy estimator. It uses the read depth and tumour BAF to estimate the purity of a sample and generate a copy number profile. ([https://github.com/hartwigmedical/hmftools/tree/master/purity-ploidy-estimator](https://github.com/hartwigmedical/hmftools/tree/master/purity-ploidy-estimator)). PURPLE is one of a number of tools that can be utilised to provide relatable information as a function of copy number and sample purity. The 5 basic steps in the computation of values form PURPLE include (Preistley, Baber et al 2017)

1. Calculate BAF in tumour at high confidence heterozygous germline loci
2. Determine read depth ratios for tumour and reference genomes
3. Segmentation – the division of the genome into uniform Copy Number (CN) regions, within the limits of breakpoints and chromosomes
4. Purity Fitting – relates ploidy and BAF to determine reliability of CN calls
5. Smoothing – reduces outlying small variances in CN to related regions not including those broken by segmentation

Diagrammatic representation of bioinformatic outputs were conceived in collaboration with and refined by Bioinformatician Dr M Gauthier.

2.6 Scoring Variant Effects

Variant calling, and subsequent variant effect prediction is the process whereby the aligned reads are subtracted from the germline, within ascribed confidence limits to identify somatic variants. Strelka was used as the variant caller (Saunders, Wong et al 2012). The output of Strelka is a .vcf file (variant calling format). The effect of the variant can be measured (or scored) by algorithms, including SIFT, Polyphen and CADD, to provide evidence regarding the likely impact of variants to help understand the biological effect.

2.6.1 SIFT (Sorting Intolerant From Tolerant)

SIFT (Ng, Henikoff 2001) predicts the likely impact of an amino acid substitution due to a somatic variant in the coding genome. It uses the genomic information of a missense variant as input. SIFT scores range from 0-1 and a score less than 0.05 means the amino acid change is likely to be damaging, based on probability. Between 0.05 and 0.1, the biological effect of the change in protein function due to the amino acid change is judged to be possibly damaging.
2.6.2 PolyPhen

PolyPhen (Adzhubei, Schmidt et al 2010) uses a similar mechanism to predict the effect of missense variants in the coding genome. It uses protein information to arrive at an assessment of biological effect. The possible outputs are probably damaging, possibly damaging and benign, all given with a confidence score, where closer to 1.0 is the highest level of confidence.

2.6.3 CADD (Combined Annotation Dependent Depletion)

CADD (Kirsher, Witten et al 2014) is a more recent addition to the predictive tools for genomic variants. CADD combines other scoring systems (including SIFT and PolyPhen) with integration with both large genomic datasets (such as 1000 Genome) to cover both coding and non-coding regions. CADD scores imply that for a score of 10, the variant is in the most damaging 10% of variants, and a score of 20 means the variant is in the highest 1% of damaging variants. Both raw and scaled CADD scores may be derived. Raw CADD scores are best to compare overall effect between 2 groups, eg control and study, whereas scaled scores are more useful for looking between small groups or individuals.

Without translated data, these numerical interpretations of implied biological effect are not entirely accurate as identified in a recent review of this topic (Misoge, Field et al 2015). Nonetheless, these parameters serve as an independent measure of risk of biological impact.

For each significant short variant reported, IGV was used to confirm the coverage in the region and the true nature of the variant in a given sample to trust the finding. Some GC rich areas seen in our samples had problems with coverage due to the challenges of genome assembly on the Illumina platform. Because coverage effects the reporting of variant allele
fractions and because coverage can be affected by such bias, high variant areas, or areas that we would expect to show variation but didn’t need to be reviewed for each sample to establish the reliability of the individual call. This is a manual process and took some time to learn and subsequently complete.

2.7 Gene Expression Analyses

Gene expression was explored using the NanoString nCounter Sprint system using the 770 gene PanCancer Progression panel with 25 ng of RNA extracted as above from fresh frozen tumours as per the manufacturer's instructions (NanoString Technologies). Nanostring uses a hybrid probe to allow RNA in solution to be identified with a capture probe and subsequently reported by a reporter probe. This highly automated process delivers molecule counts for each gene of interest as a measure of expression. The molecule counts are compared and calibrated to that of housekeeping genes. Results were analysed using NanoString nSolver 4.0 and Advanced Analysis Module, which normalizes gene expression to a set of positive and negative controls genes built into the platform. The housekeeping genes selected for Content Normalization are selected based on low level of overall change in reads across the samples and also represent low, medium, and high expression levels. Differential expression of key gene pathways was compared between specimens. The parameters for constructing heatmaps were derived in NSolver Basic Analysis program. These included background subtraction of the geometric mean of negative controls and including only those housekeeping genes with an average numerical count of greater than 50. Variation in expression between samples for given genes was expressed across Z-scores (how many standard deviations below or above the population mean a raw score is).
3 Clinicopathological characteristics and broad genomic characterisation

3.1 Tumor cellularity and clinicopathological parameters

As outline in Methods, funding allowed for 20 matched pairs of blood and tumour to undergo WGS. Despite review of case notes and pathological specimens, 1 of the 20 cases identified, processed and sequenced was ultimately found to have be inaccurately categorised, and was actually a large primary tumour which had no nodal spread following neck dissection. This left us with 19 sequenced blood; tumour pairs. Initial bioinformatic analysis using Sequenza to derive cellularity (as part of the overall bioinformatic pipeline “Refynr”) revealed significant variance between the histopathological yield (purity) and the bioinformatic (Table 3.1) Furthermore, despite attention to quality control, ultimately 4 of the 19 specimens were not useful for somatic variant analysis due to poor cellularity as judged by Purple. The discrepancy between the histopathological yield, that derived by Sequenza and that derived by Purple was the subject of much analysis.

So far as the difference between the yield determined by the Histopathologist and Sequenza, the specimen used for DNA extraction can never be the same as that used by the Pathologist for cellularity estimation, they can at best be adjacent, which invites microscopic variance between the two. As a result, the potential exists for a specimen used for DNA extraction to have less viable tumour, more stroma or even non-cellular areas of necrosis.

Strategies for decreasing the likelihood of this in future experiments of this kind could include using a block tissue as samples for cellularity estimation on either side of the block of tissue used for nucleic acid extraction, or by using a SNP array to identify suitability of the
tumour tissue for downstream processing. The SNP array would incur further processing costs, but perhaps only 10% of the costs associated with WGS.

Table 3.1: Variance between histopathological cellularity and that defined by Sequenza

<table>
<thead>
<tr>
<th>Specimen</th>
<th>Pathology cellularity</th>
<th>Sequenza cellularity</th>
<th>Specimen</th>
<th>Pathology cellularity</th>
<th>Sequenza cellularity</th>
</tr>
</thead>
<tbody>
<tr>
<td>184136</td>
<td>35%</td>
<td>64%</td>
<td>35649</td>
<td>70%</td>
<td>46%</td>
</tr>
<tr>
<td>321773</td>
<td>30%</td>
<td>33%</td>
<td>35818</td>
<td>80%</td>
<td>69%</td>
</tr>
<tr>
<td>193958</td>
<td>70%</td>
<td>34%</td>
<td>37130</td>
<td>60%</td>
<td>82%</td>
</tr>
<tr>
<td>48585</td>
<td>70%</td>
<td>28%</td>
<td>38532</td>
<td>80%</td>
<td>24%</td>
</tr>
<tr>
<td>184577</td>
<td>95%</td>
<td>60%</td>
<td>35562</td>
<td>90%</td>
<td>46%</td>
</tr>
<tr>
<td>4447</td>
<td>60%</td>
<td>59%</td>
<td>658492</td>
<td>*</td>
<td>32%</td>
</tr>
<tr>
<td>4699</td>
<td>70%</td>
<td>52%</td>
<td>33432</td>
<td>90%</td>
<td>70%</td>
</tr>
<tr>
<td>9120</td>
<td>70%</td>
<td>27%</td>
<td>34366</td>
<td>70%</td>
<td>26%</td>
</tr>
<tr>
<td>200971</td>
<td>80%</td>
<td>34%</td>
<td>183410</td>
<td>50%</td>
<td>35%</td>
</tr>
<tr>
<td>34934</td>
<td>60%</td>
<td>73%</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Furthermore, in samples where the true cellularity was very low (in the order of 10-20%) Sequenza tends to overcall variation. In these cases, Purple was a more reliable determinant of the true cellularity, ultimately exhibited by almost no somatic variation, either as short variants or CNV.

The 4 cases that were judged by Purple to be of inadequate cellularity had all passed quality control measures as part of the routine initial assessment of cases worthy of inclusion in that tissue quality and tumour cellularity within the specimen was assessed as adequate by a Specialist Histopathologist highly experienced in cutaneous malignancy. Nonetheless, ultimately it became clear that the cellularity as determined by the bioinformatics algorithms
was significantly different to the clinical and histopathological assessment, and these tumours were judged as having cellularity that was too low (less than 20%) to allow for calling of low variant allele mutations, and therefore could not be relied upon to faithfully represent the true mutational status of the original tumour. The clinicopathologic data of these 15 cases is summarised in Table 3.2.

**Table 3.2. Clinicopathologic data of 15 specimens cleared for bioinformatic analysis.**

Staging according to AJCC 8th edition: (Grade 1: Well differentiated; Grade 2: Moderately differentiated; Grade 3. Poorly differentiated. Tumour mutational burden is genomic. Units mutations/Megabase.)

<table>
<thead>
<tr>
<th>Patient</th>
<th>Age (years)</th>
<th>Sex</th>
<th>Primary location</th>
<th>Metastasis location</th>
<th>Nodal stage</th>
<th>Grade</th>
<th>Tumour mutational burden (variants/Mb)</th>
</tr>
</thead>
<tbody>
<tr>
<td>183410</td>
<td>30</td>
<td>male</td>
<td>left lip</td>
<td>left neck</td>
<td>N3b</td>
<td>1</td>
<td>468.8</td>
</tr>
<tr>
<td>184577</td>
<td>78</td>
<td>male</td>
<td>right ear</td>
<td>right parotid</td>
<td>N3b</td>
<td>3</td>
<td>261.8</td>
</tr>
<tr>
<td>193958</td>
<td>74</td>
<td>male</td>
<td>unknown</td>
<td>right parotid</td>
<td>N3b</td>
<td>3</td>
<td>169.2</td>
</tr>
<tr>
<td>200971</td>
<td>64</td>
<td>male</td>
<td>bilateral lip</td>
<td>bilateral neck</td>
<td>N2c</td>
<td>2</td>
<td>49.3</td>
</tr>
<tr>
<td>321773</td>
<td>78</td>
<td>male</td>
<td>left forehead</td>
<td>left parotid</td>
<td>N2a</td>
<td>3</td>
<td>173.7</td>
</tr>
<tr>
<td>33432</td>
<td>69</td>
<td>male</td>
<td>left cheek</td>
<td>left neck</td>
<td>N3b</td>
<td>3</td>
<td>50.6</td>
</tr>
<tr>
<td>34366</td>
<td>87</td>
<td>male</td>
<td>unknown</td>
<td>left neck</td>
<td>N2b</td>
<td>3</td>
<td>202.1</td>
</tr>
<tr>
<td>34934</td>
<td>87</td>
<td>male</td>
<td>unknown</td>
<td>right parotid</td>
<td>N3b</td>
<td>2</td>
<td>44.8</td>
</tr>
<tr>
<td>35562</td>
<td>66</td>
<td>male</td>
<td>Bilateral forehead</td>
<td>right neck</td>
<td>N3b</td>
<td>2</td>
<td>122.1</td>
</tr>
<tr>
<td>35649</td>
<td>64</td>
<td>male</td>
<td>Left scalp</td>
<td>left neck</td>
<td>N3b</td>
<td>3</td>
<td>151.3</td>
</tr>
<tr>
<td>35818</td>
<td>69</td>
<td>male</td>
<td>unknown</td>
<td>right parotid</td>
<td>N3b</td>
<td>3</td>
<td>406</td>
</tr>
<tr>
<td>38532</td>
<td>77</td>
<td>male</td>
<td>Right nose</td>
<td>right neck</td>
<td>N3b</td>
<td>2</td>
<td>75.3</td>
</tr>
<tr>
<td>4699</td>
<td>77</td>
<td>male</td>
<td>Right pinna</td>
<td>right parotid</td>
<td>N3b</td>
<td>2</td>
<td>159.8</td>
</tr>
<tr>
<td>48585</td>
<td>79</td>
<td>female</td>
<td>Left cheek</td>
<td>left perifacial</td>
<td>N3b</td>
<td>3</td>
<td>426.9</td>
</tr>
<tr>
<td>9120</td>
<td>66</td>
<td>male</td>
<td>Left scalp</td>
<td>Left scalp</td>
<td>N2b</td>
<td>2</td>
<td>355</td>
</tr>
</tbody>
</table>
One sample initially designated as metastatic was found at a later time to have been from a primary cSCC with no evidence of metastasis. This specimen, and its matched whole blood, underwent WGS to 60X/45X prior to its status being understood. As such, it was not included in this analysis, but the variant calls and all data from this analysis will be used by a collaborator who is mirroring this overall study design, but in the primary cSCC setting.

### 3.2 Sequencing coverage

Of the 15 samples that passed quality control, sequencing coverage is presented in Table 3.3. Overall tumour coverage was 78.5X and germline was 34.4X. Coverage and tumour cellularity combines to enable confidence in variant calling, and the ability to also identify low VAF variants.

#### Table 3.3 Sequencing coverage for each sample tumour and blood(germline).

<table>
<thead>
<tr>
<th>Sample</th>
<th>Tumour coverage</th>
<th>Blood coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>184310</td>
<td>90.8</td>
<td>27.3</td>
</tr>
<tr>
<td>184577</td>
<td>65.2</td>
<td>42.7</td>
</tr>
<tr>
<td>193958</td>
<td>64.9</td>
<td>42.3</td>
</tr>
<tr>
<td>200971</td>
<td>65.6</td>
<td>49.2</td>
</tr>
<tr>
<td>321773</td>
<td>67.8</td>
<td>38.6</td>
</tr>
<tr>
<td>33432</td>
<td>85.4</td>
<td>35.5</td>
</tr>
<tr>
<td>34366</td>
<td>104.3</td>
<td>36.5</td>
</tr>
<tr>
<td>34934</td>
<td>109.4</td>
<td>33.4</td>
</tr>
<tr>
<td>35562</td>
<td>96.5</td>
<td>27.1</td>
</tr>
<tr>
<td>35649</td>
<td>103.9</td>
<td>30.9</td>
</tr>
<tr>
<td>35818</td>
<td>105.2</td>
<td>35.1</td>
</tr>
<tr>
<td>38532</td>
<td>96.8</td>
<td>37.4</td>
</tr>
<tr>
<td>4699</td>
<td>59.1</td>
<td>41.9</td>
</tr>
<tr>
<td>48585</td>
<td>63.3</td>
<td>38.6</td>
</tr>
<tr>
<td>9120</td>
<td>69.8</td>
<td>32.4</td>
</tr>
<tr>
<td><strong>Average</strong></td>
<td><strong>78.3</strong></td>
<td><strong>34.3</strong></td>
</tr>
</tbody>
</table>

### 3.3 Mutational burden and signatures

Across the 15 cases with matched metastatic cSCC and matched whole blood for germline, a striking incidence of mutational burden was observed (Figure 3-2) with an average mutation load of 207.8 mutations/Mb. This is greater than previously published data from primary
cSCC where the mutational burden was 45.2 mutations/Mb (Chalmers, Connelly et al. 2017).

No mutational burden data for a cohort of metastatic cSCC has previously been published.

Mutational burden of our cohort compared to previously published data is presented in figure 3.1 and is now also published (Mueller et al 2019).

The scale of the variants identified within this cohort is significant. In total, WGS identified 14681149 variants. Of these, 5801 were classified High Impact (including Stop gained, Splice site, Frame shift or Stop lost) on the basis of either functional or algorithmic (CADD, SIFT, Polyphen) implication. A further 74612 were classified as being of medium impact. Of
these, 45067 had a CADD > 15. A CADD score of 20 means the variant is within the top 1% of deleterious variants in the entire genome, a CADD of 30 put the variant in the top 0.1% of the most deleterious variants (Kircher, Witten et al. 2014). All high and medium impacts were in the coding genome, as non-coding variants are classified low impact on the established scoring systems.

For each specimen, it was possible to report on the percentage of nucleotide variants that best fit one of the 30 COSMIC mutational signatures (Alexandrov, Nik-Zainal et al. 2013) (Figure 3-2). Cosmic signatures are derived from major catalogues of cancer associated exome and genome level sequencing. Where tumours are predicted to have a clear aetiological relationship with some known or unknown factor, they are grouped on the basis of predominant nucleotide variants. For some of the mutational signatures, the aetiological agent is known and ascribed, for other groups of tumours with a recurrent pattern of mutations, the agent is unknown. For instance, Signature 4, seen in tumours of the head and neck (mucosal), liver, lung and oesophagus, which is characterized by C>A mutations, and CC>AA dinucleotide substitutions has tobacco mutagens as its proposed aetiological agent. Signature 7, seen in skin cancers, is characterized by C>T, and CC>TT dinucleotide substitution, and is ascribed ultraviolet radiation as its proposed aetiology. The predominant mutation signature in our disease cohort was Signature 7 (Figure 3-2). As predictable as this may sound, this finding has not been reported in metastatic cSCC and confirms the metastatic process does not overly influence the underlying likely mechanism and pattern of nucleotide variation. Compared to other tumour groups, the percentage of mutations showing concordance to the underlying signature is striking. On average, greater than 65% of variants seen in our cohort were C>T, or CC>TT. The fit with Signature 7 is somewhat clouded by the similarity between it and Signature 11, ascribed to the exposure to alkylating agents (first described in
exposure to temozolomide). Temozolomide acts by its ability to alkylate or methylate DNA, particularly on guanine residues. This effect triggers apoptosis.

Figure 3-2 Mutation density and signature analysis across the cohort of 15 cSCC lymph node metastases. (Top) Base change mutation distribution at single base level shows predominance of C→T transitions. (Middle) Boxplot showing median number of mutations per megabase (Mb) in the coding and non-coding DNA. Mutation burden per patient in coding and non-coding DNA. (Lower) Signature profiles using the updated signature repertoire by Alexandrov et al. (Alexandrov et al., 2018).
3.4 Discussion

Tumour mutational burden is a complex and potentially confusing descriptor. Traditionally, TMB has been described as the rate of mutations/Mb of the coding regions, either as a percentage of the total coding region, or as a percentage of the entire genome. This rule is not universally followed. The melanoma genomic analysis paper (Hayward, Wilmott et al 2017) cites TMB for various subtypes of melanoma. The authors used TMB in a genomic context, i.e. looking at mutation across the entire genome. This is different to most previous TMB papers, although truly is a measure of genomic instability. In our paper describing TMB in metastatic cSCC (Mueller, Lauthier et al 2019), TMB is reported as both non-coding (non-coding mutations/Mb of the entire genome) and coding (coding mutations/Mb of the entire genome) to illustrate the predominance of non-coding variants within our samples. In this paper, we report coding TMB as 1.2 mutations/Mb and non-coding TMB as 206.6 mutations/Mb. If we report the same dataset as TMB of the coding region with the total Mb of the coding region as the denominator, the result is 43.0 mutations/Mb. Similarly, if we report non-coding TMB with the non-coding genome as the denominator, the result is 212.5 mutation/Mb. It is therefore important to compare like with like, and to establish the parameters of TMB prior to drawing conclusions.

Whilst we suspected the genomic mutational burden in metastatic cSCC (total mutations of any type across the entire genome) would be high, at an average of 207.8 mutations/Mb, this figure vastly eclipses the rates of any other malignancy, including metastatic melanoma and primary cSCC. UV associated melanoma and primary cSCC had the highest rate of mutations/Mb described. In a study outlining the genomic landscape of subtypes of melanoma, sun exposed skin occurring melanoma had an overall tumour burden averaging 49 mutations/Mb. Non UV associated tumours (uveal and acral) showed far lower rates of
overall mutational burden (Hayward, Wilmott et al. 2017). Rates for a mixed cohort of cSCC (including 7 cases of metastatic cSCC) were 61.2 mutations/Mb (Pickering, Zhou et al. 2014).

The UV associated signature of pyrimidine substitution (C>T) was strikingly consistent throughout our samples. This confirms the penetration of the UV associated genomic base substitution through EMT, with subsequent genotype of the metastatic clone/s closely mirroring the predicted effect.

We assume that the increased mutational load in metastatic cSCC is due to the ongoing and prolonged burden of UV associated damage. However, it is clear from analysis of microsatellite instability (MSI) in non-polyposis colorectal cancer, that mutation rates are higher in so called MSI-High (MSI-H) tumours (Pawlik 2004). These cancers in HNPCC are the result of germline mutations in mismatch repair genes, in particular MSH2, MLH1, MSH6, PMS2, and PMS1.

The unprecedented level of mutational burden seen in our tumours makes the case for the use of checkpoint inhibitors as probable therapeutic agents in this disease. Whilst the consideration of therapies is well beyond the scope of this project, the utility of any therapy needs to be based on scientific observation, and it is established that tumours with a greater mutational burden should respond with greater effect to checkpoint inhibition (Yarchoan, Hopkins et al. 2017). The authors describe response rates to checkpoint inhibition as a function of coding (exomic) TMB. According to the formula established by Yarchoan et al for assessing predicted response rate to PD-1 inhibition, Response rate = 10.8 x logₑ(X) - 0.7, where “X” is the mutational burden (mutations/Mb), our cohort should have a 39.6% response rate. The results of the phase 2 EMPOWER CSCC-1 trial of a monoclonal antibody
to PD-1 Cemiplimab which showed that about half of patients responded to therapy with a mean follow-up of 7.9 months, have lead to the release of this drug in the US in the advanced cSCC setting, for both metastatic and locally advanced disease (Midgen, Rischin et al 2018). Our mutational burden findings support the use of PD-1 inhibitors in metastatic cSCC.
4 Major structural variation including copy number variation

4.1 Overview

Structural variation (SV) covers major chromosomal events including inversions, translocations, large deletions and amplifications. Visual representation of SV events can be figuratively exemplified in Circos plot format - a means by which multiple data points and types can be represented on a single figure able to recreate the full, in this case, genomic picture. Each circus plot is built upon data derived from variant and structural analysis through the bioinformatic pipeline. Each circus plot thus also contains information relating to short variants including allele frequency, copy number variation as well as major structural variants. Such a format allows for a comparison within a cohort of the extent of genomic variation, and for each sample, can guide areas of the genome for interrogation with respect to a given category of variation. After exploring a number of formats to represent our structural variant data in circos, Purple, a bioinformatics program that estimates purity and copy number by using read depth and tumour variant allele frequency, was chosen. This allowed for the most accurate visualisation of all the elements deemed useful.

4.2 Major structural variation

4.2.1 Results

As outlined above Circos plots provide an overall impression of the somatic genome of an individual sample, in comparing between samples, or identifying congruence between versions (tumour, cell line, passages) of a single sample. An annotated example of a circos
plot derived by Purple is presented in Figure 4-1. Circos plots of all 15 specimens, with a clinicopathologic summary are presented below.

Figure 4-1 Circos plot sample 4699. Cellularity 52%

The utility of the circos plot is to enable an overview of all categories of variants within a sample’s genome. Sample 4699 is from a 78 year old male who had a right pinna cSCC treated with cryotherapy only and developed ipsilateral nodal disease. The patient eventually developed contralateral neck disease from a forehead cSCC moderately differentiated with no LVI or PNI, treated with surgery and adjuvant RT. He eventually succumbed to soft tissue recurrence within the RT field and developed lung metastasis (DOD).
The overwhelming majority of the second shell (SNVs) of the circos shows C>T (red). When compared to Figure 5-1, detailing Indels, shell 3 in the circos appears mostly insertions. The contrast between these colours could be improved to allow for greater resolution. CNV is demonstrated in the 4th shell and is able to be constrained by block size. The minor allele CN in shell 5 here shows, for example, amplification of both major and minor allele in 5p. The proximity of 4q and 5p are useful for displaying a copy number gain despite the loss of the minor allele(4q) and a copy number gain of both major and minor allele (5p). In 8p, a CN neutral LOH is seen due to loss of the minor allele. The inner shell herein shows both inversions and translocations. Particular concentration is seen in the proximal long arm of 15 and in 8q.
Figure 4-2: Circos plot of sample 9120. Cellularity 27%.

Sample 9120 is from a 66 year old male. The patient initially had an excision of recurrent scalp tumour of 8.5mm thickness, moderately differentiated with clear margins and no LVI or PNI. The patient developed metastases in the draining posterior scalp lymph nodes and posterior triangle 12 months following the initial excision. Currently the patient is alive with no evidence of disease 2 years post lymphadenectomy.

Striking within this circos plot is the deletion of chr3p, chr4, chr11 and chr18 and loss of copy number. There was an oscillating CN artefact identified during SV analysis. This artefact resulted in, within established parameters as above, an excessive CNV count which
was not consistent with the actual CN and can be seen in the difference between Figures 4-1 and 4-2 in the green track.

Figure 4-3: Circos plot sample 33432. Cellularity 70%.
Sample 33432 is from a 69 year old male with a background of rheumatoid arthritis treated with pharmacologic immunosuppression (Azathioprine). The patient in initially had a temporal scalp SCC. He underwent surgery and adjuvant RT and then had recurrence of poorly differentiated metastatic cSCC. This patient died of disease. The lack of significant CNV and SV in this sample is striking.
Figure 4-4: Circos plot of sample 34366. Cellularity 26%.

Sample 34366 is from an 86 year old male. Metastatic cSCC to parotid treated with surgery and then adjuvant RT. Thereafter recurred in soft tissues within radiated field. Died of disease.

Prominent loss of minor allele copy is demonstrated across much of the genome in this sample, as well as areas of very high overall CN gain (chr2q and chr18p). Inversions are more common than translocations in this sample.
Figure 4-5: Circos plot sample 34934. Cellularity 73%.

Sample 34934 is from an 87 year old male who previously had a maxillary mucosal SCC. Subsequently he developed a scalp cSCC and left parotid metastasis. Surgery revealed a single metastatic deposit. He underwent post operative radiotherapy and was alive with no evidence of disease 2 years later.

Once again, despite being high cellularity there is relatively less amplification in this sample. CN loss is once again the predominant CNV despite some area of amplification (chr14). This sample has a very low indel incidence (shell 3).
Figure 4-6: Circos plot sample 35562. Cellularity 46%.

Sample 35562 is from a 66-year-old male, 15 years post liver transplant on tacrolimus and prednisone for immunosuppression. He initially had excision of a 20mm forehead skin lesion (moderately differentiated cSCC 7mm thickness, with both LVI and PNI) with right neck dissection. Neck dissection 2/29 lymph nodes in parotidectomy and neck dissection. Post operative radiotherapy. Alive with no evidence of disease at 2 years.

Whether the formal immunosuppression has impacted the CNV in this case is not clear. Apart from chr8q, there are not large scale amplifications, but areas of CN loss and LOH.
Sample 35649 was from a 63 year old male with chronic lymphocytic leukaemia. Recurrent scalp cSCC with metastases to parotid and neck. Primary mod differentiated 8.5mm thick with LVI but not PNI. Post operative radiotherapy with early recurrence. Died of disease. Again this sample, although having more areas of amplification, including with very high CN in an isolated area of chr4p, is more characterised by CN loss with LOH. A focus of translocation is in chr19.
Sample 35818 is from a 69 year old male. Underwent a right radical parotidectomy (including sacrifice of facial nerve and reconstruction) and then had post operative radiotherapy. Late recurrence ipsilateral lymph node level Ia/b junction. Clinically very high risk for recurrence but alive with no evidence of disease.

This circos is more typical of the trend with widespread CN gain with some areas of CN neutral loss of minor allele (chr8p, chr5q). Most of the areas of CN gain have contribution to overall CN by amplification of the minor allele (blue in shell 5). Note also the concentration of inversion in chr1q and chr5p.
Figure 4-9 Circos plot sample 38532. Cellularity 24%.

Sample 38532 is from a 78 year old male with Protein S deficiency. Nasal cSCC with prominent PNI. Metastasis to left parotid and eventually right neck. Post op radiotherapy.

Alive with evidence of disease 1 year post op.

Isolated regions of amplification can be seen with total CN>10 in chr3 and 7. Overall limited amplification across the genome. Prominent CN loss with LOH in chr3, 8, 12 and 15.

Amplified area of chr3 also corresponds to a concentration of both inversion and translocation.
Figure 4-10 Circos plot sample 48585. Cellularity 28%.

Sample 48585 is from a 78 year old female (note sex chromosomes) with multiple recurrent cSCC bilateral forehead and face. She had an acceleration of cSCC over the last 5 years of her life without formal immunosuppression. This patient developed bilateral neck node metastases from different primaries. This invites the potential for her having a predisposing germline or acquired immunodeficiency. She eventually succumb to left sided recurrence at skull base post salvage surgery and radiotherapy.

This is a heavily amplified sample. Some of the amplified regions have minor allele loss. A focus of translocation in chr22 is typical of that seen with the TTC28 transposon (see below).
Figure 4-11 Circos plot sample 183410. Cellularity 35%.

Sample 183410 if from a 30 year old male with limited UV exposure history. Left lip cSCC excised with emergence of ipsilateral Level Ib metastasis. Post operative chemoradiotherapy. He developed tinnitus with cisplatin and was switched to cetuximab. Alive with no evidence of disease 2 years post treatment.

This sample had an average ploidy of 4. This is a heavily amplified genome, with only very few areas of minor allele loss. This sample has the highest incidence of CN >10 (4th shell
with circle on top of green CN gain). Additionally, significant SV with again a focus of translocation in chr22.

**Figure 4-12 Circos plot sample 184577. Cellularity 60%.

Sample 184577 is from a 78 year old male. Right ear cSCC to ipsilateral parotid. Nodal deposit 45mm poorly differentiated with positive margin on facial nerve. Post op RT. Alive with no evidence of disease beyond 3 years.

This is a similar plot to 183410 except for the extent of translocation. Some areas of LOH (chr10q and 13). Again some concentration of SV in chr22 (*TTC28*) is observed.
Figure 4-13 Circos plot sample 193958. Cellularity 34%.

Sample 193958 (Parent tumour of cell line UW01- Jay Perry) is from a 74 year old male. cSCC from right ear excised. Eight months later emergence of parotid metastasis. 50mm deposit with PNI and 2 other positive nodes. Clear margins. Post op radiotherapy. Alive with no evidence of disease 3 years post treatment.

This circos shows significant amplification across most of the genome. There is widespread loss of the minor allele, often occurring as CN neutral, but often with an overall CN gain. A focus of SV (both inversion and translocation) can be seen in the short arm of chr7.
Figure 4-14 Circos plot sample 200971. Cellularity 34%.

Sample 200971 was from a 65 year old male who initially had a lower lip cSCC (6mm moderately differentiated) with metastasis to level Ib at 14 months post surgery to the primary. Moderately differentiated disease. Post op radiotherapy. Succumbed to pulmonary metastases within 3 years.

This sample shows isolated regions of CN gain with even more isolated regions of loss of minor allele.
Figure 4-15 Circos plot sample 321773. Cellularity 33%.

Sample 321773 was from a 78 year old male who initially had a Left forehead cSCC with widespread PNI. Post op radiotherapy to the primary site but not to draining nodal basins. Metastasis to left parotid 19 months post surgery. Parotidectomy and neck dissection with adjuvant radiotherapy to left neck and parotid bed. Alive with no evidence of disease at 3 years.

Loss of minor allele and CN loss entire chr13. Otherwise a significantly amplified genome. SV including chr22 (TTC28) and complex pattern of SV in chr8.
4.2.2 Discussion

Just as for short variants, the identification and characterization of structural variants in a highly mutated genome is challenging. As stated earlier, the mutational burden of our samples is greater than any previously published series, at 207.8 mutations/Mb. Complex bioinformatics and computational analysis is required to overcome this genomic noise, and to filter spurious findings.

Structural variants may be between non-coding or coding regions. They may involve large scale deletions or duplications, or breakpoints which then reconnect with a remote part of the genome, either intrachromosomal (inversion), or interchromosomal (translocation). Large deletions may include areas coding or impacting on transcription of tumour suppressor genes, thus promoting a carcinogenic stimulus. In addition, and often as well, large duplications (usually with overall CN >5 and block size >5000 base pairs) may include regions coding for, or impacting the transcription of oncogenes, once again with potential for malignant genomic effect.

In addition to simple descriptions of SV events (inversion, translocation, deletion, duplication), more complex patterns of structural rearrangement have been described. In a cohort of prostate cancers, Baca et al (2013) describe a pattern of translocation associated with deletion breakpoints, but also giving rise to observed deletion bridges from one chromosome to another. They termed this phenomenon “chromplexy” (pleko : “to weave” or “to braid”). Chromplexy is thought to occur throughout the progression of a cancer and is characterized by tens of chromosomal structural variants, effecting different loci and probably occurring at multiple timepoints. Such large scale chromosomal rearrangement is relatively common, particularly in prostate cancer (Shen 2013). This is in contrast to
chromothripsis. Chromothripsis entails a very localized disruption of the genome, often entailing hundreds of structural variant events, and is thought to occur as a once off phenomenon (Stephens 2011). Such a massive genomic crisis event gives rise to oscillating copy number states between one or only a few chromosomes and may include cancer causing fusions.

Within our structural variant analysis, there are no convincing episodes that can be easily labelled chromothripsis. This observation may be clouded by the volume of SV events, and it is possible that both chromoplexy and chromothripsis are occurring in the same samples, both inferring a tumour clonal evolution by different means and chronologies.

### 4.3 Gene fusions

#### 4.3.1 Results

Break points within coding regions of the genome can lead to fusions with other genes or non-coding regions. Such coding impacts are called gene fusions and may have unpredictable effects on the transcribed RNA. We identified 2004 gene fusion events structural variants having any gene involvement (gene:intron or gene:gene) across our 15 cases. Break points (including those leading to gene fusions) occurred in more than 1 sample in only 5 genes as listed below. By far the most frequently involved gene was *TTC28* which showed significant structural variation in the first intron.
Ten of fifteen samples (66%) had a SV in TTC28. In our series, the TTC28 first intronic transposon (between coordinates 22:29064630-29066160) is inserted into a number of genes as depicted in Figure 4-16.

Additionally, the following genes had break points in their coding regions: **MYLK** (2/15 samples); **PDE4D** (3/15 samples); **EPHB1** (2/15 samples). **PTK2:SLA2** gene fusion occurred in 1/15 samples.
4.3.2 Discussion

*TTC28* (chr22:28,374,002-29,075,853) is described having recurrent chromosomal translocations in colorectal, small cell lung and liver cancer (Fujimoto et al 2016). The pattern of the structural variants we have observed involving *TTC28* is most in keeping with a phenomenon known as somatic retrotransposition. The resulting genomic feature is an example of a retrotransposon in the form of a Long Interspersed Element 1 (LINE-1 or L-1). A known retrotransposon is within the first intron of *TTC28*. Retrotransposon insertion may also happen in the germline.

It is unusual for the same region in a chromosome in the same sample to be widely dispersed in a structural variant or translocation event. The model of a LINE-1 dispersion to “random” locations, even within one sample was identified and characterized during the assessment of a new bioinformatics application (MELT) looking at data from 1000 Genomes Project (Gardner, Lam et al 2017). This tool identifies and characterises LINE-1, Short interspersed nuclear elements (SINEs) and other major and structural variant events caused by mobile elements. The use of MELT with our data supports this hypothesis of LINE-1 transposon “rearrangement”.

In our study, the only gene to show a fusion with *TTC28* is *ESRRG*. *ESRRG* has recently been identified as playing a role as a tumour suppressor gene in gastric cancer (Kang, Choi et al 2018) by its impact as a suppressor on the Wnt signaling pathway, as evidenced by the down regulation of key Wnt genes in *ESRRG* overexpressing tumours.

Otherwise there appear no key known cancer genes in our observed structural variants interacting with *TTC28*. This process of somatic retrotransposition has been questioned by
others (Pitkanen, Cajuso et al 2017) as potentially over-calling the role of such LINE-1 elements, particularly so since in the oesophageal SCC series, no demonstration of an expression change was demonstrated. Nonetheless, this pattern is repeated in numerous cancers, and it will be illustrative to review once we have completed a parallel project looking at WGS within primary cSCC, and also to examine the transcriptome affect once our RNA Seq project is complete across cSCC cohorts.

**MYLK**; Functionally MYLK has transferase activity, transferring phosphorus-containing groups and protein tyrosine kinase activity. MYLK and MYL9 were found in NSCLC to be downregulated in Stage 1 and II cancers but upregulated in III and IV cancers (Tan, Chen 2014). This suggests not only a role in carcinogenesis, but also in metastasis.

**PDE4D**; a phosphodiesterase that degrades cAMP thus disrupting activation of the cAMP pathway. Interaction with Focal Adhesion Kinase (FAK) (which we did not see as a gene fusion) may promote melanoma invasion (Delyon, Servy et al 2017) and conversely the blocking of this interaction reduces invasion.

**EPHB1**; This gene codes for a receptor tyrosine kinase. Decreased expression of EPHB1 has been identified in renal cell carcinoma (Zhou, Wang 2014)

**PTK2:SLA2** translocation occurred in only 1 sample. PTK2 is a protein tyrosine kinase which if activated is an important promoter of downstream signalling processes. SLA2 is a member of the SLAP family of proteins, that play a role in downregulating inflammatory cascades, but that are also active in cancer. Silencing of SLAP promotes tumour progression in colorectal cancer, while overexpression inhibits tumour growth and invasiveness (Marton
2015). Indeed, one of the SLA2 protein domains actively binds to phosphorylated tyrosine residues.
4.4 Copy Number Variation

4.4.1 Results

Copy Number Variation detection was principally derived by Sequenza (as described earlier). Sequenza walks through the genome in 50Mb segments to identify copy number variation and variant allele frequency and is highly dependent on cellularity. Low cellularity specimens (where tumour yield was ultimately found to be low), or areas of low coverage can both impact on reliability of CNV data.

Chromosomal amplification and deletion across the entire cohort is presented in Figure 4-17. Amplification is predominant although some areas of the genome do not show amplification, and other areas are commonly and recurrently deleted.

![Figure 4-17 All samples genome view of amplification vs deletion as percentage of samples (y axis) for each chromosomal arm (x axis). Amplification shown as blue bars; Deletions shown as red bars.](image)

An output of Sequenza includes the genome view of copy number and allele frequency. An example output of a single sample with high frequency alterations is shown in Figure 4-18. Obvious alterations in this example include a large CN gain in the long arm of Chr3 (with a
CN of 1 in the beta allele), and a CN neutral (CN = 2) loss of heterozygosity (B allele CN=0) in the long arm of Chr5.

**Figure 4-18 Genome view of CNV sample 4699.** Chromosomes are listed on the X axis. Copy number along the Y axis. The colours within each chromosome represent the total CN (red) and the CN of the minor variant allele (blue).

Figure 4-19 shows CNV of a low cellularity specimen. CNV is essentially non-existent and resembles the genome of normal (blood), and thus this specimen most likely represents mainly normal tissue DNA (only minor changes from the germline). A genome view with little departure from the normal is thus one of the clues to low cellularity.

**Figure 4-19 Genome view of CNV of low tumour cellularity sample 4777**
Figure 4-20 shows CNV for commonly effected samples with total events for each gene at the top of the figure. Amplification events were recognised if they had adequate size (base pairs > $10^4$) and with total copy number >5. Deletion events were also of the same minimum size and either had loss of both alleles or a loss of heterozygosity event with complete loss of minor allele regardless of total CN.

Below are listed common gene (Fig 4-20) and chromosomal band (Fig 4-21) CNV across as many samples as they are shared by. This is the reason not all samples are included as there is a tail of less common events which are not shared by many samples. Common genetic and chromosomal amplification is a relatively more common phenomena, as compared to commonly occurring deletion, a can be seen by only DCC and SMAD4 reaching any common threshold across samples that also shared common amplification. This is also mirrored by the overall finding of combined samples in Fig 4-17, which demonstrates the prominence of amplification in CNV analysis.
Figure 4-20 Genes with most frequent Copy Number Variation with minimum Duplication (CN>4) and Deletion (CN=1) for 10 samples with common shared CNV. Histogram at top: number of events for each sample for the group of genes. Histogram at right: number of samples (of 10) with CNV for given genes.
Figure 4-21 Recurrent CNV of chromosome bands. Histograms as per Figure 4-20. As per Figure 4-20, represented are common chromosomal band CNV with CN>4 or =1.

Specimen 9120 was unable to be included in CNV analysis due to an oscillating CN artefact identified during SV analysis. This artefact resulted in, within established parameters as above, an excessive CNV count which was not consistent with the actual CN when inspected using Integrated Genome Viewer. This is most instructively seen in Figure 4-22 below, when comparing the CN track (green) of the circus plots of samples 4699 versus 9120.
Figure 4-22 Circos plots of samples 4699 (left) and 9120 (right). A comparison of the CN track (green) and also the minor allele CN (blue/orange) track in circos plot for 9120 shows an oscillating artefact represented by smaller blocks of CNV. This disallowed 9120 to be included in the CNV analysis.

4.4.2 Discussion

Overall picture is of a highly somatically amplified genome. Gene amplification has been for some time regarded as a driver of both carcinogenesis and of clonal expansion within tumours (Albertson 2006). A recent study of amplification of driver genes across multiple cancer types identified a group of 6 genes (from a total of 138 candidate oncogenes (n=64) and tumour suppressor genes (n=74)) commonly amplified across multiple cancer types (not including cSCC) (Ohshima, Hatakeyama et al 2017); MDM2, MYC, MYCL, MYCN, NKX2-1 and SKP2.

Recurrent CN amplification of 3q and 5p were identified across various SCC sub types in a recent review comparing TCGA data and mutational signatures (Campbell, Yau et al 2018). Within 3q, which showed frequent CNA in our samples are included PIK3CA, TERC, TP63 and TP73, and SOX2 (see below). 5p includes TERT, TRIP13 and FASTKD3. The role of TERT (in the context of TERT promoter variants) in our cohort is detailed in Chapter 6.
TRIP13 can promote error prone non homologous end joining, cell proliferation, survival, and resistance to cisplatin in head and neck (mucosal) SCC. Twelve of 15 samples in our cohort had copy number gains of TRIP13 with an average CN of 4.8 across the amplified samples. FASTKD3 has prosurvival affects probably via inhibition of the intrinsic mitochondrial cytochrome-mediated cell death pathway (Simarro, Gimenez-Cassina et al., 2010) was amplified again in (the same) 12/15 samples as TRIP13, with an average CN of 4.6. With an average block size of 1900Mb, the amplifications encompassing these 2 genes were the same events. Notwithstanding the observed 5p and 3q amplification, in our cohort, there were also consistently amplified segments in 7p, 8q, 14q and 20q (Figure 4-17).

In a census of amplified and overexpressed genes in cancer, the gap between amplification and over expression is highlighted (Santarius, Shipley et al. 2010). An amplification may be within a gene, or, given a cutoff of not less than 5000 BP, more often includes multiple genes over upward of 100 000BP. Within such regions, driver genes may be identified, but co-amplified genes (eg DDX1 in the amplification of MYCN) could potentially play a role in any expression effect.

4.4.2.1 Recurrently amplified genes

The following genes were amplified in percentage of samples outlined in Figure 4-20.

**NDRG1** (N-Myc Downstream Regulated 1) chr8:134,249,414-134,314,265 encodes a cytoplasmic protein involved in stress response. It is generally regarded as a tumour suppressor, but levels of the protein are prognostically unfavourable in liver, renal and brain malignancy (proteinatlas.org).

**PIK3CA** chr3:178,865,902-178,957,881 is a key participant in cellular signalling in response to the binding of numerous ligands to receptor tyrosine kinases. Whilst hotspots for mutation
at codons 542/545 and 1047 are described (Vorkas, Poumpouridou et al 2010), we did not see any examples of this SNV. Only 2 samples had missense mutations. By far the greatest likely impact on PIK3CA was amplification, which, in keeping with its role as an oncogene with AKT and mTOR pathway impacts, offers an opportunity for potential therapeutic intervention.

**SOX2** chr3:181,429,712-181,432,224 involved in embryonic development, stem cell maintenance in the central nervous system and for expression of gastric epithelial expression. There is no clear role for its amplification being a driver in cancer, although it thought to have a role in adult stem cell differentiation (Karamboulas and Ailles 2013). With 11/15 samples showing amplification, with an average CN of 4.3, SOX2 is a good example of the 3q amplification characteristic across various subtypes of (non-cutaneous) SCC(Campbell, Yau et al. 2018)

**ABL1** chr9:133,589,268-133,763,062 encodes a protein (non-receptor) tyrosine kinase and is a proto oncogene, highly expressed in many cancers (proteinatlas.org). We observed no major structural variations in ABL1 as described in other cancers including the BCR-ABL1 fusion in chronic myeloid leukaemia (CML).

**ASXL1** chr20:30,946,147-31,027,122 encodes a chromatin-binding protein which binds and then disrupts chromatin in specific regions to enhance transcription. Its expression in endometrial cancer in a negative prognostic indicator, and conversely provides a favourable prognosis in head and neck (mucosal) SCC (proteinatlas.org) (Chung, Guthrie et al 2015).
**BAI1/ADGRB1** chr8:143,530,791-143,626,370 encodes a protein which acts as an inhibitor of angiogenesis, perhaps as a member of the secretin receptor family. Its transcriptional regulation is by p53. It has low levels of expression in cancer (proteinatlas.org).

**CAP2** chr6:17,393,447-17,558,023 probably plays a role in actin binding and ectoderm differentiation. It is not normally expressed in skin (proteinatlas.org). There is no reported association with skin malignancy although overexpression in hepatocellular carcinoma may indicate a poor prognosis (Fu, Li 2015).

**COL5A1** chr9:137,533,620-137,736,689 encodes for collagen type V alpha chain. Type V collagen is found with type I collagen (fibrillar collagen found in most tissues) and this gene helps regulate fibre assembly (proteinatlas.org).

**CYCI** chr8:145,149,930-145,152,428 encodes a protein involved in mitochondrial respiratory chain electron transfer, Upregulation of this process may help the cell overcome the deleterious effects of oxidative phosphorylation and overexpression has been identified in breast cancer cell lines as a marker of worse prognosis and metastasis (Han, Sun et al. 2016).

**DSP** chr6:7,541,808-7,586,950 encodes a protein that anchors desmosomes in the internal surface of the cell membrane. Mutations are associated with keratoderma, or abnormal thickening of the skin. It has been shown to be underexpressed in some Non small cell lung cancers (NSCLC) and its antitumoural properties might be the result of its impact on the expression of Wnt/ß-catenin genes Axin2 and MMP14(Yang, Chen et al. 2012).

**EPPK1** chr8:144,935,822-144,952,632 encodes a protein which is a member of the plakin family, generally involved in cytoskeletal architectural organisation, exclusively in epithelial
cells. This enables normal epithelial differentiation but may also effect cell migration in injured tissues as part of repair (Yoshida, Shiraki et al 2008).

**EDN1** chr6:12,290,529-12,297,427 encodes a preprotein that is then converted to a vasoactive peptide, enabling vasoconstriction. Its is overexpressed in colorectal cancer probably by interaction with β-catenin (Kim, Xiong et al 2005).

**EGFL7** chr9:139,553,308-139,567,130 also encodes for a vasoactive peptide. Most malignancies show expression and overexpression carries an unfavourable prognosis in colorectal and renal carcinoma (proteinatlas.org).

**PLEC** chr8:144,989,321-145,050,913 Plectin is an important cytoskeletal structural protein. It is known to be expressed in skin cancers, and its expression predicts a worse prognosis in renal, lung and colorectal cancer (proteinatlas.org).

**MYC** chr8:128,747,680-128,753,680 encodes a proto-oncogene that complexes with **MAX** to act as a transcription factor and impacts cell cycle, apoptosis and cellular transformation. It may also bind to VEGFA promoter to drive transcription and subsequent angiogenesis. Known to be amplified in cancers, can also shows translocation in both Burkitts Lymphoma and Multiple myeloma. Samples from both proteinatlas.org and from COSMIC from TCGA skin cancer (melanoma samples) suggest that overexpression is not universal or profound.

### 4.4.2.2 Recurrently deleted genes

**DCC** (Deleted in Colon Cancer) encodes for a protein which is a membrane receptor for netrin-1 (dependence receptor). When not activated by netrin-1, DCC has a pro-apoptotic effect. Once bound, it can block apoptosis by activation of MAPK pathway and by Caspase
3. As outlined above, DCC in our cohort was most significantly affected by LOH events in 8/15 samples.

**SMAD4**

SMAD4 is a member of a family of signal transduction proteins and is a tumour suppressor. Activated by TGF-β binding to serine protein kinases on the cell surface, the products of SMAD4 accumulate in the nucleus to regulate target gene transcription. Deletions have been associated with pancreatic malignancy, neuroendocrine tumours (Simbolo, Vicentini et al. 2018), juvenile polyposis syndrome, and hereditary haemorrhagic telangiectasia syndrome. Alternative splicing which also leads to decreased expression due to the predominance of an inactivated isoform has recently been described in a keratinocyte cell line subject to UVB radiation exposure (Ullah, Liao et al. 2018).
5 Short Variants

5.1 Results: Short variants

The overwhelming number of short variants detected were single nucleotide variants (SNV) in non-coding regions (including introns, 5’ and 3’ UTR and regulatory regions) (See Figure 3-2). The overwhelming majority of coding SNVs were missense mutations, where the resultant codon encodes for an alternate amino acid (Figure 5-1). The next most common class of short variant was a nonsense mutation, where the altered codon results in premature shortening of the resultant transcribed and translated protein (eg due to the introduction of a stop codon). This class of mutations represents less than 5% of the missense class.

The predominant pattern of SNV was C>T (refer to Figure 3.2). This made up more than 80% of SNV and is consistent with the dominant effect of UV radiation on pyrimidine bases, in keeping with our general findings on mutational signature (Section 3.3) and recently published (Mueller, Gauthier et al 2019).

Figure 5-1: Coding short variant classification. Short coding variants were assessed for rates of recurrence and likely impact on the basis of various predictors using SIFT, PolyPhen and CADD (see Chapter 2.6). These tools provide evidence of the likely impact of variants on biological activity and are best used in combination.
A gene list was collated that included known tumour suppressor genes, oncogenes, other genes falling within genomic hotspots in other cancer surveys (Gonzalez-Perez, Perez-Llamas et al. 2013) and genes of specific relevance to surveys of cutaneous malignancy (Su, Viros et al. 2012). In total, 1365 genes comprised the list that was used to assess for variants within genes (coding and non-coding regions) using the Seave platform. The gene list used for identifying short variants, copy number variation and structural variant effects of the coding genome is listed in Appendix 3. Other lists of specific non-coding regions, including specific promoter regions, long non-coding RNA and micro-RNA are described separately below.

5.2 Highly recurrent somatic mutations

The most highly recurrent short somatic variants with scaled CADD >10 are represented, combined with concomitant CNV in Figures 5-2 and 5-3. Two figures are presented combining single nucleotide variation and copy number variation in the same samples for the same genes/chromosomal bands. These two figures differ in thresholds for calling of variants and the filtering applied in Figure 5-3 to include those genes with at least 4 COSMIC variants and to exclude larger genes (eg MUC16, CSMD3) where variant frequency is high but may not be as meaningful (Lawrence, Stojanov et al. 2013).
Figure 5-2 Recurrent short variants in genes also subject to CNV– including PCLO, CSMD3 and MUC16. Top histogram: short variant events per sample colour coded. Histogram at right: Number of samples effected. Colour coding shows type of alteration.
Figure 5-3 Common short variants (and co-existent CNV) excluding PCLO, CSMD3 and MUC16. Histograms as per Figure 5-2.
Following is a detailed description of the types of mutation for each of the genes listed in Figures 5.2 and 5.3.

**TP53** (chr17:7,565,097-7,590,863). In our cohort, a mixture of high impact variants as well as copy number events were observed. A total of 79 SNV were identified across the cohort, 31 with a CADD >20.

**CDKN2A** (chr9:21,967,751-21,995,300). Eight samples showed high impact (mostly stop-gained) SNV, all within the coding region and in 4 cases occurring in multiple samples. All of these high impact variants had COSMIC IDs and had an average scaled CADD of 36. A further 6 examples of medium impact SNV were seen in **CDKN2A** including a missense variant occurring in 3/15 samples.

**MECOM** chr3:168,801,287-169,381,563). The predominant pattern of variation in **MECOM** was amplification. It was also the most highly amplified oncogene with 10 samples (66%) harbouring a CNV with average CN of 4, and no evidence of any LOH. Of the 6 samples with high and medium impact SNVs, one sample (4699) had a stop gained (COSMIC IDs COSM1420480, COSM4948202, COSM5829617) and the others were missense mutations. Of these variants, the average scaled CADD was 20.53.

**NOTCH1** chr9:139,388,896-139,440,314 Like **MECOM** the predominant pattern of variation in **NOTCH1** was amplification. Notwithstanding this, numerous high impact SNV were observed including 5 samples with stop-gain mutations in the second half of the gene (negative strand), and numerous significant medium impact missense and splice region variants.
**PTPRD** chr9:8,314,246-10,612,723 In our series, *PTPRD* showed a mixture of variants, dominated by high impact (inactivating) SNV, and CNV, predominantly by deletion and LOH. When filtered for impact, SNV in *PTPRD* were the most plentiful of any in our study, with 293 variants with a CADD > 20 across all samples.

**PLCB4** chr20:9,049,410-9,461,889 Within our cohort, one specimen had both a stop gained and a splice acceptor SNV with high impact. In total, 56 SNV with a scaled CADD > 20 were identified across all samples. We identified amplification in 9 of 15 samples, to an average CNV of 4. Two samples had LOH with CN =1.

**PCLO** chr7:82,383,321-82,792,246). The predominant pattern of variation for this gene was SNV, with a single sample showing amplification. Of the SNV, 4 samples displayed stop gained, while one other sample had a high impact splice variant as well as two high impact frameshift deletion events.

**PPP6C** chr9:127,908,852-127,952,218 A single stop-gained SNV and 5 other missense SNV were identified in our series for *PPP6C*.

**FAT4** chr4:126,237,554-126,414,087). *FAT4* SNVs were observed in 5 samples. Of these short variants, there was a single high impact stop-gain, and 57 other mutations with a minimum scaled CADD of 20. CNV analysis revealed 2 samples with LOH (CN =1), both samples also harbouring missense SNV.
CSMD3 chr8:113,235,157-114,449,328. Five samples had a stop gained SNV in this gene, effecting codons 22-67/71. One stop gain mutation effected 2 samples. This mutation has 2 COSMIC IDs: COSM1721763, COSM1721764. These catalogue to melanoma and rectal adenocarcinoma, across 5 curated samples. A further 152 SNVs with a minimum scaled CADD of 20 were identified. These included both coding and non-coding regions.

SYNE1 chr6:152,442,819-152,958,936 A single sample in our study showed 3 high impact mutations in this gene, including a stop gained and 2 splice acceptor variants. A further sample had a stop gained (high impact). In total 50 SNVs were identified in 12/15 samples with a CADD > 20. A single LOH event effecting SYNE1 was noted in one sample. There were 8 amplification events, only one of which had a CN >4.

PTCH1 chr9:98,205,262-98,279,339 PTCH1 mutations were common, but it was the least effected by high impacting variants. Only 3 variants with a CADD > 20 were identified, all were missense SNV, and none were classified as “probably damaging”.

DCC chr18:49,866,542-51,062,273 In our cohort, SNV in DCC was prominent with a stop gained truncating mutation in one sample, but in all other samples another 60 variants with scaled CADD >20. Of particular note, DCC in our cohort was most significantly effectted by LOH events in 8/15 samples.

AMPH chr7:38,423,297-38,671,167 We observed prominent amplification of this gene with 10 samples having duplications with CN form 3-5.
**MUC16 chr19:8,959,520-9,092,018.** MUC16 was subject to a high rate of inactivating mutations, including a stop gained in 8/15 cases, and another 9 cases of coding variants with CADD >20, including a missense SNV 9:8997432 (glutamate to lysine) effecting 2 samples with COSMIC ID COSM4546866/7, catalogued by COSMIC in the context of aggressive cSCC (Pickering, Zhou et al. 2014, South, Purdie et al. 2014), although not mentioned in these publications. We identified 6 samples with duplication events.

**NRXN1 chr2:50,145,643-51,259,674** High impact missense variants were identified in 8/15 samples, including with COSMIC ID COSM 4475612 seen in high risk cSCC (Pickering, Zhou et al. 2014).

**USH2A chr1:215,796,236-216,596,738.** Three samples with a total of 5 stop gained SNVs were identified for this gene with another 46 missense variants with CADD > 20 evident. Within cutaneous malignancy, USH2A variants have been documented in multiple COSMIC samples, predominantly melanoma, but also in a cSCC (COSM4547692) (Pickering, Zhou et al. 2014). COSMIC data shows an overwhelming proportion of variants of USH2A to be C>T (94% of 15408). We would have expected this to be the same, but only 25/46 (54%) were C>T. This anomaly, given the overwhelming nature of SNV being C>T in the overall cohort, is difficult to account for.

**SEMA3D chr7:84,624,869-84,816,171** Only one sample had a stop gained and splice accepter high impact variant, and another 6 samples exhibited 10 separate missense SNV, including 7:84628811 C>T which is shared by 4 samples. This variant has a COSMIC ID COSM170328, seen in cSCC (Pickering, Zhou et al. 2014), melanoma, prostate and bowel carcinoma. It has a scaled CADD of 29 and is probably damaging as assessed by PolyPhen.
One sample had a stop gained, 8 samples showed missense SNV (19 SNV events ie most samples had more than 1 missense variant) and a further 15 non-coding short variants of high impact (CADD > 20), 6 of which were short deletions for this gene. CNV revealed 2 samples with large deletion events.

5.3 Other genes of interest based on known association with metastasis

*PLAU* chr10:75,668,935-75,677,259 encodes the protein urokinase plasminogen activator which converts plasminogen to the broad spectrum serine protease plasmin responsible for the degradation of extracellular matrix and tumour migration and proliferation (Ranson and Andronicos 2003). There were no coding short variants in *PLAU* in any sample. There were numerous large amplification events on Chr10 spanning the entire gene (Figure 4-17) and see also circos plots 4699, 9120, 35649, 35818, 48585, 183410, 184577, 193958, 321773 above.

*FGFR2* chr10:123,237,844-123,357,972. *FGFR2* encodes for a cell surface growth factor receptor, which when upregulated can activate both the RAS-MAPK pathway and the PI3K-AKT pathway. *FGFR2* variants with scaled CADD scores > 20 were largely missense variants. Coding SNV resulting in codon changes were seen at codon numbers:

216 – S > L (Sample 9120) within the Ig-like C2-type 2 domain

572 – G > E (Sample 4699) within protein kinase domain

591 – H > Y (Sample 183410) within protein kinase domain

654 – S > F (Sample 48585) within protein kinase domain

778 – S > L (Sample 9120) within the cytoplasmic domain

All were assessed as Deleterious on Provean and Damaging on SIFT.
**RIPK4** chr21:43,159,529-43,187,266. *RIPK4* has been postulated as a potential driver gene in cSCC (Pickering, Zhou et al. 2014), by activation of Wnt/Hedgehog/Notch signalling pathways. Five samples (33%) showed medium impact SNV, including COSM21051 which annotates a C>T substitution at 21:43176851 with a scaled CADD score of 28.6.

**RASAI** chr5:86,563,705-86,687,748. *RASAI* probably acts as a tumour suppressor gene by its inhibitory regulation of the Ras-cyclic AMP pathway (Pickering, Zhou et al. 2014). Three samples (20%) showed inactivating missense SNV with scaled CADD scores >25 for this gene.

**HRAS** (chr11:532,242-537,287). See above in Chapter 1.5.3. Four samples (27%) showed damaging and deleterious missense SNV. Only 1 sample harboured an amplification CN >= 5.

**PARD3** chr10:34,398,488-35,104,253. High impact short variants (1 stop gained and 1 splice acceptor region variant) were identified in 2 samples for this gene. Five samples showed missense SNV (all with scaled CADD > 20), with one of these samples also showing an inframe deletion. One of the missense SNV had a COSMIC ID catalogued to BCC. All samples showed at least one non-coding but scaled CADD > 10 SNV. There were no examples of amplification in this cohort and two samples had large scale deletion events. Neither of these large deletions were in samples with inactivating missense coding SNV.
Mismatch repair genes

To identify susceptibility to microsatellite instability (MSI) due to mutations in mismatch repair genes, we analysed variation in a subset of genes included in a review of MSI in melanoma (Chae 2016). These genes were the top 10 mutated DNA repair genes when comparing COSMIC and TCGA data for diseases including lung, breast, liver, large intestine and skin (melanoma). The variant data for these 10 genes is presented in Table 5-1. All genes had some degree of inactivating variant. A large and significant amplification of *TP53BP1* was seen in a single specimen (4699).

<table>
<thead>
<tr>
<th>Gene</th>
<th>Missense SNV/total samples</th>
<th>Stop gain or splice variant/total samples</th>
<th>Indel/total samples</th>
<th>CNV</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>TP53</em></td>
<td>11/15</td>
<td>7/15</td>
<td>In 1/15</td>
<td>1 x LOH</td>
</tr>
<tr>
<td><em>KMT2C</em></td>
<td>8/15</td>
<td>3/15</td>
<td>In 3/15 Del 3/15</td>
<td>4 x LOH CN1</td>
</tr>
<tr>
<td><em>POLQ</em></td>
<td>6/15</td>
<td>-</td>
<td>-</td>
<td>No large deletion events</td>
</tr>
<tr>
<td><em>ATM</em></td>
<td>6/15</td>
<td>1/15</td>
<td>Del 1/15</td>
<td>1 x LOH with CN gain, 1 x deletion</td>
</tr>
<tr>
<td><em>ATR</em></td>
<td>10/15</td>
<td>1/15</td>
<td>Del 2/15</td>
<td>9/15 CN gain</td>
</tr>
<tr>
<td><em>BRAC2</em></td>
<td>8/15</td>
<td>1/15</td>
<td>In 2/15</td>
<td>2 x LOH CN1</td>
</tr>
<tr>
<td><em>SLX4</em></td>
<td>4/15</td>
<td>-</td>
<td>-</td>
<td>No large deletion events</td>
</tr>
<tr>
<td><em>TP53BP1</em></td>
<td>4/15</td>
<td>-</td>
<td>In 1/15</td>
<td>1 x LOH CN1 and 1 x amplification CN8 (7:1)</td>
</tr>
<tr>
<td><em>PRKDC</em></td>
<td>7/15</td>
<td>1/15</td>
<td>-</td>
<td>1 x large deletion</td>
</tr>
<tr>
<td><em>CENPE</em></td>
<td>3/15</td>
<td>-</td>
<td>-</td>
<td>2 x LOH CN1</td>
</tr>
</tbody>
</table>

Table 5-1 Mismatch repair gene SNV/Indel and CNV

5.4 Mismatch repair genes

To identify susceptibility to microsatellite instability (MSI) due to mutations in mismatch repair genes, we analysed variation in a subset of genes included in a review of MSI in melanoma (Chae 2016). These genes were the top 10 mutated DNA repair genes when comparing COSMIC and TCGA data for diseases including lung, breast, liver, large intestine and skin (melanoma). The variant data for these 10 genes is presented in Table 5-1. All genes had some degree of inactivating variant. A large and significant amplification of *TP53BP1* was seen in a single specimen (4699).
5.5 Discussion

5.5.1 Short Variants

The overwhelming majority of short variant events in metastatic cSCC are non-coding (Figure 3-2). This underlines the utility of WGS in this study compared to the largest previous study of aggressive and high risk cSCC which employed WGS (Pickering, Zhou et al. 2014), and a recent major work using targeted NGS analysis of 10 000 tumours including metastatic cSCC. The conclusion could be drawn from our findings that it is simply our ignorance of the non-coding genome that limits the application of WGS to this disease and has little to do with the mutational landscape.

5.5.2 Recurrent SNV

Somatic SNV identification was one of the key aims of this project. We were interested in a metastatic cohort to establish the mutational landscape of cSCC. As with BRAF v600e, we wondered whether a single variant might be recurrent and provide an avenue for therapy in this disease. Of note, there were no BRAF coding SNV in our series.

The extent of somatic variation in “normal” UV exposed skin (Martincorena 2014) is significant. Nonetheless, the burden of mutation and variants seen in cancer associated genes is different between their cohort and ours, being at either end of the range of sun exposed and UV implicated pathologies. What is to be made of the extent of SNV and how can we best make sense of short variants in the face of the mutational burden in metastatic cSCC? The answer is obviously to more mindfully interrogate hotspots of impact in the non-coding
genome, some of which are known, and some will depend on findings from transcriptome analysis of our cohort.

The short variants that were recurrent and high impact (coding, scaled CADD >10) included many known tumour suppressors or other genes previously described in the cancer context. These genes and their known roles in carcinogenesis and metastasis include:

**TP53** was the gene in which most mutations occurred. *TP53* encodes a protein (p53) that acts as a tumour suppressor, with effects on apoptosis and cell cycle regulation, including by p21, GADD45 and 14-3-3 activation. Germline mutations cause Li Fraumeni Syndrome, characterized by early onset of many varieties of malignancy. Somatic variations in *TP53* are seen in all cancer types. Mode of variant may be large scale impact such as frameshift or premature stop codons, or by missense mutation.

**CDKN2A**—Cyclin-dependant kinase inhibitor 2A (p16) acts as a tumour suppressor, acting by inhibiting cyclin-dependant kinases. It is capable of inducing cell cycle arrest in G1 and G2. Its actions are many and can be related to its effect on p53 degradation (by binding to MDM2), or independent of p53 (by activation of cyclin complexes). The usual mechanism of inactivation is via deletion, but this was not the case in our series, whereby SNV was predominant. *CDKN2A* (p16) expression acts as a surrogate marker for HPV infection and is of particular interest in mucosal oropharyngeal SCC. It is thought that p16 expression may be a marker of disease more sensitive to therapy. The application of this theory is currently the subject of numerous clinical trials. Its role in cutaneous malignancy is less clear. There is good evidence that p16 expression in the lymph node metastases of cSCC is common (Beadle, William et al. 2013). Practical application of this finding supports a sceptical
approach to the routine attribution of p16 +ve SCC in cervical lymph node metastases to an occult mucosal primary.

**MECOM**– a gene involved in haematopoiesis. It has at least three transcribed isoforms, MDS1-EVI1, EVI1 and EVI1Δ324. Generally MDS1-EVI1 behaves as a tumour suppressor, whereas EVI1 and EVI1 324(commonly co-expressed with EVI1) act as oncoproteins, binding as transcription factors to ETS binding sites(Sayadi, Jeyakani et al. 2016), and are associated with aggressive cancers with poor prognosis. *MECOM* is a transcriptional regulator and oncogene which plays a role in development, cell proliferation and differentiation. It has anti-apoptotic effects by suppressing the JNK-1 mediated phosphorylation of c-Jun. c-Jun is usually activated (phosphorylated) in response to UV to protect against UV associated apoptosis. The role of an amplification (and therefore over-expression) of *MECOM* could be to decrease c-Jun activity in UV exposed cells and to impact on normal apoptotic regulation. *MECOM* overexpression is associated with worse prognosis in glioblastoma multiforme (Hou, Zhao et al. 2016). Translocations of this gene with *AML1* can occur in Acute Myeloid Leukaemia. In a study of irinotecan resistant colorectal cancer cell lines, MECOM was identified as a differentially expressed gene, acting through evasion of apoptosis and the MAPK pathway, to worsen prognosis in this therapy resistant group.

**NOTCH1** encodes a transmembrane protein with EGF like receptors which, once ligand bound, releases an intracellular component which has numerous roles in the regulation of transcription and subsequent proliferation, differentiation and apoptosis. The Notch signalling pathway is upregulated in murine mammary oncogenesis, and increased expression of Notch receptors has been associated with many malignancies.
PTPRD encodes a protein of the protein tyrosine phosphatase receptor family, that have roles in cell growth and differentiation and oncogenic transformation, their action opposing that of the tyrosine kinases. Large scale genomic events impacting CDKN2A can also affect PTPRD due to their proximity. PTPRD dephosphorylates STAT3, deactivating its tumourogenic activity. STAT3 hyperactivation is associated with decreased survival and resistance to EGFR-targeted therapy (Peyser, Du et al. 2015). PTPRD inactivation was demonstrated to significantly increase levels of STAT3 in HNSCC. PTPRD is a tumour suppressor that exhibits putative inactivating somatic variants in >50% of GBM and between 10-20% of head and neck mucosal SCC (Veeriah 2009).

PLCB4 encodes for a protein that catalyses the formation of inositol 1,4,5-trisphosphate and diacylglycerol from phosphatidylinositol 4,5-bisphosphate. Amplification of PLCB4 and subsequent overexpression has been associated with overall more aggressive disease and worse prognosis in primary GIST (Li et al 2017). PLCB4 is also one of a group of genes frequently mutated in uveal melanoma.

PCLO encodes Piccolo, which is a presynaptic cytomatrix protein. There is evidence for its role stabilising and preventing breakdown of EGFR, and in the progression of disease in oesophageal SCC (Zhang 2017). In this study, in both a knockdown mouse model, as well as using a monoclonal antibody targeting PCLO, tumour progression was inhibited. As such, in oesophageal SCC it behaves as an oncogene.

PPP6C encodes a catalytic subunit of the protein phosphatase that regulates IL2 receptor stimulation by removing phosphate groups and activating the cytoplasmic receptor tyrosine kinase. This gene was recently identified as altered in review of BCC, occurring in 15% of
primary tumours (Bonilla, Parmentier et al. 2016). Herein it is cited as an inhibitor of Cyclin D1. Loss of regulation or inhibition of Cyclin D1, and its overexpression is established in many cancers, including head and neck SCC (mucosal).

**FAT4** encodes a protocadherin that plays a role in regulating planar cell polarity, the Hippo signalling pathway, the Wnt signalling cascade and the expression of YAP1. In gastric cancer cell lines and in a mouse xenograft model, *FAT4* downregulation increases lymph node metastasis and worse survival, and increases growth and invasion of gastric cancer cells, with over expression of mesenchymal markers and decreased epithelial phenotype (Cai, Feng et al. 2015).

**CSMD3** encodes a transmembrane protein with CUB and Sushi Multiple Domains. Available data from multiple sources supports a tumour suppressor role for *CSMD3* and the other CSMD genes. Loss of function mutations of *CSMD3* were identified in a knockout airway epithelial cell line model to increase cell turnover and probable role in lung tumourogenesis in a non-small cell lung cancer model (Liu, Morrison et al. 2012).

**SYNE1** encodes for a spectrin repeat containing protein that localises to the cell membrane and assists in maintaining subcellular spatial organisation. It has influence in both meiosis and cell cycle pathways. Cutaneous melanoma exhibits *SYNE1* variants in 24% of samples, while HNSCC shows SNV in 18% (Intogen : Barcelona Biomedical Genomics Lab (Gonzalez-Perez, Perez-Llamas et al. 2013)).

**PTCH1** encodes a transmembrane protein receptor (Patched) for Sonic Hedgehog (SHH). Unbound, it suppresses the activity of Smoothened (and keeps it cytoplasmically bound
within and endosome), but once bound by SHH, it releases SMO to promote cell proliferation by nuclear GLI gene activation. Indeed, upon SHH binding, Patched is degraded. *PTCH1* is a key component of the Hedgehog pathway, and its protein’s action on SMO is the key event in contemporary Hedgehog pathway inhibition. Indeed, vismodegib acts to bind and inhibit SMO in effect, similar to the action of Patched.

*DCC* encodes for a protein which is a membrane receptor for netrin-1 (dependence receptor). When not bound, DCC has a pro-apoptotic effect. Once bound, it can block apoptosis by activation of MAPK pathway and by Caspase 3.

*AMPH* encodes a protein (amphyphysin) associated with the cytoplasmic surface of synaptic vesicles. There is a potential role for AMPH expression changes in cancer. A subset of patients with Stiff-Person Syndrome and breast cancer have autoantibodies to amphyphysin.

*MUC16* encodes CA-125, which is a transmembrane 0-glycosylated protein which helps to protect the apical aspect of epithelial cells. MUC16 interacts with mesothelin (MSLN) to activate matrix metalloproteinases to enhance invasion in pancreatic cancer (Chen, Hung et al. 2013). CA-125 is used as a marker of disseminated disease in the blood of ovarian cancer patients, and its role in other malignancies is emerging. Its cancer association is generally assumed to be by overexpression, probably as a result of amplification. Given the role of *MUC16* in invasion is likely to be by amplification, expression of *MUC16/CA-125* and MSLN relative to MMP is likely to determine any role in this disease.

*NRXN1* encodes for neurexin-1-beta, which is a cell-surface protein that binds to neuroligins and is involved in synapse communication between cells, and likely plays a role in cell
adhesion interactions. Deletions in this gene are associated with neurodevelopmental and neurobiological abnormalities (genecards.org).

**USH2A** encodes for a protein involved in hearing development. It has been implicated (perhaps as a tumour suppressor) as a result of observations of missense mutations of this gene in HCC, and its related genes **GPR98, PCDH15** and **MYO7A** in a review of 88 HCC by WGS. (Kan, Zheng et al. 2013).

**SEMA3D** encodes a protein involved in axon guidance in neural development and diseases associated with **SEMA3D** deactivating SNV include Meniere and Hirschsprung Disease.

**NAV3** is another gene involved in axon guidance and neurone development and impact IL2 production by T cells (Karenko, Hahtola et al. 2005) and in neural regeneration. Additionally, it plays a role in microtubule regulation and in a breast cancer xenograft model with **NAV3** knockdown, metastasis was increased, and patients with breast cancers expressing normal levels of **NAV3** show longer survival.

**FGFR2** encodes a critical receptor tyrosine kinase, one of four for FGFR. FGF plays a role in cell division, regulation of growth and maturation, angiogenesis and wound healing. Targeted next generation sequencing in a cohort of non-metastatic primary cSCC identified variants in **FGFR2** to be exclusively seen in tumours with perineural infiltration (PNI) of tumour cells (Zilberg, Lee 2018). Histopathology of lymph nodes infrequently displays PNI due to the absence of neural tissue which makes this finding difficult to correlate in our cohort.

**RIPK4** encodes a key serine/threonine protein kinase involved in keratinocyte and stratified epithelial differentiation. It is involved in both NK-kappaB signalling interactive and
Wnt/Hedgehog/Notch signalling pathways. Five samples (33%) in our series showed medium impact SNV, including COSM21051 which annotates a C>T substitution at 21:43176851 with a scaled CADD score of 28.6. This variant was also seen in a series of RIPK4 in primary head and neck cSCC (Pickering, Zhou et al. 2014). In that series, RIPK4 variants were identified in 28% cases. They found variants only in exon 2 and 8, whereas in our cohort, they included 2, 5, 7 and 8.

**RASA1** encodes an inhibitory regulator of the Ras-cyclic AMP pathway, acting via weak GTP-ase action to derive the inactive GDP bound RAS on the cytoplasmic extent of RTK. It thus acts as a tumour suppressor. In our cohort, three samples showed inactivating missense SNV with scaled CADD scores >25. This is in keeping with previously reported findings in cSCC (Pickering, Zhou et al. 2014).

**HRAS** encodes for one of the Ras activating genes, involved in signal transduction and the MAPK pathway. It is classified as an oncogene.

**PARD3** encodes an adaptor protein involved in asymmetrical and cell polarization processes. It is involved in epithelial tight junctions (Chen, An et al. 2017) and may also targets PTEN to the same tight junctions. Inherited defects have been implicated in neural tube defects as a result of the disruption of neuroepithelial morphogenesis. Its role in cancer has mostly been described as a tumour suppressor in breast cancer. Recently, evidence for a role of decreased expression of PARD3 in facilitation of the invasion of malignant cells within the breast tumour microenvironment by enhancing the sliding property of tumour cells within stroma (Milano, Ngai et al. 2016). In our samples, high impact short variants (1 stop gained and 1 splice acceptor region variant) were identified in 2 samples. Five samples showed missense SNV (all with scaled CADD > 20), with one of these samples also showing an
inframe deletion. One of the missense SNV had a COSMIC ID catalogued to BCC. All samples showed at least one non-coding but scaled CADD > 10 SNV. There were no examples of amplification in this cohort and two samples had large scale deletion events. Neither of these large deletions were in samples with inactivating missense coding SNV. These data support a high incidence of inactivating short variants potentially enhancing the metastatic process.

With such a range of high impact short variants (as well as CNV and SV) within our cohort, the question is how to interpret this against what we understand to be drivers of cancer. A recent update to the COSMIC database is the Cancer Gene Census (Sondka, Bamford et al 2018). This is not new data but a review and representation of existing COSMIC data and revision of the original paper (Futreal, Lachlan et al 2004).

Within the census, genes are categorised according to Tiers. Tier 1 genes possess documented activity relevant to cancer, along with evidence of mutations in cancer which change the activity of the gene product in a way that promotes oncogenic transformation. COSMIC also reviewed the existence of somatic mutation patterns across cancer samples in COSMIC. For instance, tumour suppressor genes often show a broad range of inactivating mutations and dominant oncogenes usually demonstrate well defined hotspots of missense mutations. Genes involved in oncogenic fusions are included in Tier 1 when changes to their function caused by the fusion drives oncogenic transformation, or in cases when they provide regulatory elements to their partners (e.g. active promoter).

Tier 2 genes are those that, whilst being implicated in cancer, have less robust evidence and often relate to genes that are more recent targets of interest.
Within our cohort, Table 5-2 represents those genes for which we observed either significant SV, CNV or high impact SNV compared to the same genes annotation within the Cancer Gene Census.

Of note is that for our reported SNV, we have a minimum scaled CADD of 10 for predominantly missense variants. Other higher impact short variants including stop gained, frameshift and splice donor or acceptor variants are also present. The same is not the case for the data from the CGS. No such minimum impact is required. Rather, for each gene, following a search of the COSMIC database and the available literature, data is compared from multiple reported series (at least 2), and each piece of evidence must satisfy at least 2 expert (post-doctoral scientists) reviewers. The gene is then ascribed a Tier.

Within the CNV/SV group, concordance in this comparison is seen within SOX2, MYC, DCC, SMAD4, PDE4D. Additionally, most of our high impact (missense) SNV share this observation with the CGS annotations.
Table 5-2 Comparison of metastatic cSCC CNV/SV genes and high impact SNV with annotations from COSMIC Cancer Gene Census.

<table>
<thead>
<tr>
<th>GENE</th>
<th>Met cSCC</th>
<th>CGS (Tier)</th>
<th>GENE</th>
<th>Met cSCC</th>
<th>CGS (tier)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NDRG1</td>
<td>Amp</td>
<td>Trans (1)</td>
<td>TP53</td>
<td>FS Splice</td>
<td>Mis SNV</td>
</tr>
<tr>
<td>PIK3CA</td>
<td>Amp</td>
<td>Mis SNV(1)</td>
<td>CDKN2A</td>
<td>FS Mis SNV</td>
<td>Del/FS/Splice/Mis/Non SNV (1)</td>
</tr>
<tr>
<td>SOX2</td>
<td>Amp</td>
<td>Amp (1)</td>
<td>MECOM</td>
<td>Mis SNV</td>
<td>Trans (1)</td>
</tr>
<tr>
<td>ABL1</td>
<td>Amp</td>
<td>Mis SNV/Trans (1)</td>
<td>NOTCH1</td>
<td>Mis SNV</td>
<td>Trans Mis SNV (1)</td>
</tr>
<tr>
<td>ASXL1</td>
<td>Amp</td>
<td>FS, Mis/Non SNV (1)</td>
<td>PTPRD</td>
<td>FS Mis SNV</td>
<td>Del Mis/Non SNV (1)</td>
</tr>
<tr>
<td>BAI1</td>
<td>Amp</td>
<td>-</td>
<td>PLCB4</td>
<td>Mis SNV</td>
<td>-</td>
</tr>
<tr>
<td>CAP2</td>
<td>Amp</td>
<td>-</td>
<td>PCLO</td>
<td>Mis SNV</td>
<td>-</td>
</tr>
<tr>
<td>COL5A1</td>
<td>Amp</td>
<td>-</td>
<td>PPP6C</td>
<td>Mis SNV</td>
<td>Mis/Non SNV (1)</td>
</tr>
<tr>
<td>CYC1</td>
<td>Amp</td>
<td>-</td>
<td>FAT4</td>
<td>Mis SNV</td>
<td>Mis/Non SNV (1)</td>
</tr>
<tr>
<td>DSP</td>
<td>Amp</td>
<td>-</td>
<td>CSMD4</td>
<td>Mis SNV</td>
<td>-</td>
</tr>
<tr>
<td>EPPK1</td>
<td>Amp</td>
<td>-</td>
<td>SYNE1</td>
<td>Mis SNV</td>
<td>-</td>
</tr>
<tr>
<td>EDN1</td>
<td>Amp</td>
<td>-</td>
<td>PTCH1</td>
<td>Mis SNV</td>
<td>FS/Splice/Mis/Non SNV (1)</td>
</tr>
<tr>
<td>EGFL7</td>
<td>Amp</td>
<td>-</td>
<td>AMPH</td>
<td>Mis SNV</td>
<td>-</td>
</tr>
<tr>
<td>PLEC</td>
<td>Amp</td>
<td>-</td>
<td>MUC16</td>
<td>Mis SNV</td>
<td>Mis SNV(2)</td>
</tr>
<tr>
<td>MYC</td>
<td>Amp</td>
<td>Amp/Trans (1)</td>
<td>NRXN1</td>
<td>Mis SNV</td>
<td>-</td>
</tr>
<tr>
<td>DCC</td>
<td>Mis SNV/Del</td>
<td>Mis/Non SNV Del (1)</td>
<td>USH2A</td>
<td>Mis SNV</td>
<td>-</td>
</tr>
<tr>
<td>SMAD4</td>
<td>Del</td>
<td>FS Mis/Non SNV Del (1)</td>
<td>SEMA3D</td>
<td>Mis SNV</td>
<td>-</td>
</tr>
<tr>
<td>TTC28</td>
<td>Trans</td>
<td>-</td>
<td>FGFR2</td>
<td>Mis SNV</td>
<td>Mis SNV(1)</td>
</tr>
<tr>
<td>MYLK</td>
<td>Trans</td>
<td>-</td>
<td>RIPK4</td>
<td>Mis SNV</td>
<td>-</td>
</tr>
<tr>
<td>PDE4D</td>
<td>Trans</td>
<td>Trans</td>
<td>RASA1</td>
<td>Mis SNV</td>
<td>-</td>
</tr>
<tr>
<td>EPHB1</td>
<td>Trans</td>
<td>-</td>
<td>HRAS</td>
<td>Mis SNV</td>
<td>Mis SNV(1)</td>
</tr>
<tr>
<td>PTK2</td>
<td>Trans</td>
<td>-</td>
<td>PARD3</td>
<td>Mis SNV</td>
<td>-</td>
</tr>
</tbody>
</table>
5.5.3 Mismatch repair genes

Mismatch repair defects give rise to microsatellite instability (MSI) due to hypermutation. Microsatellites are small recurring repeats, usually one to six base pairs in length. MSI was originally described in Lynch syndrome, a non-polyposis colorectal cancer, the phenotype having mainly right sided cancers occurring in the 6th decade and having a signet ring histopathology (Thibodeau 1993). This disease is characterised by germline mutations in the MSI genes MSH2, MLH1, MSH6, PMS2, and PMS1 (Pawlik 2004). In particular SNPs in MSH2 and MLH1 were seen most frequently in tumours exhibiting MSI.

A comprehensive review of somatic mutations, CNV and expression frequencies of DNA repair genes within COSMIC identified recurrently somatically varied genes in melanoma (Chae 2016). The top 10 genes associated with subsequent MSI in melanoma were TP53, KMT2C, POLQ, ATM, ATR, BRCA2, SLX4, TP53BP1, PRKDC, CENPE. Somatic variants within our cohort across these genes is presented in Table 6-3.

As stated earlier, in a series of high risk cSCC, Pickering et al reported inactivating mutations of KMT2C in 39 cSCC, which encodes a histone methylation complex to alter transcriptional regulation. This mutation has been identified in other malignancies and was associated with increased incidence of bone invasion and a shorter time to recurrence in cSCC (Pickering, Zhou et al. 2014). Once again, in our series, it was a key MMR gene with a high rate of inactivating mutations. This confirms the earlier findings of a likely role in advanced and metastatic disease.
5.5.4 Pathway impacts from short variants

5.5.5 Canonical Wnt/β-catenin and Hippo interaction

The canonical Wnt/β-catenin pathway is one of 3 Wnt pathways (canonical, non-canonical (cell polarity and calcium homeostasis). Wnt signalling is responsible for many regulatory and growth limiting effects with obvious cancer implications. Canonical Wnt pathway is activated via Wnt ligand binding to the Frizzled receptor (Fz) with subsequent intracytoplasmic accumulation of β-catenin and translocation into the nucleus to act as a transcription factor for genes including c-myc and cyclin D1. When the Wnt ligand is not bound, β-catenin does not accumulate, but rather is phosphorylated and ubiquinated to be degraded by proteasomes.

Hippo is a pathway that acts to regulate the size of organs. It is impacted by cell density and so in a high cell density is in the activated state. Such a state is meant to regulate further cell division and organs growth. This is ultimately by phosphorylation (and degradation) of YAP/TAZ, the final drivers of anti-apoptotic and proliferation genes including diap1 and cyclin E (Kim and Jho 2014).

Mechanisms for interaction between Wnt and Hippo pathways include the blocking of β-catenin’s nuclear localization (and therefore activation of regulators of proliferation) by YAP/TAZ, remembering that YAP/TAZ activation (non-phosphorylation and escape from degradation) occurs in the low-cell density state. Additionally, β-catenin levels, and their ubiquitinisation complex levels may also impact on the levels of TAZ in the cytoplasm (Azzolin Cell 2012). FAT4 is one of the previously unknown upstream activators of Hippo. Its deactivation should lead ultimately to less phosphorylation of YAP/TAZ with less resultant degradation. This would then replicate the low-cell density state and drive anti-
apoptotic and proliferative genetic drivers. As reported earlier, \textit{FAT4} showed significant missense SNV in our cohort. Conversely, we have identified an amplification of \textit{DSP} which, in NSCLC acts as tumour suppressor by inhibiting the Wnt/\beta-catenin pathway.

### 5.5.6 Notch signalling

Of the 4 NOTCH receptor encoding genes, \textit{NOTCH1} showed the most variation in metastatic cSCC. Activation of the NOTCH pathway is highly conserved, and binding of ligands and activation of the transmembrane NOTCH receptor transfers the Notch Intracellular domain (NICD) to the nucleus to activate transcription of specific target genes and activation of downstream targets including PI3K, AKT and p21.

### 5.5.7 Hedgehog pathway signalling

\textit{PTCH1} was recurrently varied in metastatic cSCC although perhaps not to the same deleterious biological effect of some of the other highest impacting genes by SNV. Nonetheless, there is a degree of variant activity in this gene which is likely to impact on its role as a key transmembrane receptor in the Hh pathway. Inactivation of \textit{PTCH1} to disallow reception of the Hh ligand will impact disinhibition of \textit{SMO}. \textit{SMO} affect to induce expression and post translational modification of GLI zinc finger transcription factors (Karamboulas and Ailles 2013).

### 5.5.8 Cell Cycle

Genes in the cell cycle pathway predominate in the most recurrent single nucleotide variants in our samples. \textit{TP53} and \textit{CDKN2A} are genes most impacted by high impact (eg stop gained and splice region) and missense variants. \textit{SMAD4} is the gene with the second highest rate of large scale deletion events. Crossover effects to the regulation of apoptosis magnify the loss
of not just the cell cycle but also programmed cell death as a result of these inactivating mutations.

Figure 5-4 Cell cycle Kegg diagram highlighting TP53, CDKN2A and SMAD4 interaction. Source https://www.kegg.jp/kegg-bin/show_pathway?hsa04110

5.5.9 MAPK and MEK inhibition

The Ras-Raf-MEK-Erk pathway starts with the bindings of a receptor tyrosine kinase and ends with the activation of transcription factors in the nucleus to drive cell growth, differentiation, proliferation, apoptosis and migration functions. Key genes within the pathway showing somatic variation in our cohort include the tyrosine kinase regulator PTPRD and PPPC6. In recognised key MAPK genes DUSP4, DUSP6, MAP2K1, MAP3K1, MAP3K2, MAP3K3, MAP3K7, MAP3K9, MAPK1, MAPK3, MAPK8, MAPK12, MAPK14 we found multiple missense variants across 10 of 15 samples. These were most prominent in DUSP4, MAP3K1, MAP3K3, MAP3K9 and MAPK14. No such mutations were seen in DUSP6, MAPK1 or MAP2K1. The usual targets of MEK pathway inhibition, BRAF and RAS
did not show significant variation, and there were no examples of V600E (or V600K) mutation in our samples.
6 Results: Non-coding regions

6.1 TERT promoter variants

*TERT* promoter mutations (TPM) were seen in 13 (93%) metastatic tumours (Figure 6-1).

These included common and widely reported *classic* mutations C228T (chr5: 1295228) (n=3) and C250T (chr5:1295250) (n=8), but also the less common but described variants A161C(chr5:1295262), C205T(chr5:1295205) (n=7), C242T(chr5:1295242), C243T(chr5:1295243) and C252T(chr5:1295252). Of those cases with TPM, 85% showed either C228T or C250T, but none showed both.

The classic TPM C228T was seen in 21% of all samples and 23% of those with TPM. In this series the previously infrequently observed mutation C205T, was seen in 50% of all samples and 54% of samples with TPM (Figure 6-1).

Amplification (CN=>4) of *TERT* itself was demonstrated in 13/15 samples (Figure 6-1), with a minimum block size of 5000bp.
In order to understand whether any of the observed TPM impacted on telomerase expression, paraffin blocks of all metastatic tumour samples were cut and stained for immunohistochemical (IHC) assessment (TERT Antibody (A-6), Santa Cruz Biotechnology, California). Of those cases that had TPM, only 25% stained positive for telomerase (Figure 6-2). The greatest association between TPM and positive IHC was for the C205T variant, whereby 75% of cases that stained positive had this mutation, and 43% of the C205T cases stained positive. Only 1 out of 8 cases with C250T stained positive, with this case also having C205T TPM. There was no correlation between positive IHC and clinicopathological features. Vinagre et al found a trend toward higher TERT expression as assessed by IHC in gliomas with TPM, but this failed to reach significance. Furthermore, there was no association between TPM or amplification pattern, and differentiation of tumours, nodal
ratio, resection status or survival in our cohort. Thus, telomerase IHC is unhelpful in considering TPM effect.

![Image of immunohistochemistry analysis showing positive telomerase staining in tumour cells of sample 35818 with no staining (negative control) of stroma evident. 35818 had a TERT CN of 9 and was the only sample with C205T and A161C. TERT (A-6): sc-393013 Santa Cruz Biotechnology Inc.](image)

Figure 6-2 Immunohistochemistry analysis showing positive telomerase staining in tumour cells of sample 35818 with no staining (negative control) of stroma evident. 35818 had a TERT CN of 9 and was the only sample with C205T and A161C. TERT (A-6): sc-393013 Santa Cruz Biotechnology Inc

### 6.2 NFKBIE

Across the NFKBIE promoter co-ordinates 6:44233377-6:44233437, SNVs were identified in 9/15 specimens (60%), with 5 of these samples having more than one promoter variant.

Chr6:44233400 C>T was seen in 5/15 (33%) samples with an average VAF of 0.25.

In total, more than 50% of cases showed an amplification, one in combination with an LOH event.

Noteworthy is that all samples with a CNV other than 48585 also had a promoter variant. 48585 herein shows a major amplification over 370000bp with a CN of 6 (5:1).
6.3 MicroRNA

We selected a group of microRNA (miRNA) that had been identified in SCC of the vulva (Melo-Maia, Lavorato-Rocha et al 2013) and cervix (Ding, Wu et al 2014) as potential biomarkers for progression of disease and also as potential serum biomarkers. These miRNA included miR-1246, miR-20a, miR-2392, miR-3147, miR-3162-5p miR-4484. To assess a wider block size (5000bp) than confined only to the miR co-ordinates (in the region of 22 nucleotides each), manual co-ordinates were entered into Seave for analysis. These co-ordinates were:

miR-1246 Chr 2: 176817825-178208520
miR-3147 Chr 7: 56738644-58335801
miR-4484 Chr 10: 125523232-125649350
miR-3162 Chr 11: 59265408-59409255
miR-20a Chr 13: 91984898-92091600
miR-2392 Chr 14: 101233633-101308604

Copy number analysis of the regions spanning these miRNA showed both CN loss and gain. miR-3147 showed deletion in 5 samples (33%). This same miRNA was duplicated in 100% of samples with an average CN of 4.6 and average block size of 19.7 Mb. 13/15 samples showed a duplication of miR-2392 with an average CN of 3.7.

A selection of microRNA implicated in cSCC (miR-21, miR-184, miR-31, miR-203 and miR-205) (Dziunycz, Iotzova-Weiss et al 2010) were analysed for CNV and SNV in our cohort. (Table 6.2). As discussed in Chapter 9.5.3, miR-203 and 205 are heavily amplified in our samples. In particular, miR-203 showed highly consistent amplification (in all but 1 sample) and no evidence of deletion. Within this group miR-31 (known to suppress growth,
invasion and colony formation in metastatic cSCC cell lines (Wang, Landen et al 20140) had the highest average CN in those samples showing amplification.

**Table 6-1** cSCC associated microRNA Copy number variation.

<table>
<thead>
<tr>
<th>microRNA</th>
<th>Amplified (n)</th>
<th>Ave CN amplification</th>
<th>Deleted (n)</th>
</tr>
</thead>
<tbody>
<tr>
<td>miR-21</td>
<td>6/15</td>
<td>3.5</td>
<td>4/15</td>
</tr>
<tr>
<td>miR-184</td>
<td>5/15</td>
<td>3.9</td>
<td>2/15</td>
</tr>
<tr>
<td>miR-31</td>
<td>6/15</td>
<td>4.5</td>
<td>2/15</td>
</tr>
<tr>
<td>miR-203</td>
<td>14/15</td>
<td>3.7</td>
<td>-</td>
</tr>
<tr>
<td>miR-205</td>
<td>10/15</td>
<td>3.6</td>
<td>3/15</td>
</tr>
</tbody>
</table>

### 6.4 Long non-coding RNA

A number of cancer associated lncRNAs, including HOTAIR, LINC00568, TERC, LINC00657 (NORAD), TCF7, TINCR, MALAT1 (Schmitt and Chang 2016) PVT1, HULC, ,RP11-65J3.1(Xie, Jiang et al. 2018) and PICSAR (LINC00162) (Pipponen et al 2016)were analysed for both SNV and CNV. SNV (with minimum scaled CADD >5) were common across all samples (Table 6-2). 200971 was the least varied sample with only 2 SNV of lncRNA PVT1 and TINCR.

**Table 6-2** Frequency of lncRNA SNV in 15 metastatic cSCC specimens. Scaled CADD >5

<table>
<thead>
<tr>
<th>lncRNA</th>
<th>SNV # samples (of 15)</th>
<th>Average scaled CADD</th>
</tr>
</thead>
<tbody>
<tr>
<td>HOTAIR</td>
<td>5</td>
<td>10.8</td>
</tr>
<tr>
<td>LINC00568</td>
<td>2</td>
<td>9.3</td>
</tr>
<tr>
<td>TERC</td>
<td>1</td>
<td>6.7</td>
</tr>
<tr>
<td>LINC00657</td>
<td>9</td>
<td>8.7</td>
</tr>
<tr>
<td>TCF7</td>
<td>9</td>
<td>9.1</td>
</tr>
<tr>
<td>TINCR</td>
<td>5</td>
<td>9.5</td>
</tr>
<tr>
<td>MALAT1</td>
<td>12</td>
<td>9.9</td>
</tr>
<tr>
<td>PVT1</td>
<td>15</td>
<td>7.7</td>
</tr>
<tr>
<td>HULC</td>
<td>2</td>
<td>9.1</td>
</tr>
<tr>
<td>LINC00162</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>RP11-65J3.1</td>
<td>11</td>
<td>6.9</td>
</tr>
<tr>
<td>MIRLET7A1</td>
<td>1</td>
<td>16.3</td>
</tr>
</tbody>
</table>
CNV of these lncRNA was predominantly amplification (Table 6-3). In particular 13 of the 15 samples had amplification in PVT1 with an average CN of 4.8 and no deletions.

Table 6-3 Copy number variation long non-coding RNA

<table>
<thead>
<tr>
<th>lncRNA</th>
<th>CN</th>
<th>Amplification # samples (of 15)</th>
<th>Ave Size BP for Amp (Mb)</th>
<th>Deletion # samples (of 15)</th>
</tr>
</thead>
<tbody>
<tr>
<td>HOTAIR</td>
<td>3.5</td>
<td>8</td>
<td>24.61</td>
<td>2</td>
</tr>
<tr>
<td>LINC00568</td>
<td>3.9</td>
<td>9</td>
<td>19.13</td>
<td>2</td>
</tr>
<tr>
<td>TERC</td>
<td>4.1</td>
<td>11</td>
<td>14.47</td>
<td>0</td>
</tr>
<tr>
<td>LINC00657 (NORAD)</td>
<td>4.2</td>
<td>9</td>
<td>12.88</td>
<td>1</td>
</tr>
<tr>
<td>TCF7</td>
<td>3.2</td>
<td>5</td>
<td>36.12</td>
<td>3</td>
</tr>
<tr>
<td>TINCR</td>
<td>3.2</td>
<td>8</td>
<td>9.89</td>
<td>2</td>
</tr>
<tr>
<td>MALAT1</td>
<td>3.9</td>
<td>11</td>
<td>13.01</td>
<td>2</td>
</tr>
<tr>
<td>PVT1</td>
<td>4.8</td>
<td>13</td>
<td>20.80</td>
<td>0</td>
</tr>
<tr>
<td>HULC</td>
<td>4.0</td>
<td>10</td>
<td>19.13</td>
<td>0</td>
</tr>
<tr>
<td>RP11-65J3.1 (linc01503)</td>
<td>4.0</td>
<td>8</td>
<td>15.64</td>
<td>0</td>
</tr>
<tr>
<td>MIRLET7A1</td>
<td>4.0</td>
<td>8</td>
<td>32.02</td>
<td>0</td>
</tr>
<tr>
<td>LINC00162</td>
<td>3.6</td>
<td>6</td>
<td>5.81</td>
<td>3</td>
</tr>
</tbody>
</table>

6.5 Discussion: Non-coding Regions

Most of the genome is non-coding. Non-coding regions include 3’ and 5’ untranslated regions (UTR), promoter and other regulatory regions, long non-coding RNA, introns, either in genes or intergenic with no clear functional role. Interrogating non-coding regions is a
mammoth task. This task is most sensibly approached either in piecemeal fashion targeting known hotspots or regions with transcription factor binding sites, or in conjunction with expression data which identifies over or under expressed regions not obviously the result of coding variants. And just as not all somatic variants are coding, not all expression changes will be the result of genomic variance, highlighting the impact of epigenetic modulation and stromal interaction.

Non-coding regions discussed here are as a result of assessment of genes in interest only and are not presented as comprehensive. A further review of non-coding regions will follow a more thorough analysis of the expression pattern in cSCC generally. Promoter and enhancer variants probably impart effect by changing the affinity of transcription factor binding sites. Promoter and enhancer (regulatory) regions are rich in transcription factor binding sites. University of California Santa Cruz (UCSC) genome browser includes a track which includes TFBS in regulatory regions. From this resource, each TFBS can be interrogated for the evidence behind each claim, including the cell line or disease in which the enhancement is seen. Given the lack of genomic assessment of cSCC (moreover for metastatic cSCC), the application of data derived from ICGC or 1000 Genomes is not always clear.

6.5.1 TERT promoter mutations
Telomerase is a ribonucleoprotein polymerase, maintaining telomere ends by addition of the telomere repeat TTAGGG. In most normal somatic cells, telomerase is repressed allowing for progressive shortening of telomeres and eventual senescence. Telomerase is made up of an internal telomerase RNA template (encoded by gene TERC chr3) and the enzyme, telomerase reverse transcriptase (encoded by gene TERT chr5). The TERT promoter spans the coordinates chr5:1295154-1295376 (50-270BP upstream from the transcription start site).
Deregulation of telomerase expression in somatic cells may be involved in oncogenesis, and TPM are perhaps the most common variants in cancer.

Vinagre et al (2013) evaluated the presence of TERT promoter mutations (TPM) across a number of malignant cell lines and tumours. They identified rates of TPMs in various malignancies; glioblastoma (49%), urothelial carcinoma of the bladder (59%), differentiated thyroid cancer (10%) and melanoma (29%). Corresponding upregulation of expression was noted in these groups. In the same study, no mutations were found in renal cell carcinoma, GIST or phaeochromocytoma. Within melanocytic tumours, only sun exposed cutaneous melanoma showed recurrent mutations, with no such evidence in either benign nevi or ocular melanoma. Huang (2013) in a study of 329 cell lines form various tissues and pathologies found a high rate of recurrent TERT mutations C228T and C250T within the melanoma group. Most mutations observed were either C250T or C242T. Hugdahl, Kalvenes et al (2018) recently described the rate of TPM in matched primary and metastatic melanoma. They found a rate of 68% and 64% respectively, with 24% of mutations being discordant.

Some telomerase activity persists in normal skin. Burnworth (2007) observed that over the spectrum from normal skin, through keratoacanthoma to invasive SCC, TERT expression was equally upregulated. In a focused examination of TERT promoter variants of SCC from various sites Cheng et al (2015) showed a similar pattern of recurrent and mutually exclusive mutations to that reported by Huang, but most marked in UV associated tumours. These mutations had no impact on clinicopathologic behaviour. Scott et al (2013) surveyed TPM in primary skin cancers compared to benign skin conditions, without matched normal germline control. They found the highest rate of mutations in basal cell carcinoma (78%). Mutations
were found in SCC (50%) and Bowens Disease (cSCC insitu) (9%), but no TERT promoter variants were found in benign skin. Zehir (2017) in 27 cases of metastatic nodal cSCC all with DNA extracted from FFPE samples and subjected to targeted NGS (including TERT promoter) found 32% of patients harboured TERT promoter variants; however, none of the patients exhibited C228T or c205T. TPM have been identified (Jung, Kim et al 2017) as a negative prognostic indicator in Non Small Cell Lung Cancer, being associated with both regional lymph node metastasis and worse overall survival.

This mutual exclusivity of C228T and C250T was first described by Huang (2013). Breakpoint analysis showed no translocations or inversions effecting either promoter or TERT in any samples. One sample showed a 3Mb tandem duplication of TERC.

The classic TPM C228T has not been previously described in metastatic cSCC. Herein it was seen in 21% of all samples and 23% of those with TPM. In this series the previously infrequently observed mutation C205T, was seen in 50% of all samples and 54% of samples with TPM. In their collection of a variety of solid organ metastases, Zehir et al observed C205T mutation in only 0.05% of all cases (5 of 10336) and 0.4% of those cases with TPM (5 of 1232).

TERT regulatory variants outside of the promoter region (total co-ordinates span chr5:263002-2208200 [Ensembl.org]) were identified in all samples with promoter region mutations. Some of these variants occurred in multiple samples. These occurred in Enhancer, TF, CTCF (CCCTC-binding factor) bindings sites, open chromatin and Promoter Flanking regions. Of the 6 cases without TPM, 2 also had a variant in an open chromatin TERT regulatory region.
This high rate and novel pattern of TPM may be in response to the high mutational burden and may give rise to the high level of structural variation, including chromothripsis, although this paradoxical theory of telomere crisis is theoretical. Telomere crisis is characterised by the shortening of telomeres due to increased cell division (unregulated growth occurring early in carcinogenesis) to the extent that telomeres are completely lost to expose the chromosome proper (Chin, de Solorzano et al. 2004). Subsequent replication of isolated chromosomal fragments may give rise to chromothripsis and katagesis (Maciejowski, Li et al. 2015).

*ATRX* can act to reduce telomere shortening via an alternate mechanism to the action of *TERT*. Its loss, by either mutation or deletion, might lead to telomere length loss. This process is described in glioma (Wiestler, Capper et al. 2013) to be mutually exclusive to TPM. In our series, this was not the case and the only coding mutation in *ATRX* was seen in a single sample that also had C250T TPM. Nonetheless, the missense mutation (C>T) with amino acid change (leucine to phenylalanine) seen was probably deleterious and had a scaled CADD of 29.8, and has been described in cholangiocarcinoma with COSMIC ID COSM4767440, COSM4767441.

### 6.5.2 NFKBIE

*NFKBIE* is a gene with 2 isoforms (Shain, Garrido et al. 2015). The longer isoform is present mainly in brain tissue, whereas the shorter isoform is present in all other tissues. Shain et al identified mutation hotspots within the promoter region of the shorter isoform, particularly in desmoplastic melanoma. One of these variants is chr6:44233400 C>T (G>A on coding negative strand). The mutation hotspots within the promoter of *NFKBIE* are within consensus binding sites for multiple transcription factors. Across the promoter co-ordinates 6:4423377-
we identified SNV in 9/15 samples (60%), with 5 of these samples having more than one promoter variant. Chr6:44233400 C>T was seen in 5/15 (33%) samples with an average VAF of 0.25. In the original paper outlining this finding, 2 of 20 samples of desmoplastic melanoma also showed \textit{NFKBIE} amplification.

\subsection*{6.5.3 microRNA}

MicroRNA are small (19-25nt) fragments of the genome that are transcribed to a primer and then pre miRNA before exiting the nucleus as mature miRNA. In the cytoplasm they are activated by Dicer and then interact with mRNA in a complex known as the RNA-induced silencing complex (RISC). The effect of the RISC can be to either completely silence the translation of the mRNA or to reduce its expression and enhance its degradation, depending on the match of the miRNA and the 3’ UTR of the target gene/mRNA. The first human miRNA (let-7) was discovered in 2000 (Reinhart, Slack et al 2000). miRNA are often located in close proximity to genes (eg \textit{TP53} and \textit{miR-34}) a trait that enables or augments their activity to reduce protein expression and in turn to potentially play a role in pathology. miRNA can be thought of in much the same fashion as cancer associated genes, ie having proto-oncogenic or tumour suppressing effect. Because miRNA through the RISC opposes gene expression by disabling mRNA (a miRNA may be involved in the RISC for a number of different genes), an over expression of a miRNA that bind to a tumour suppressor gene will have a pro-cancer effect, and similarly, under expression of a miRNA that binds to the mRNA of an oncogene will have the same effect. The converse would be the same in the tumour suppressing context. So miRNA are of interest in their interaction with cancer associated genes and might be able to be used as a biomarker or potentially as a therapeutic target. Not insignificant in assessing impact of miRNA in disease is the identification of the target genes for each.
Matched tumour and serum miRNA were investigated in oral SCC (OSCC) (Schneider, Victoria et al. 2018). They found amongst a group of 48 miRNA differentially expressed between healthy tissue and OSCC (25 down and 23 up-regulated), 30 were also able to be identified in the serum samples. They promoted the utility of miR-32 as it was up-regulated in both states compared to healthy tissue. miR-32 was identified as associated with colon, pancreas and prostate cancer in one of the earliest studies matching miRNA with solid tumour types (Volinia, Calin et al 2006). miR-32 is located on the long arm of chr1, near PTPN3, a protein tyrosine phosphatase which has been implicated in numerous solid organ malignancies, both to inhibit and to promote (Gao, Zhao et al. 2014, Li, Lai et al. 2015). In this context of liquid biopsy biomarkers, the microRNA miR-1246, miR-20a, miR-2392, miR-3147, miR-3162-5p miR-4484 were identified as serum markers of metastatic cervical SCC(Chen, Yao et al. 2013). These miR cluster together in pathways by using miR-Path (Vlachos, Zagganas et al. 2015)

In our cohort, copy number analysis of the regions spanning these miRNA showed both CN loss and gain. miR-3147 showed deletion in 5 samples (33%). This same miRNA was duplicated in all 100% of samples with an average CN of 4.6 and average block size of 19.7 Mb. 13/15 samples showed a duplication of miR-2392 with an average CN of 3.7.

A study of HPV negative vulval cSCC described overexpression of miR-3147 in cancers compared to non dysplastic adjacent skin. Levels of overexpression were proportional to depth of invasion, but not metastatic spread (Yang and Guo 2018).
We also identified a highly amplified state of miR-203, a miRNA previously identified as under-expressed in cSCC when compared to normal skin (Dziunycz, Iotzova-Weiss et al 2010). This report did not analyse metastatic cSCC. Indeed in our samples, of the five miRNAs analysed in Dziunycz et al (miR-21, miR-184, miR-31, miR-203 and miR-205) miR-203 stood out due to there being no deletion events but having sizeable amplification in 14/15 specimens. In SNV terms, miR-203 showed downstream SNV in 10/15 samples, though of debatable impact without total RNASeq data.

This amplification of miR-203 in metastatic cSCC is at odds with most of the body of evidence surrounding this microRNA. In colorectal cancer for example, depletion of miR-203 is associated with a higher grade and lymph node metastasis (Deng, Wang et al 2016). Conversely, overexpression was associated with improved survival and suppression of growth in vitro on CRC cell lines. Using a bioinformatic algorithm called TargetScan, Deng et al inferred the target of the miR-203 blockade was EIF5A2. Expression of miR-203 and EIF5A2 were essentially opposite. The one exception from the published series is that of epithelial ovarian cancers, which generally displayed the opposite effect, that miR-203 levels are proportional to aggressiveness and are prognostically adverse (Iorio, Versone et al 2007).

In our cohort, with significant amplification of miR-203, we found that EIF5A2 (located on 3q – a highly amplified region – see above) was amplified in 11/15 samples with an average CN of 4.2 in the amplified samples. There were no deletions of this gene. This is not the same as either transcription influence or protein translation, but the amplification of both seems to suggest there is more to this story, particularly in our cohort.

TP63 is a target of miR-203 in effect reducing p63 expression (Yi, Poy 2008). This means that in vitro it reduces stemness in the suprabasal epithelial layers and promotes
differentiation. Conversely, miR\_205 is thought to target E-cadherin expression, and in doing so, be associated with a less well-differentiated form. In comparing miR-203 and miR-205 and their expression in high risk cSCC, again, miR-203 was more likely to be expressed in better outcome disease, when compared to miR-205, which was associated with more aggressive biology in terms of recurrence and locoregional metastasis (Canueto, Cardenoso-Alvarez et al 2017). Again, in terms of purely genomic observations of amplification and deletion, this difference is not seen in our cohort, with amplification of miR-203 being the most prominent amplified cSCC associated microRNA. Total RNA Seq will qualify and quantify any expression difference.

Also within the list of miRNA of Dziunycz is miR-31. It has been further analysed by Wang et al (Wand, Landen et al 2014) wherein they describe its apparent role in both primary and metastatic cSCC, utilizing a previously described metastatic cSCC cell line UT-SCC-7. They found that blocking of miRNA expression reduces invasion, migration and colony formation of UT-SCC-7. miR-31 was not overexpressed in normal tissue or actinic keratosis. Within our samples, miR-31 amplification was seen in 6/15, with an average CN of 4.5 (Table 6-1), perhaps providing genomic evidence of over-expression.

### 6.5.4 Long non-coding RNA

lncRNA are a type of noncoding RNA that have transcripts greater than 200 nucleotides in length without protein coding function. These are thought to be highly specific to differentiated tissues and diseases including cancer. lncRNA influence gene expression by a variety of mechanisms. The physiological and pathological mechanism of miRNA action is more fully understood and there are a number of miRNA inhibitors and mimics currently being developed for use in cancer and chronic diseases (Matsui and Corey 2017). lncRNA
situated between genes are referred to as long intergenic ncRNA, and there is a subclass of lincRNA that is termed very long intergenic non-coding RNA as their size is greater than 50kb.

PVT1 (lincRNA) upregulation is associated with worse prognosis and a more aggressive phenotype in many cancers. It shares a locus with the known oncogene MYC on chr8. There is significant correlation between amplification of PVT1 and MYC due to their close proximity. In breast cancer, PVT1 may interact with the miRNA-200 family to promote early disease events (Colombo, Farina et al. 2015). It may exert its role in gastric cancer by interfering with the ubiquitinisation of STAT3 to reduce its degradation, and in doing so, to promote angiogenesis (Zhao, Du et al 2018).

In this study, we have shown amplification in chr8q24.21 which includes both MYC and PVT1. In addition, CNV identified PVT1 as amplified (to an average CN of 4.8) in 13/15 samples of metastatic cSCC (87%). There was no evidence of any deletions in this lincRNA. The highest rate of amplification frequency seen in TCGA data is for ovarian cancer at 43% (Colombo, Farina et al. 2015). Of note is that all samples with amplification of PVT1 showed corresponding amplification of MYC presumably due to the block size (average size of amplified element affecting PVT1 20.80 Mb) covering both regions in the long arm of chr8. PVT1 holds some promise, not just in cSCC, as a potential biomarker of disease and/or progression.
7 Germline variants

While the focus of this study has been an analysis of the landscape of somatic variants in metastatic cSCC, we have, in the process of interrogation of genes of interest, identified a number of recurrent inherited variants that may be implicated in the development of cSCC, and may play a role in the progression of cSCC.

7.1 Mismatch repair genes

In our cohort, significant germline events were identified in 14/19 whole blood samples. The most recurrent of these variants, seen in 9/20 samples, was a splice region SNP of MSH2 (chr2:47641560 (A>C) (dbSNP ID rs11309117), with a scaled CADD of 11.63. This SNP was not identified in any of the more than 4000 genomes sequenced in the Medical Genome Reference Bank healthy elderly adult germline database, compiled and curated by the, Kinghorn Centre for Clinical Genomics/Sydney Genomics Collaborative. (https://www.garvan.org.au/research/kinghorn-centre-for-clinical-genomics/research-programs/sydney-genomics-collaborative/mgrb). One other SNP, effecting one specimen, had a probably damaging and deleterious variant in PMS2 (chr7:6045634 T>C) rs63750123 (scaled CADD 26.9), seen in 1% of MGRB samples. This SNP also has a COSMIC ID (COSM601786), catalogued as a missense somatic variant in tumour sample of a sample of metastatic cSCC (Li, Hanna et al. 2015)

7.2 TERT promoter

Germline analysis of all included samples reveals recurrent Chr5p15.33 SNPs. We identified 3 recurrent and novel variants in the TERT/CLPTM1L super-enhancer SE_66421 in 16 of 19
samples. The incidence of these events in normal elderly healthy adult genome library (MGRB – as discussed above) (n=5678) are 43%, (rs466502/Chr 5:1325767 A>G), 40% (rs31488/Chr 5:1342156 A>G) and 44% (rs27070/Chr 5:13465303 G>C). This combination of SNPs has not been previously described and they have neither independently nor in combination been described in cSCC.

These germline events occur less than 1500BP upstream of the TSS of CLPTM1L, which in turn is immediately upstream of TERT. They also sit within the bound region of ENSR00000177543 which is the CLPTM1L regulatory element with binding sites for promoting transcription factors including PKNOXI and ATF1.

Germline analysis reveals 93% of our cases have a G>C SNP at 5:1346303 (rs27070), less than 1500BP upstream of the TSS of CLPTM1L, within the TERT associated super-enhancer SE_66421. This SNP also sits within the bound region of ENSR00000177543 which is the CLPTM1L regulatory element with binding sites for promoting transcription factors including PKNOXI and ATF1. This SNP is present in 43% of 5678 sequenced germlines in our database of healthy subjects. SNPs in the TERT/CLPTM1L.

### 7.3 Discussion

Much effort in this analysis of cSCC has been on tumour variation. We know, using xeroderma pigmentosum as a model (see above Chapter 1.2) that predisposing inherited conditions can mirror the effect of drug induced immunosuppression in increasing the incidence of cSCC. There has been previously no investigation into whether novel inherited predisposing genetic factors might play a role in determining a landscape for either a more
aggressive form of cSCC or for disease which readily metastasizes. These initial results suggest peculiarities in the germline of this cohort of patients with metastatic disease and offers an avenue for a more formal interrogation. This idea is supported by recent work by Ioannides and Wang et al (2018) who have identified through genome wide and transcriptome wide analysis loci of likely predisposition for cSCC. These genes include CTSS, HORMAD1, GOLPH3L and ANXA9 at 1q21, CASP8 at 2q33, AHI1 at 6q23, HAL at 12q23, and ORMDL3 at 17q21. Further analysis of larger germline samples within an expanded group of our cases will allow some insights into any relevance of these genes in susceptibility in the metastatic disease setting.
8 Expression analysis

The Nanostring platform using the Pancancer Progression Gene Panel for multiplex gene expression analysis was used. This catalogue of cancer associated genes (n=770) was selected to match best with predicted patterns of variants associated with the metastatic process. The genes analysed in the panel are listed in Appendix 6 and are known drivers of carcinogenesis, metastasis or epithelial mesenchymal transition.

8.1 Results

Expression data for across all 770 genes is shown in Figure 8-1. The samples and genes are clustered in an unsupervised manner and show considerable variation in gene expression across all genes and specimens.

Within each of the following heatmaps, sample 34943 shows significant difference from the remained of the cohort. This sample is from a patient who had left sided parotid lymph node metastasis. Of all the patients included, this patient was the only to have previously had a mucosal SCC, in addition to the included metastatic cSCC. They had undergone a contralateral maxillary resection for an alveolar SCC 6 years prior to the emergence of their cutaneous disease. It is unclear as to the significance of the observed expression difference and how much previous treatment may have influenced subsequent behavior of a new, anatomically related, malignancy.
Figure 8-1 Heatmap analysis of expression of all 770 genes in the Nanostring Pancancer Progression panel for all 15 metastatic cSCC specimens. Lower panel shows (log2)*gene expression.

Next, gene expression analyses of tumour samples based on genes with recurrent short variants across our cohort (refer to Figure 5-2) that are also included in the Nanostring Pancancer Progression Panel was performed (Figure 8-2). For each of these genes, there was demonstrable expression changes for high impact SNV. Within TP53, the only samples with
high impact SNV (eg stop-gained or deletion) were 321773, 48585 and 4699. Conversely, 193958 had no high or medium impact SNV in TP53. With respect to PTPRD, while 38532 has evidence herein of decreased expression, and while it does have a high/medium impact variant is a missense SNV, the SNV is classified as benign and tolerated, but with a scaled CADD of 23. Within each sample, SNVs can be implicated in some of what is observed in the expression analysis, but certainly not all. This supports the conventional wisdom that expression is the summation of multiple influences including both genetic and epigenetic.

Figure 8-2 Heatmap analysis of expression of genes with recurrent short variants across 15 metastatic cSCC specimens. Lower panel shows (log2)*gene expression.
Expression of those genes that are included in the Nanostring Pancancer Progression Panel with significant CNV as reported in Figure 5-3 is shown in Figure 8-3. DCC and SMAD4 show recurrent deletions, all other genes were amplified.

Figure 8-3 CNV and expression effect for the genes listed in Figure 5-3 that were included in the 770 gene Cancer Progression Panel. Samples with SMAD4 and DCC deletions listed beneath the heatmap. The samples in which SMAD4 is deleted include 321773, 183410 and 184577. The samples which showed deletion of DCC include 34366, 321773, 183410 and 184577.
As discussed earlier (Chapter 1.8), many genes are potentially implicated in EMT. Expression differences in these genes are not necessarily driven by genetic variants. Epigenetic influence may be a key driver of EMT. We examined a group of genes previously identified as exhibiting expression differences in cSCC animal models (Pastushenko, Brisebarre et al. 2018). This study defined tumour cells on the basis of their expression of EPCAM markers CD61, CD51 and CD106, and differential expression of genes including CDH1, EPCAM, KRTN14, ASPN, MMP1, TWIST1, VCAM, ZEB1, LOX.

The expression heatmap of these genes across our cohort is presented in Figure 8-4.
MYLK, EPHB1 and PTK2 were amongst the genes that were shown to have gene break points in more than one sample. To examine the effect of these coding break points and/or gene fusions, all samples were tested for MYLK, EPHB1 and PTK2, as these genes were amongst the 770 genes included in our Nanostring Pan Cancer Progression Panel. The findings are presented in Figure 8-5.

![Figure 8-5 Expression of genes with breakpoints (samples with breakpoints highlighted in blue for given gene). Samples with MYLK breakpoints were 9120 and 184577. Samples with EPHB1 breakpoints were 4699 and 321773. Samples with PTK2 breakpoints were (including sample 34934 which had gene fusion with SLA2). These expression differences are inconclusive.](image-url)
8.2 Discussion

It is ultimately expression that determines a biological effect. Variants in the genome may be transcribed and translated in an altered fashion to affect a pathological outcome. Various methods exist for the correlation of genomic observations with gene expression. These include formal sequencing of the entire transcribed RNA (transcriptome), a more limited coding RNA sequencing, PCR (polymerase chain reaction) of isolated regions of transcribed RNA by identification using primers, newer digital counting technologies and more clinical means such as immunohistochemical assessment and in situ hybridization. There are pros and cons to each platform, derivatives mainly of cost and technical demands, and within any experiment, one or more of these may be applicable to different components of the analysis.

NanoString NCounter technology uses a hybrid probe to allow RNA in solution to be identified with a capture probe and subsequently reported by a reporter probe. This highly automated process delivers molecule counts for each gene of interest as a measure of expression. The molecule counts are compared and calibrated to that of several housekeeping genes.

The extent of intracohort variability for selected genes with either CNV or high impacting SNV was significant. Some of these findings were counterintuitive. For example, in Fig 8-2, the sample with the least expression of TP53, relative to other samples, was 48585, which showed TP53 amplification. However, it is worth noting that this sample also harboured a high impact SNV, so the expression affect could be due to, for example, a premature stop-gained or a splice variant curtailing transcription. The correlation between our identified genomic events and the NanoString expression is not linear, and ultimately the employment of RNASEq to more accurately measure the entire transcriptome will provide for greater assurance around discrepancies. There are other instances where a clear expression change
cannot be accounted for with only genomic data, without an understanding of transcriptional translational influences beyond the observed variation.
9 Conclusion and reflections

9.1 Conclusion

Cutaneous squamous cell carcinoma (cSCC) ranks as the most common lethal malignancy. Primary cSCC ranks second only to BCC in prevalence, and together they comprise more numbers than all other cancers combined. But most cSCC are able to be treated with simple excision with clear margins. This is made more complex in head and neck regions, given anatomical constraints and aesthetic considerations. Nonetheless, less than 5% of cSCC will ever metastasize.

There are no clinicopathologic predictors of risk of metastasis. Large databases of high risk and metastatic cSCC, such as that maintained at the Sydney Head & Neck Cancer Institute, support the idea that tumours of the lip and ear are more likely to spread to cervical lymph nodes. However, by no means do all tumours behave in this aggressive way, and therefore most can be treated simply and not subjected to rigorous surveillance, other than for further primary cutaneous lesions.

The impact of metastasis on the patient with cSCC is profound. Spread to lymph nodes usually requires multimodal treatment including the surgical resection of effected lymph node basins and adjuvant external beam radiotherapy. A large study of the effect of such therapy reveals the quality of life costs (Wang, Palme et al 2013). The issue is not the metastasis perse, but rather that most patients with metastatic disease have not been the subject of surveillance to risk and with the aim of identifying regional disease spread. This is not due to some dereliction of care, but rather that no reliable means of prediction of risk, and therefore application of surveillance, has been established. The result of this clinical dilemma impacts not just the individual patient, whereby lymph node metastases are often only identified at an
advanced stage, but also the community, where effective strategies for surveillance and pre-
eemptive management could save not just morbidity to individuals but could more effectively
and efficiently use public money and resources.

The gaps in our knowledge of the genomic landscape of cSCC have lagged behind that of
many other less common cancers. In particular, as a marker of activity, publications in cSCC
compare unfavourably to those on melanoma (Figure 9.1) despite far greater prevalence.

![Figure 9-1 Publication rates per year for the last 10 years comparing cSCC, Skin SCC
and Melanoma (PubMed).](image)

This project is the first dedicated examination of the whole genome in metastatic cSCC. This
disease has the highest somatic genomic mutational burden of any disease ever described.
The overwhelming majority of this variation is in the non-coding genome. Within the coding
genome, the observed mutational burden provides evidence for the application of checkpoint
inhibiting immunotherapeutics. Metastatic cSCC is characterized by a highly disrupted
genome with significant structural variation. The predominant structural variation is
amplification. Within the coding genome, a pattern of highly impacting and recurrent SNV of
both tumour suppressor genes and oncogenes is evident. Within the non-coding genome, novel patterns of variation have been identified within the TERT promoter and also within noncoding RNA.

This project is an exploration of the genomic landscape of metastatic cSCC. This subject underpins the work now commenced on looking at expression changes in metastatic versus primary disease, with a view to eventually providing validated biomarkers to discriminate cancer behaviour and identify those tumours at greatest risk of metastasis from those at little risk.

The suite of genes with high impact variants, and subject to copy number variation and or structural rearrangement is significant, both in terms of synergy with other disease, and for their potential to provide both a measurable biomarker, and a potential for therapy for advanced or unresectable disease, or indeed for distant metastatic disease.

The subjects of investigation that have become obvious future areas if interest from this work include:

Interrogation of non-coding variants- of note was the extent to which WGS illuminates the variation in the non-coding genome. We have herein identified some noteworthy regions which have stemmed from findings in other diseases, including the TERT promoter and miRNA and lncRNA. This is however by no means an exhaustive list and a concentration on these findings as well as other known hotspots of influence must be undertaken.

Analysis of germline variants – this project was not designed as an investigation of the germline looking for predisposing inherited variants. However, even within the CLPTM1L/TERT super-enhancer, we were able to identify SNPs of unusual combination and recurrence. We understand the role of the immune system in predisposing to cSCC. It would therefore seem illogical to imagine that components of the genetic makeup of individuals
could not play a role in predisposing to an exaggerated effect of UV exposure, either through innate immunodeficiency or through stromal variation.

*Transcriptome analysis* – this work must be combined with a thorough analysis of the transcription and expression of the genomic events we have discovered. Outside of using circulating DNA remnants as a biomarker of disease, an examination of the genomic state of tumours does not provide any targets of use as a tool for assessing risk. Certainly the discoveries herein provides clues to targets, but the transcriptome will potentially provide a more useful picture of high risk and metastatic cSCC.

Already planned future work with this dataset will include the abovementioned transcriptome analysis and integration and will also need to be enhanced by increasing the overall numbers of both primary and metastatic deposits, including those cases that have synchronous primary and metastatic disease.

### 9.2 Reflections

It has not been possible to extrapolate from the genomic analysis of 15 samples any particular clinicopathologic behaviour attributable to somatic variation, other than to observe that the metastatic stage of cSCC has a high mutational burden with prominent amplification. Within our patient group, there was a wide variation in age, comorbid disease and immunocompetency, all of which potentially confound a direct line between any given genomic event and a clinical effect. So it would be fair to say that small numbers in this study have not helped overcome the challenge of drawing specific conclusions with regard to clinicopathologic behaviour.

Studies such as this need to be based on a watertight and meticulously archived tumour bank. It is very challenging, even for a dedicated clinical team, to annotate and care for specimens
to that level required of such a project. At every stage, we have been reminded of the absolute requirement for record keeping and sample movement rigour, and have, on occasions, needed to repeat experiments and retrace our steps, with time and money sacrificed. During this project, the establishment of a highly sophisticated and research enabling tumour bank (CONCERT) has taken over the ongoing collection and curation of our precious clinical specimens and combines this with a prospective clinicopathologic database.

A single sequencing facility and team would have reduced opportunities for error and overall would have reduced the work of sample delivery and downstream processing. Whilst the AFP at ANU were as helpful and accommodating as the geographic dislocation allowed, the delivery of tissues overseas to Macrogen did incur delay and further handling steps. Subsequent delivery of raw data to allow for alignment, variant calling and eventual bioinformatics analysis was more cumbersome than would have been afforded had we done all the sequencing and processing at the KCCG, or any other single quality assured sequencing provider.

I would have spent more time and consulted more widely in the establishment of an exact protocol for cellularity estimation. I failed to discuss this process with other groups who may have been in the position to help with a protocol to diminish the rate of poor tumour cellularity and subsequent unusable sequencing effort. We had designed a system that relied too heavily on potentially inexact and subjective measures, based on tissue samples that may not have eventually matched exactly that which was subsequently used for DNA extraction and sequencing. The protocol we eventually landed on was to take a small block of tissue either side of that to be sequenced (30mg blocks) and to then have a Specialist Pathologist give a tissue cellularity. If the tissue is assessed at >50%, the extracted DNA should then be
submitted to a SNP array to confirm the suitability for WGS. One WGS undertaken on low
cellularity (unusable) tissue would pay for 15 SNP arrays. This modification of process has
educated a subsequent round of WGS of DNA extracted from primary cSCC.

Despite the learning curve associated with the technical and cognitive gains required, this
project has been transformative. It has allowed me to further understand the biology of a
challenging disease, to appreciate the scientific method and to build a strong collaboration of
interested clinicians and scientists to progress this work.


Brunner, M., Ng, B. C., Veness, M. J., Clark, J. (2014) "Assessment of the new nodal classification for cutaneous squamous cell carcinoma and its effect on patient stratification." Head Neck 36


de Melo Maia B, Lavorato-Rocha AM, Rodrigues LS, Coutinho-Camillo CM, Baiocchi
Fu J, Li M, Wu DC, Liu LL, Chen SL, Yun JP.(2015)"Increased Expression of CAP2 Indicates Poor Prognosis in Hepatocellular Carcinoma".Transl Oncol; 8(5):400-6


Horn, S. et al. (2013) "TERT Promoter Mutations in Familial and Sporadic Melanoma". Science 339:959-961


Martincorena, I., A. Roshan, M. Gerstung, P. Ellis, P. Van Loo, S. McLaren et al. (2015)
by increased lymphatic density and enhanced expression of macrophage-derived VEGF-C." J Invest Dermatol **131**(1): 229-236.


Rachakonda, P. S. et al. (2013) "TERT promoter mutations in bladder cancer affect patient survival and disease recurrence through modification by a common polymorphism" Proc Natl Acad Sci U S A 110(43) : 17426-31


Tan X, Chen M (2014) "MYLK and MYL9 expression in non-small cell lung cancer identified by bioinformatics analysis of public expression data" Tumour Biol;35(12):12189-200


Veness MJ, Quinn DI, Ong CS, Keogh AM, Macdonald PS, Cooper SG, Morgan GW (1999) "Aggressive cutaneous malignancies following cardiothoracic transplantation: the Australian experience" Cancer; 85(8): 1758-64

Vinagre J.et al. (2013). "Frequency of TERT promoter mutations in human cancers." Nature Communications 4: 2185  https://doi.org/10.1038/ncomms3185


Xu N, Zhang L, Meisgen F, Harada M, Heilborn J, Homey B, Grandér D, Ståhle M,


### Appendix 1 AJCC Staging 8th Edition Head and Neck SCC

<table>
<thead>
<tr>
<th>Tm</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tx</td>
<td>Primary tumor cannot be identified</td>
</tr>
<tr>
<td>Tis</td>
<td>Carcinoma in situ</td>
</tr>
<tr>
<td>T1</td>
<td>Tumor &lt;2 cm</td>
</tr>
<tr>
<td>T2</td>
<td>Tumor ≥2 cm but &lt;4 cm</td>
</tr>
<tr>
<td>T3</td>
<td>Tumor &gt;4 cm or minimal bone erosion or PNI or deep invasion</td>
</tr>
<tr>
<td>T4a</td>
<td>Tumor with gross cortical bone/marrow invasion</td>
</tr>
<tr>
<td>T4b</td>
<td>Tumor with skill base invasion and/or skull base foramen involvement</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>pN category</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>N1</td>
<td>Metastasis in a single ipsilateral lymph node ≤3 cm and ENE (−)</td>
</tr>
<tr>
<td>N2a</td>
<td>Metastasis in a single ipsilateral or contralateral lymph node ≤3 cm and ENE (+) or a single ipsilateral node &gt;3 cm but ≤6 cm and ENE (−)</td>
</tr>
<tr>
<td>N2b</td>
<td>Metastasis in a multiple ipsilateral lymph nodes ≤6 cm and ENE (−)</td>
</tr>
<tr>
<td>N2c</td>
<td>Metastasis in bilateral or contralateral lymph nodes ≤6 cm and ENE (−)</td>
</tr>
<tr>
<td>N3a</td>
<td>Metastasis in a single lymph node &gt;6 cm and ENE (−)</td>
</tr>
<tr>
<td>N3b</td>
<td>Metastasis in a single lymph node &gt;3 cm and ENE (+) or multiple ipsilateral, contralateral, or bilateral nodes, any with ENE (+)</td>
</tr>
</tbody>
</table>

### TNM stage

<table>
<thead>
<tr>
<th>Stage</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Tx, N0, M0</td>
</tr>
<tr>
<td>I</td>
<td>T1, N0, M0</td>
</tr>
<tr>
<td>II</td>
<td>T2, N0, M0</td>
</tr>
<tr>
<td>III</td>
<td>T3 or N1, M0</td>
</tr>
<tr>
<td>IV</td>
<td>T4 or N ≥ 2 or M1</td>
</tr>
</tbody>
</table>

AJCC, American Joint Committee on Cancer. HNcSCC, head and neck cutaneous squamous cell carcinoma. PNI, perineural invasion. ENE, extranodal extension.
### Appendix 2: Nucleic acid extraction results

<table>
<thead>
<tr>
<th>Sample ID</th>
<th>Extraction Method</th>
<th>Extraction Quality</th>
<th>DNA Concentration (ng/μL)</th>
<th>DNA Integrity Score</th>
<th>Gene Expression</th>
<th>Data Quality</th>
</tr>
</thead>
<tbody>
<tr>
<td>1A</td>
<td>RNA</td>
<td>High</td>
<td>20</td>
<td>8</td>
<td>9</td>
<td>100%</td>
</tr>
<tr>
<td>2B</td>
<td>DNA</td>
<td>Low</td>
<td>5</td>
<td>4</td>
<td>5</td>
<td>75%</td>
</tr>
<tr>
<td>3C</td>
<td>RNA</td>
<td>Medium</td>
<td>15</td>
<td>6</td>
<td>6</td>
<td>88%</td>
</tr>
<tr>
<td>4D</td>
<td>DNA</td>
<td>High</td>
<td>30</td>
<td>8</td>
<td>8</td>
<td>95%</td>
</tr>
<tr>
<td>5E</td>
<td>RNA</td>
<td>Medium</td>
<td>25</td>
<td>7</td>
<td>7</td>
<td>90%</td>
</tr>
<tr>
<td>6F</td>
<td>DNA</td>
<td>Low</td>
<td>10</td>
<td>3</td>
<td>3</td>
<td>60%</td>
</tr>
<tr>
<td>7G</td>
<td>RNA</td>
<td>High</td>
<td>18</td>
<td>9</td>
<td>9</td>
<td>98%</td>
</tr>
<tr>
<td>8H</td>
<td>DNA</td>
<td>Medium</td>
<td>12</td>
<td>5</td>
<td>5</td>
<td>80%</td>
</tr>
</tbody>
</table>

*Note: Data quality scores range from 0% to 100%, with higher scores indicating better data quality.*
Appendix 3 Gel Electrophoresis
Appendix 4 Gene List

AHNAK, BMP1, CALD1, CDH2, COL1A2, COL3A1, COL5A2, FN1, FOXC2, GNG11, GSC, IGFBP4, ITGA5, ITGAV, MMP2 MMP3, MMP9, MSN, SERPINE1, SNAI1, SNAI2, SNAI3, SOX10, SPARC, STEAP1, TCF4, TIMP1, TMEFF1, TMEM132A, TWIST1, VCAN, VIM, VPS13A, WNT5A, WNT5B, CAV2, CDH1, DSP, FGFBP1, IL1RN, KRT19, MITF, MST1R, NUDT13, PPPDE2, RGS2, SPP1, TFPI2, TSPAN13, AKT1, CAV2, CDH2, CTNNB1, FN1, FZD7, GNG11, GSK3B, IGFBP4, ILK, ITGA5, MAP1B, MITF, RGS2, SNAI1, SNAI2, SPARC, TCF4, TGFB1, TGFB2, TGFB3, TIMP1, TMEFF1, TSPAN13, VIM, VPS13A, WNT5A, AKT1, BMP2, BMP7, CTNNB1, DSP, ERBB3, F11R, FZD7, GSC, KRT14, MITF, MST1R, NODAL, NOTCH1, PTP4A1, SMAD2, TGFB2, TGFB3, TMEFF1, TWIST1, VCAN, WNT11, WNT5A, WNT5B, CTNNB1, PPP3R1, RAC1, SMAD2, SOX10, TGFB1, TGFB2, TGFB3, TWIST1, WNT11, WNT5A, AKT1, BMP2, BMP7, CAV2, CTNNB1, EGFR, ERBB3, FGFBP1, HIF1A, IGFBP4, ILK, MST1R, NODAL, PDGFRB, TGFB1, TGFB2, TGFB3, TIMP1, VCAN, CAV2, EGFR, FN1, ITGB1, MSN, MST1R, NODAL, PDGFRB, RAC1, STAT3, TGFB1, VIM, BMP2, BMP7, CDH1, CDH2, CTGF, CTNNB1, DSC2, EGFR, ERBB3, F11R, FN1, ILK, ITGA5, ITGAV, ITGB1, PTK2, RAC1, SPP1, TGFB1, TGFB2, TIMP1, VCAN, CAV2, ESR1, KRT19, TGFB3, NOTCH1, EGFR, ERBB3, PDGFRB, RGS2, SPARC, BMP2, BMP7, SMAD2, SMAD4, TGFB1, TGFB2, TGFB3, CCT3, CENPF, KCNK1, CCL14, GREB1, ADH1B, ANLN, RELN, RELA, AAMP, ABI3BP, ACHE, ACTG2, ACVR1, ACVR1C, ACVRL1, ADAM15, ADAM17, ADAM28, ADAM8, ADAM9, ADAMTS1, ADAMTS12, ADAMTS8, ADAP1, ADD1, ADM2, ADRA2B, AEBP1, AGGF1, AGR2, AGRN, AGT, AKAP12, AKAP2, AKT1, AKT2, AKT3, ALB, ALDOA, ALOX5, AMH, ANG, ANGPT1, ANGPT2, ANGPTL2, ANGPTL4, ANPEP, ANXA2, AP1M2, APC, APOD, APOE, APOH, AQP1, ARAP2, AREG, ARHGAP32, ARHGDIB, ASPN, ATP1F1, B3GNT3, BAD, BAG2, BAI1, BAI3, BCAS1, BGN, BICC1
SSTR2 ST14 STAB1 STAB2 STAT1 STAT3 SULF1 SV2B SYK SYNE1 TACSTD2 TAL1 TBX1 TBX4 TBXA2R TCEB1 TCEB2 TCF20 TCF3 TCF4 TDGF1 TEK TF TFDP1 TFPI2 THBS1 THBS2 THBS4 THY1 TIE1 TIMP1 TIMP2 TIMP4 TJP2 TJP3 TLR4 TMC6 TMEM100 TMEM30B Tmprss2 Tmprss4 Tmprss6 Tnc TNF Tnfrsf12a Tnfrsf1a Tnfrsf10 Tnfsf12 Tnfsf13 Tnmd Tnn Tns1 Tnxb Tom1l1 Tp53 Tpm2 Tp53b Tpsd1 Tshr Tspan1 Twist1 Twist2 Txnip Tymp Uba52 Uts2 Vamp8 Vash1 Vav2 Vav3 Vcam1 Vcan Vegfa Vegfb Vegfc Vezf1 Vhl Vim Vit Vpsi13a Vsig4 Vwa1 Vwa2 Wars Wipf1 Wnt5a Wnt5b Wwtr1 Zc3h12a Zcchc24 Zeb1 Zeb2 Zfpm2 Zfyve16 Zfyve9 Agk Ammecri1l Cc2d1b Cnot10 Cnot4 Cog7 Ddx50 Dhx16 Dnajc14 Edc3 Eif2b4 Ercc3 Cfc1 Gpatch3 Hdac3 Mrp5 Mtrm14 Nol7 Nubp1 Prpf38a Sap130 Sf3a3 Tlk2 Tmub2 Trim39 Usp39 Zc3h14 Zkscan5 Znf143 Znf346, Cdh2, Fn1, Focx2, Gng11, Gsc, Igbp4, Itga5, Itgav, Msn, Sox10, Sparc, Steap1, Tcf4, Timp1, Tmeff1, Tmem132a, Twist1, Vcan, Vim, Vpsi13a, Wnt5a, Wnt5b, Cav2, Cdhi, Dsp, Fgfbp1, Il1rn, Krt19, Mitf, Mst1r, Nudt13, Rgs2, Spp1, Tfp12, Tspan13, Akt1, Cav2, Cdhi2 Ctnnb1, Fn1, Fzd7, Gng11, Gsk3b, Igbp4, Ilk, Itga5, Map1b, Mitf, Rgs2, Sparc, Tcf4, Tgfb1, Tgfb2, Tgfb3, Timp1, Tmeff1, Tspan13, Vim, Vpsi13a, Wnt5a, Akt1, Bbmp2, Bmp7ctnnb1, Dsp, Erbb3, F11r, Focx2, Fzd7, Gsc, Krt14, Mitf, Mst1r, Nodal, Notch1, Ptp4a1, Smad2, Sox10, Tgfb2, Tgfb3, Tmeff1, Twist1, Vcan, Wnt11, Wnt5a, Wnt5b, Ctnnb1, Focx2, Ppp3r1, Rac1, Smad2, Sox10, Tgfb1, Tgfb2, Tgfb3, Twist1, Wnt11, Wnt5a, Akt1, Bmp2, Bmp7, Cav2, Ctnnb1, Egfr, Erbb3, Fgfbp1, Focx2, Hif1a, Igbp4, Ilk, Mst1r, Nodal, Pdgfrb, Timp1, Vcan, Cav2, Egfr, Fn1, Itgbi, Msn, Mst1r, Nodal, Pdgfrb, Rac1, Stat3, Tgfb1, Vim, Bmp2, Bmp7, Cdhi, Cdhi2, Col5a2, Ctgf, Ctnnb1, Dsc2, Egfr, Erbb3, F11r, Fn1, Focx2, Ilk, Itga5, Itgav, Itgbi, Mmp3, Ptk2, Rac1, Spp1, Tgfb1, Tgfb2,
TIMP1, VCAN, CAV2, ESR1, KRT19, TGFβ3, PDGFRβ, RGS2, SPARC, SMAD4, CCT3, CENPF, KCNK1, CCL14, GREB1, ADH1B, ANLN, RELN, PCLO, CSMD3, MUC16, ST18, USH2A, AMPH, FAT4, PDE1C, SHANK1, ABCA13, APOB, THSD7B, CYP2C8, DNAH10, SEMA3D, IL1RAPL1
Appendix 5 Nanostring Pancancer Progression Panel Gene List

ZFYVE9, ZFYVE16, ZFPM2, ZEB2, ZEB1, ZCCHC24, 2C3H12A, WWTR1, WNT5B, WNT5A, WIPF1, WARS, VWA2, VWA1, V5IG4, VPS13A, VIT, VIM, VHL, VEZ1F, VEGFC, VEGFB, VEGFA, VCAN, VCAM1, VAV3, VAV2, VASH1, VAMP8, LTS2, UBA52, TYMP, TXNIP, TWIST2, TWIST1, TSPAN1, TSHR, TPSD1, TPSB2, TPM2, TPS3, TOM1L1, TNXB, TNS1, TNN, TMND, TNSF13, TNSF12, TNSF10, TNSRF1A, TNSRF1A2, TNC, TMCPR55, TMRPS54, TMRPR25, TMEM30B, TMEM100, TMC6, TLR4, TP3, TP2, TIMP4, TIMP2, TIMP1, TIE1, THY1, THBS4, THBS2, THBS1, TGFB2, TGFB1, TGFBI, TFPI2, TFDP1, TF1, TGF1, TCEF, TCF3, TCF20, TCEB2, TCEB1, TXBA2R, TXB4, TXB1, TAL1, TACSTD2, SYNE1, SYK, SYB, SULF1, STAT3, STAT1, STAB2, STAB1, ST14, SSTR2, SRPX2, SRPK2, SRGB, SRF, SRC, SPP1, SPOCK3, SPINT1, SPINK5, SLPK2, SPDEF, SPARCL1, SPAR, SP1, SOX9, SOX2, SOX17, SORD, SOD1, SNRPF, SNAI3, SNAI2, SNAI1, SMURF2, SMURF1, SMOC1, SMCO1, SMAD9, SMAD5, SMAD4, SMAD3, SMAD2, SMAD1, SLPI, SLIT2, SLC4A4A, SLC37A1, SLC35A3, SLC2A1, SLC12A6, SKP1, SIRT1, SHB, SH3YL1, SH2D3A, SH2B3, SFRP2, SFRP1, SETD2, SET, SERPINH1, SERPIN1, SERPINF1, SERPINE1, SERPINA1, SERINC5, SEMA3E, SELC4, SCN11A, SCG2, SAMSN1, SACS, S1PR1, S100A7, S100A14, RUNX1T1, RUNX1, RTN4, RAAS, RPS6KB2, RPS27A, RORB, RORA, ROCK2, ROCK1, ROBO4, RHOAI, RHOA, RGGC, RELN, RBX1, RBPI, RBM47, RBL2, RBL1, RB1, RAMP2, RAMP1, RAF1, RAC2, RAC1, RAB25, QKI, PYCA, RD, PXDN, PTX3, PTTG1, PTF8, PTPRM, PTPBR, PTPB, PTK6, PTK2B, PTK2G, PGTSG, PGTDS, PTE, PRSS8, PRSS2R, PRSS22, PRSS15L, PROM1, PROK2, PRKCKZ, PRKCG, PRKCB, PRF1, PRELP, PPP3R1, PPP2R1A, PPP2CB, PPP1R16B, PPL, PFPB2, POSTN, POPDC3, PNPLA6, PMP22, PLXND1, PLXNCL, PL, S1, PLEKH01, PLCG2, PLCG1, PLAUR, PLA2G3, PLA2G2D, PLA2G2A, PLA2G10, PKNOX1, PKN1, PKM, PITU2, PIK3R2, PIK3R5, PIK3R2, PIK3R1, PIK3CG, PIK3CD, PIK3CA, PGK1, PKFB4, PKFB1, PECAM1, PEAP4, PDPN, PDK1, PDGFRB, PDGFC, PDGA, PDA, PDCD10, PCOLCE, P3H2, P3H1, Ovol2, OLFM12B, OGN, OCLN, OAS1, NTRK1, NRRX3, NRRX1, NRP2, NRP1, NRCAM, NR4A3, NR4A1, NR4C1, NR1X5, NOTCH1, NOS3, NODAL, NME4, NME1, NID2, NFKB1, NFATC2, NFAT5, NF2, NF1, NDRG1, NDF, NDFD, NCL, NCAM1, NAP1L3, NAA15, MYOC, MYO1D, MYLK, MYH11, MYCL1, MYC, MUC1, MTOC, MTDH, MIBP4, MTA1, MT3, MS4A6A, MS4A4A, MRC1, MPD2, MMRN2, MMP9, MMP3, MMP4, MMP2, MMP17, MMP14, MMP13, MMP10, MMP1, MISP, MGP, MGAT5, MFA4, MTF1, MEOX2, MEG3, MED23, MED1, MCM, MANKAPK3, MAP4K3, MAPK1, MAP3K7, MAP2K4, MAP2K2, MAP2K1, MAP2A, MA, YLF96, LUM, LTB4, LRG1, LOX1, LOX2, LOX, LLLG2, LIFR, LHFP, LGALS1, LEFTY1, LDAH, LAMC2, LAMC1, LAMB3, LAMA5, LAMA4, LAMA3, LAMA1, LAD1, KRT7, KRT19, KRT14, KRT1, KRT11, KASL, KIS1, KIAA1462, KDR, KDM1A, KCNN8, JUN, JAM3, JAM2, JAG1, ITM2A, ITGB8, ITG7, ITGB6, ITGB4, ITGB3, ITGB2, ITGB1B1, ITGAM, ITGA9, ITGA8, ITGA7, ITGA6, ITGA5, ITGA3, ITGA2, ITGA11, ITGA1, ISL1, ISL1, IRF6, INHBE, INHBA, ILK, IL6, IL1RN, IL1R1, IL1B, IL1A, IL1B, IL1A, IL1B, IL1A1, IL1B1, IL1R1, IL1A1, IL1B1, IL1A1, IL1B1, IL1B2, IL1A2, IL1A2, IL1A2, IL1R1, IGFBP7, IGFBP4, GF1, FNG, ID4, ID4, ID2, ID1, ID1, ICAM1, IBS, HUNK, HSPG2, HSPB1, HSP90B1, HSD17B12, HRAS, HPSE, HOXB3, HOXB3, HOXA7, HOXA5, HMOX1, HLA-DBP1, HKDK1, HK3, HK2, HIPK2, HIPK1, HIP1A, HGF, HEG1, HDH3, HDACE3, HSA1, HAPLN1, GZMK, GF2, GSN, GRHL2, GREM1, GPRX1, GPR56, GPR124, GPI, GLYR1, GJAX, GIMAP6, GIMAP4, GDF6, GDF5, GDF15, GATA4, GALT7, GXYD6, FUT3, FSTL1, FST, FREGM, FREGM1, FRAS1, FOXO4, FOXC2, FN1, FMD, FLT4, FLT1, FL1, FIGF, FHL1, FLG2, FLG4, FLG3, FLG4, FLG2, FGFR1, FGFF9, FGFF, FGFB, FGFE, FERMT2, FBP1, FBN2, FBN1, FBLN5, FBN1, FASLG, FAP, FAM174B, F3, F11R, EVP1, EVP2, EVP4, ESRF1, ERMP1, EREG, ERBB3, ERBB2IP, ERBB2, EPBSL1, EPNS1, EPHB4, EPHB3, EPHB1, EPHA1, EFCAM, EPA1, EP300, ENPP2, ENPEP, EPO3, ENO3, EPO2, ENO1, EMP3, EMLIN3, EMLIN1, EMCN, ELK3, ELF3, EIF4E
BP1, EIF4E2, EIF2AK3, EGLN3, EGLN2, EGF, EGFL7, EGF, EDN1, ECSCR, ECM2, ECM1, DST, DSC2, DPY
SL3, DPT, DLL4, DLG1, DLC1, Dicer1, DESI1, DENR, DENND5A, DDR2, DCP, DCC, DAG1, CYP1B1, CYBB
, CYB561, CXCR4, CXCR3, CXCL8, CXCL17, CXCL13, CXCL12, CXCL11, CXADR, CX3CL1, CUL1, CTSK, CTSN1, CTSN1, CTNNB1, CST7, CSPG4, CSF2RB, CRISPLD2, CRIP2, CREBBP,
COMP, COL7A1, COL6A3, COL6A2, COL6A1, COL5A2, COL5A1, COL4A6, COL4A2, COL4A1, COL3A1,
COL1A2, COL1A1, COL1A1, COL1A1, CNN1, CMA1, CLU, CLIC4, CLEC3B, CLEC2B, CLDN7, CLDN4, CLDN3, CLDN1, CKMT1A, CIB1, CCHR7, CHRD1L, CHP2, CHP1, CHI3L1, CHD4, CHAD, CGN, CFP, CEP295, CEP170,
, CEACAM6, CEACAM5, CEACAM1, CDS1, CDKN2A, CDKN1A, CDK14, CDH2, CDH13, CDH11, CDH1, C
DC42, CD82, CD46, CD44, CD36, CD34, CD2AP, CD24, CD163, CCR3, CCR2, CCL8, CCL7, CCL5, CCL21, C
CL11, CCDC80, CCBE1, CBLC, CAV1, CASP8, CAMP, CAMK2D, CAMK2B, CAMK2A, CALD1, CALCRL, CAD
M1, C3AR1, C3, C1S, BTG1, BRMS1, BNC2, BMPR2, BMPR1B, BMPR1A, BMP7, BMP5, BMP
4, BICC1, BGN, BCAS1, BAI3, BAI3, BAG2, BAD, B3GNT3, ATP1F1, ASPN, ARHGID1, ARHGAP32, AREG,
ARAP2, AQP1, APOH, APOE, APOD, AP6, AP1M2, ANXA2, ANPEP, ANGPTL4, ANGPTL2, ANGPT2, ANGPT1, ANG, AMH, ALOX5, ALDOA, ALB, AKT3, AKT2, AKT1, AKAP2, AKAP12, AHNAC, AGT, AGRN, AGK
2, AGGF1, AEBP1, ADRA2B, ADMA, ADD1, ADAP1, ADAMTS8, ADAMTS12, ADAMTS1, ADAM9, ADAM8, ADAM28, ADAM17, ADAM15, ACVR1L, ACVR1C, ACVR1, ACTG2, ACHE, AAMP