The Genomics of Drug Response in Head and Neck Cancer

by

Ren X. Sun

A thesis submitted in conformity with the requirements for the degree of Master of Science

Graduate Department of Pharmacology and Toxicology
University of Toronto

© Copyright by Ren X. Sun, 2017
Head and neck cancer (HNC) is the sixth most common malignancy worldwide with a consistently poor five-year survival rate of $\sim 50\%$. A major contributor to these poor clinical outcomes is the broad scope of therapeutic guidelines that fail to address underlying molecular heterogeneities. To elucidate genomic correlates of drug sensitivity for a targeted approach, we characterized the molecular profiles of 33 HNC cell lines for their somatic single nucleotide variants (SNVs) and copy number aberrations (CNAs). In this process, we developed a novel and reliable method for calling SNVs in cell lines with high precision and recall. Using a selected panel of drugs derived from high-throughput screening, we computed half maximal inhibitory concentration values and activity areas and correlated these sensitivity metrics with recurrent genomic features. We found several candidate CNA and SNV predictors of drug response of interest and demonstrate the feasibility of pharmacogenomics studies for HNC.
Acknowledgements

I want to start off by expressing my deepest of thanks to my supervisor, Dr. Paul Boutros. These past years have been nothing short of humbling and inspirational. In a time when the reality of doing science involves money, ambition and competition across so many avenues, his sincere dedication and firm stance on doing honest science and passion for doing it for the right reasons have been enlightening. I am truly privileged to have been a receiver of his guidance, an audience to his wisdom and am unendingly grateful for his support, devotion and patience throughout. Thank you, for taking a chance on that struggling, inexperienced undergrad all those years ago.

I want to extend my gratitude to all members of the Boutros Lab, past and present. Very few are fortunate in life to meet a handful of truly amazing people and I’ve been blessed with knowing more than three dozen. Each and every one of them deserves so much more than what I can convey in words and without their continuous support, insights and friendship for the past years, I wouldn’t be where I am today. Especially, I want to thank Connie Li and Natalie Fox – for cheering me on during times of doubt, for putting up with me during times of cynicism and for celebrating with me during times of triumph – and Cindy Yao – for teaching me everything I know. Thank you guys, for turning all my lowlights into highlights.

I want to thank my advisor, Dr. Denis Grant for his help, encouragement, kindness and forgiveness when I’ve given short notices. I also want to acknowledge Dr. Anthony Nichol’s lab in London, Ontario, for being such patient and generous people and for the extraordinary collaboration that resulted this project – and Morgan Black in particular, for generating the drug screening results.

Finally and most importantly, I want to thank my mom. She is the strongest person I know and the epitome of who I want to be as a person. For longer than I can remember and longer more, I’ve aspired to embody her work ethics and uphold her moral principles. She’s seen me at my best and worst, and cared for and loved me through all of it. This and all future endeavours are dedicated to her. Thank you mom, for teaching me, for loving me. For you.
## Contents

1 Introduction .................................................. 1
   1.1 Head and Neck Cancer ................................... 1
       1.1.1 Overview .............................................. 1
       1.1.2 Risk Factors ......................................... 2
       1.1.3 Molecular Heterogeneity .............................. 4
       1.1.4 Diagnosis, Treatment and Challenges ................ 5
   1.2 Next Generation Sequencing ............................... 8
       1.2.1 Analysis .............................................. 10
   1.3 Drug Screening and Discovery ............................ 11
       1.3.1 Cancer Cell Lines .................................... 12
   1.4 Biomarkers ................................................. 13
   1.5 Rationale and Aims ....................................... 15

2 Materials and Methods ......................................... 17
   2.1 Genomics .................................................. 17
       2.1.1 Copy Number Profiles ................................. 17
       2.1.2 Reference-free SNV Classifier: S22S ............... 19
       2.1.3 Mutational Profiles ................................... 32
   2.2 Drug Screening ............................................ 34
       2.2.1 Drug Selection ....................................... 34
       2.2.2 Drug Sensitivity ..................................... 35
       2.2.3 Statistical Testing ................................... 38
       2.2.4 Sensitivity Modelling ................................ 38
       2.2.5 Power Analysis ....................................... 41
       2.2.6 Visualization ........................................ 42

3 Results ....................................................... 43
   3.1 Overview .................................................. 43
   3.2 Copy Number Analysis .................................... 43
       3.2.1 Experimental Design ................................. 43
3.2.2 Ploidy Estimation and Adjustments ........................................ 45
3.2.3 Consensus Clustering ......................................................... 46
3.2.4 Copy Number Profiles ....................................................... 48
3.2.5 Significant Regions .......................................................... 51
3.3 SNV Profile ............................................................................ 53
  3.3.1 Reference free SNV Classifier: S22S ..................................... 53
  3.3.2 Recurrent SNVs of HNC Cell Lines ...................................... 62
3.4 Drug Response ....................................................................... 64
  3.4.1 Sensitivity Assessment ......................................................... 66
  3.4.2 Statistical Testing ............................................................... 69
  3.4.3 Sensitivity Modelling and Model Power ............................... 71
  3.4.4 Power Analysis: Wilcoxon Test .......................................... 73

4 Discussion .................................................................................. 76
  4.1 Importance ............................................................................ 78
    4.1.1 Resource ....................................................................... 78
    4.1.2 S22S ........................................................................... 78
  4.2 Limitations ............................................................................ 79
  4.3 Future Directions .................................................................. 80

Appendix 1: Accurate Reference-Free Somatic Variant-Calling by Integrating Genomic, Sequencing and Population Data .................................................. 82

Appendix 2: Supplementary Materials ................................................. 101
  1 OncoScan Profiles ................................................................. 101
  2 ASCAT Profiles ..................................................................... 117
  3 Consensus Clustering .............................................................. 133
  4 Drug Profiling ....................................................................... 137
  5 Drug Modelling ..................................................................... 140
  6 Power Analysis ...................................................................... 142
    6.1 Activity Areas .................................................................. 142
    6.2 IC<sub>50</sub> ...................................................................... 169
  7 Tables ...................................................................................... 196
List of Figures

1.1 SEER Incidence of Oral Cancers ............................................. 3
1.2 Prognosis based on HPV status .............................................. 6

2.1 Procedure Pipeline ................................................................. 20

3.1 Cell Lines-Technology Overlap .............................................. 44
3.2 HPV-positive Samples Overlap .............................................. 44
3.3 OncoScan: RF15B ................................................................. 46
3.4 OncoScan: RF22B ................................................................. 47
3.5 Consensus Matrix ................................................................. 48
3.6 CDF .................................................................................. 49
3.7 Delta CDF .......................................................................... 49
3.8 CNA Profile: HNC Cell Lines ............................................... 50
3.9 Significant CNAs ................................................................. 51
3.10 Overlap of SNV calling methods ......................................... 55
3.11 S22S Performance ............................................................... 57
3.12 S22S vs. MuTect on Significant Genes from TCGA Study ........ 58
3.13 S22S vs. GATK on Significant Genes from TCGA Study .......... 59
3.14 Effect of Training Size .......................................................... 60
3.15 Model Variable Importance ............................................... 61
3.16 Recurrent SNVs: Full ........................................................... 63
3.17 Recurrent SNVs: Selected ..................................................... 64
3.18 Model Goodness of Fit ......................................................... 66
3.19 Cell Lines per Drug ............................................................... 67
3.20 Overview: Activity Area ......................................................... 68
3.21 Activity Area: p-value distribution ....................................... 69
3.22 IC_{50}: p-value distribution ................................................. 70
3.23 Response Category: p-value distribution ............................. 71
3.24 Modelling: Activity Areas .................................................... 72
3.25 Wilcoxon Power: AZD4547 .................................................. 74
3.26 Wilcoxon Power: SCS ................................................. 75

S1 OncoScan: Ref103 .................................................. 101
S2 OncoScan: 93VU147T .............................................. 102
S3 OncoScan: BICR56 .................................................. 102
S4 OncoScan: Cal27 .................................................... 103
S5 OncoScan: Cal33 .................................................... 103
S6 OncoScan: Detroit562 ............................................ 104
S7 OncoScan: FaDu .................................................... 104
S8 OncoScan: HMS001 ............................................... 105
S9 OncoScan: HSC2 .................................................... 105
S10 OncoScan: JHU006 ............................................... 106
S11 OncoScan: JHU011 ............................................... 106
S12 OncoScan: JHU029 ............................................... 107
S13 OncoScan: PCI6A ............................................... 107
S14 OncoScan: PCI6B ............................................... 108
S15 OncoScan: PCI13 ............................................... 108
S16 OncoScan: PCI30 ............................................... 109
S17 OncoScan: PE/CA:PJ-49 ........................................ 109
S18 OncoScan: RF15A ............................................... 110
S19 OncoScan: RF15B ............................................... 110
S20 OncoScan: RF22A ............................................... 111
S21 OncoScan: RF22B ............................................... 111
S22 OncoScan: RF37A ............................................... 112
S23 OncoScan: RF37B ............................................... 112
S24 OncoScan: SCC4 ............................................... 113
S25 OncoScan: SCC9 ............................................... 113
S26 OncoScan: SCC15 .............................................. 114
S27 OncoScan: SCC25 .............................................. 114
S28 OncoScan: SCC47 .............................................. 115
S29 OncoScan: UPCI:SCC090 ...................................... 115
S30 OncoScan: UPCI:SCC152 ...................................... 116
S31 OncoScan: UPCI:SCC154 ...................................... 116
S32 ASCAT: Ref103 .................................................. 117
S101 Wilcoxon Power: Pyrrolidine . . . . . . . . . . . . . . . . . . . . . . . . . . . 161
S102 Wilcoxon Power: Rapamycin . . . . . . . . . . . . . . . . . . . . . . . . . . . 162
S103 Wilcoxon Power: RDEA119 . . . . . . . . . . . . . . . . . . . . . . . . . . . 163
S104 Wilcoxon Power: Ro31-8220-mesylate . . . . . . . . . . . . . . . . . . . . . 164
S105 Wilcoxon Power: Ryuvidine . . . . . . . . . . . . . . . . . . . . . . . . . . . 165
S106 Wilcoxon Power: SCH79797 . . . . . . . . . . . . . . . . . . . . . . . . . . . 166
S107 Wilcoxon Power: SP600125 . . . . . . . . . . . . . . . . . . . . . . . . . . . 167
S108 Wilcoxon Power: TCS2312-dihydrochloride . . . . . . . . . . . . . . . . . . 168
S109 Wilcoxon Power: A-443654 . . . . . . . . . . . . . . . . . . . . . . . . . . . 169
S110 Wilcoxon Power: Atopaxar . . . . . . . . . . . . . . . . . . . . . . . . . . . 170
S111 Wilcoxon Power: BEZ-235 . . . . . . . . . . . . . . . . . . . . . . . . . . . 171
S112 Wilcoxon Power: BI78D3 . . . . . . . . . . . . . . . . . . . . . . . . . . . . 172
S113 Wilcoxon Power: BNTX maleate . . . . . . . . . . . . . . . . . . . . . . . . . 173
S114 Wilcoxon Power: Bosutinib . . . . . . . . . . . . . . . . . . . . . . . . . . . . 174
S115 Wilcoxon Power: BYL719 . . . . . . . . . . . . . . . . . . . . . . . . . . . . 175
S116 Wilcoxon Power: Cisplatin . . . . . . . . . . . . . . . . . . . . . . . . . . . . 176
S117 Wilcoxon Power: ER27319-maleate . . . . . . . . . . . . . . . . . . . . . . . 177
S118 Wilcoxon Power: Everolimus . . . . . . . . . . . . . . . . . . . . . . . . . . . 178
S119 Wilcoxon Power: Foretinib . . . . . . . . . . . . . . . . . . . . . . . . . . . . 179
S120 Wilcoxon Power: GDC-0941 bismesylate . . . . . . . . . . . . . . . . . . . . 180
S121 Wilcoxon Power: JTC801 . . . . . . . . . . . . . . . . . . . . . . . . . . . . 181
S122 Wilcoxon Power: MK-1775 . . . . . . . . . . . . . . . . . . . . . . . . . . . 182
S123 Wilcoxon Power: NSC95397 . . . . . . . . . . . . . . . . . . . . . . . . . . . 183
S124 Wilcoxon Power: NSC146109-hydrochloride . . . . . . . . . . . . . . . . . . 184
S125 Wilcoxon Power: OSI-906 . . . . . . . . . . . . . . . . . . . . . . . . . . . . 185
S126 Wilcoxon Power: Ouabain . . . . . . . . . . . . . . . . . . . . . . . . . . . . 186
S127 Wilcoxon Power: PRIMA-1 . . . . . . . . . . . . . . . . . . . . . . . . . . . . 187
S128 Wilcoxon Power: Pyrrolidine . . . . . . . . . . . . . . . . . . . . . . . . . . . 188
S129 Wilcoxon Power: Rapamycin . . . . . . . . . . . . . . . . . . . . . . . . . . . 189
S130 Wilcoxon Power: RDEA119 . . . . . . . . . . . . . . . . . . . . . . . . . . . 190
S131 Wilcoxon Power: Ro31-8220-mesylate . . . . . . . . . . . . . . . . . . . . . 191
S132 Wilcoxon Power: Ryuvidine . . . . . . . . . . . . . . . . . . . . . . . . . . . 192
S133 Wilcoxon Power: SCH79797 . . . . . . . . . . . . . . . . . . . . . . . . . . . 193
S134 Wilcoxon Power: SP600125 . . . . . . . . . . . . . . . . . . . . . . . . . . . 194
# List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>List of Model Features</td>
<td>27</td>
</tr>
<tr>
<td>3.1</td>
<td>Data Overview</td>
<td>43</td>
</tr>
<tr>
<td>3.2</td>
<td>OncoScan Quality Check</td>
<td>45</td>
</tr>
<tr>
<td>3.3</td>
<td>Significant CNAs</td>
<td>52</td>
</tr>
<tr>
<td>3.4</td>
<td>Significant Recurrent Genes</td>
<td>64</td>
</tr>
<tr>
<td>3.5</td>
<td>Drug Overview</td>
<td>65</td>
</tr>
<tr>
<td>3.6</td>
<td>Drug Response Overview</td>
<td>66</td>
</tr>
<tr>
<td>S1</td>
<td>Cell Line Characteristics</td>
<td>196</td>
</tr>
<tr>
<td>S2</td>
<td>Drug Status Overview</td>
<td>197</td>
</tr>
</tbody>
</table>
### List of Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Akt</td>
<td>(see PKB)</td>
</tr>
<tr>
<td>AA</td>
<td>Activity area</td>
</tr>
<tr>
<td>AIC</td>
<td>Akaike information criterion</td>
</tr>
<tr>
<td>ASCAT</td>
<td>Allele-specific copy number analysis of tumours</td>
</tr>
<tr>
<td>AUPRC</td>
<td>Area under the precision-recall curve</td>
</tr>
<tr>
<td>BAM</td>
<td>Binary alignment/map</td>
</tr>
<tr>
<td>BWA</td>
<td>Burrows-wheeler aligner</td>
</tr>
<tr>
<td>CDF</td>
<td>Cumulative distribution function</td>
</tr>
<tr>
<td>CN</td>
<td>Copy number</td>
</tr>
<tr>
<td>CNA</td>
<td>Copy number aberration</td>
</tr>
<tr>
<td>CV</td>
<td>Coefficient of variation</td>
</tr>
<tr>
<td>DNA</td>
<td>Deoxyribonucleic acid</td>
</tr>
<tr>
<td>FDA</td>
<td>Food and drug administration</td>
</tr>
<tr>
<td>FDR</td>
<td>False discovery rate</td>
</tr>
<tr>
<td>FN</td>
<td>False negative</td>
</tr>
<tr>
<td>FP</td>
<td>False positive</td>
</tr>
<tr>
<td>GATK</td>
<td>Genome analysis toolkit</td>
</tr>
<tr>
<td>GISTIC</td>
<td>Genomic identification of significant targets in cancer</td>
</tr>
<tr>
<td>GLM</td>
<td>Generalized linear model</td>
</tr>
<tr>
<td>GOF</td>
<td>Goodness of fit</td>
</tr>
<tr>
<td>HNC</td>
<td>Head and neck cancer</td>
</tr>
<tr>
<td>HNSC</td>
<td>Head and neck squamous cell carcinoma</td>
</tr>
<tr>
<td>HNSCC</td>
<td>Head and neck squamous cell carcinoma</td>
</tr>
<tr>
<td>IGF-1R</td>
<td>Insulin-like growth factor 1 receptor</td>
</tr>
<tr>
<td>IR</td>
<td>Insulin receptor</td>
</tr>
<tr>
<td>JNK</td>
<td>Jun N-terminal kinase</td>
</tr>
<tr>
<td>LOOCV</td>
<td>Leave one out cross validation</td>
</tr>
<tr>
<td>PAR1</td>
<td>Protease-activated receptor 1</td>
</tr>
<tr>
<td>PCR</td>
<td>Polymerase chain reaction</td>
</tr>
<tr>
<td>PKB</td>
<td>Protein kinase B (a.k.a Akt)</td>
</tr>
<tr>
<td>PKC</td>
<td>Protein kinase C</td>
</tr>
<tr>
<td>PoN</td>
<td>Panel of normals</td>
</tr>
<tr>
<td>P&lt;sub&gt;x&lt;/sub&gt;</td>
<td>Prescription</td>
</tr>
<tr>
<td>PRAD</td>
<td>Prostate adenocarcinoma</td>
</tr>
<tr>
<td>RF</td>
<td>Random forest</td>
</tr>
<tr>
<td>RNA</td>
<td>Ribonucleic acid</td>
</tr>
<tr>
<td>SNP</td>
<td>Single nucleotide polymorphism</td>
</tr>
<tr>
<td>SNV</td>
<td>Single nucleotide variant</td>
</tr>
<tr>
<td>Acronym</td>
<td>Definition</td>
</tr>
<tr>
<td>---------</td>
<td>------------</td>
</tr>
<tr>
<td>TCGA</td>
<td>The cancer genome atlas</td>
</tr>
<tr>
<td>TN</td>
<td>True negative</td>
</tr>
<tr>
<td>TNM</td>
<td>Tumour, nodes, metastases</td>
</tr>
<tr>
<td>TP</td>
<td>True positive</td>
</tr>
<tr>
<td>TS</td>
<td>Thymidine synthase</td>
</tr>
<tr>
<td>VCF</td>
<td>Variant call format</td>
</tr>
<tr>
<td>VNC</td>
<td>Virtual normal correction</td>
</tr>
</tbody>
</table>
Introduction

Head and Neck Cancer

Overview

Head and neck cancer (HNC) is a broad term used to describe epithelial malignancies arising from mucosal cells of the oral cavity, oropharynx, larynx, hypopharynx, paranasal sinuses and salivary glands [1]. Worldwide, it ranks sixth amongst tumour types in terms of incidence and eighth with respect to mortality [2]. While site-specific survival rates range anywhere from 20-90%, on average, the five-year survival for HNC is only 40-50% [3, 4]. Since HNC is a cancer affecting multiple subsites, cancer statistics and estimates tend to vary depending on the sites included in these calculations. In Canada, aggregating the most recent statistics across most affected areas, it can be estimated that ∼ 6,600 new HNC cases emerge annually with ∼ 2000 deaths occurring within the year [5]. Approximately two-thirds of newly-diagnosed patients are male; this sex bias is also reflected in the disease-specific mortality for HNC [5]. However, less commonly occurring cancers – such as those originating in the hypopharynx and nasal or paranasal sinuses – are not readily reported nor separately recorded and due to their inclusion in general categories of “other” or “unspecific” cancers, many statistical projections are actually underestimating the societal impact of HNCs [5].

The most common histological subtype of HNC is squamous cell carcinoma (SCC), which not only describes most cancers occurring in the oral cavity, oropharynx, hypopharynx and larynx, but also makes up the majority of HNCs [6]. Hence the the term head and
neck squamous cell carcinoma (HNSCC) is often used interchangeably with HNC, but more specifically to imply cancers that originate from these anatomical sites.

**Risk Factors**

Historically, the factors which impose the most risk for HNC are tobacco use and alcohol consumption, where the combination of the two actually behave more synergistically than additively \[7-10\]. Until roughly 15 years ago, incidence, prevalence and site distributions of HNCs were largely stable, where the most common types of malignancies were cancers of the oral cavity and larynx \[11\]. In fact, there has been a steady and declining trend in the prevalence of tumours at most oral sites since the early 1980s into the early 2000s – much of this attributed to better health practices and lower rates of smoking (Figure 1.1) \[12\]. However in the last decade, an unexpected surge in oropharyngeal cancers has been observed, a large fraction of these cases linked to infection by high-risk types of the human papillomavirus (HPV) (Figure 1.1) \[13-16\]. This rise in cancers of the oropharynx has been the primary factor responsible for the upward trend observed in oral cancer incidence since the early 2000s, and this recent emergence of HPV as a causal factor in the development of HNC has altered the molecular landscape, epidemiology as well as survival rates of HNCs.

**Human Papillomavirus**

HPVs are small, circular DNA viruses that are part of the *Papillomaviridae* family and are known to target basal cells of epithelial mucosa \[17-19\]. There are currently over 200 types of HPVs documented (www.hpvcenter.se accessed on 2016-09-01) and each genotype is numerically designated. HPVs are often grouped based on the types of cells they infect as well as how easily they can undergo cellular transformation – this ability and its efficiency stratifying HPVs into low-risk or high-risk types \[12\] \[19\] \[20\]. For example, HPV-1 infects cutaneous epithelial cells while HPV types 6, 11, 16 and 18 infect epithelial mucosa of the
Figure 1.1: Incidence of oral cancers (subsites included are the oral cavity, nasopharynx, hypopharynx and oropharynx) compared to just oropharyngeal cancers, from the Surveillance, Epidemiology and End Results (SEER) database, 1975-2012.

It has long been shown that HPV is a causal factor in cervical cancer – although it has also been associated with various other cancers of the uro-genital region – where $\sim 70\%$ of all cases are caused by types 16 and 18, while the remainder of cases are attributed to types 31, 33, 45 and to a much lesser extent, types 35, 39, 51, amongst others [21][25]. In HNC, molecular studies have shown that HPV-positive cancers are predominately caused by HPV-
16 (~ 90%), with a small subset caused by HPV-18, while other oncogenic types were rarely detected (<1%) [12, 20, 26–28]. HPV-16 promotes carcinogenesis through two oncogenes, E6 and E7, which inactivate p53 and retinoblastoma (RB) respectively and in turn disturbs regulation of the cell cycle in infected cells [4]. The E6 protein binds p53 and targets it for proteosomal degradation, resulting in loss of p53-mediated apoptosis as well as dysregulation of the G1/S checkpoint [29–31]. Due to the inactivation of p53 by HPV-16 E6, TP53 often remains wild-type in HPV-positive tumours, whereas most HPV-negative tumours harbour some form of inactivating mutation [32]. In a similar fashion, binding of the E7 protein to pRB catalyzes the eventual degradation of the RB protein and results in the release of E2F, a transcription factor that promotes cell proliferation [33]. Functional inactivation of pRB causes upregulation and overexpression of cyclin-dependent kinase inhibitor 2A, CDKN2A (also known as p16INK4a), which is often used as a surrogate marker for HPV infection [34, 35]. This surrogate marker, while more cost-effective, is unfortunately not 100% accurate [36, 37].

**Molecular Heterogeneity**

Since the vast majority of HNCs (~ 95%) are SCCs, it was previously regarded as a relatively homogeneous disease relative to other tumour types. However, recent studies have shown that a large degree of heterogeneity exists across HNSCCs with respect to carcinogenesis, molecular profiles, pathobiology and prognosis. First and foremost, unlike other types of cancer, HNCs encompass cancers that originate from many different sites and thus, two cancers with the same diagnosis may easily differ at the tissue-level and in the simplest manner, subtypes can be seen histologically [38]. Even for tissues that are morphologically similar, distinct molecular subgroups are present [39–41]. Karyotyping and ploidy analyses have revealed that although a subgroup of tumours are diploid or near-diploid, the majority of HNCs are aneuploid [42, 43]. Studies using genomics technologies such as RNA and
DNA profiling have further demonstrated that molecular subgroups also exhibit differential prognoses. In particular, subgroups with elevated epidermal growth factor receptor (EGFR) mRNA abundance tend to present with poorer survival rates [39, 44]. Contrastingly, a subgroup of salivary cancers with hotspot activating mutations in PRKD1 actually present more favourable survival rates [45].

Another prognostic subgroup is HPV status (Figure 1.2): owing to etiological differences where HPV-negative tumours often accumulate many mutations in TP53 whereby in HPV-positive tumours, p53 inactivation is functional and thus alterations are rarely seen at the gene level [46, 48]. This is mirrored in the genetic and epigenetic inactivation of CDKN2A, which is frequently observed for many HPV-negative HNCs but at a much lower frequency for HPV-positive ones [49, 50]. Loss of either TP53 and CDKN2A as well as a negative HPV status have been correlated with poorer clinical outcome [40, 51–55] Additional differences that emerge between these tumours are also likely factors of HPV-inactivation of TP53 being an initiating event in HPV-positive tumours while for HPV-negative ones, this event often occurs later [46–48]. Furthermore, it is worth noting that HPV-negative tumours also gain an additional set of mutations that are common and characteristic of smoking – a traditional risk factor – and biological age – a function of its prevalence in older populations [56]. Currently, HPV status is only taken into consideration by individual physicians but not clearly defined in standard practice. On the whole, while efforts are being made for more a precise classification system, it is clear that treatment approaches need to be tailored to better reflect the underlying molecular diversities of HNCs [57].

Diagnosis, Treatment and Challenges

The first signs of HNC are usually detected during physical or oral examinations [58]. Several endoscopic or image-based tests (e.g. computed tomography (CT), x-rays, magnetic
resonance imaging (MRI), etc.) exist to facilitate this process [59]. More modern aids, such as brush biopsy, toluidine blue staining, auto-fluorescence, spectroscopy and a handful of omics-based biomarkers have also been introduced in the past decade, although many of these have yet to be integrated into standard practice [58, 60–64]. Thus scalpel biopsy followed by histopathological analysis remain the gold standard for clinical diagnosis [58]. Following a positive diagnosis, standard staging takes place to characterize the size and spread of the tumour. Staging is done via the Tumour, Node, Metastasis (TNM) classification system, where the T-category describes the size and site of the primary tumour; the N-category, the extent of nodes affected; and the M-category, the presence or absence of metastases [65, 66]. Clinical staging categories are made at the initial time of diagnosis and are typically
based on the results of image-based tests, whereas pathological versions are derived from the examination of the actual tumour by a pathologist if it is removed [65][67]. Lower values in each category strongly predict better prognosis [65][66].

TNM classification and other patient-specific characteristics, such as age, swallowing ability, tumour resectability, organ preservation, are amongst the important factors that drive therapeutic decision-making. Overall, treatment of HNC is extremely complex and often involves a team of specialists, including surgeons, medical and radiation oncologists, pathologists, radiologist, plastic surgeons and dentist surgeons [3]. Surgery and radiotherapy have long been the cornerstone approaches – and surgery remains the preferred method for early-stage HNC – but for locally-advanced or recurrent/metastatic disease, studies have shown that a combination of surgery, chemotherapy and radiotherapy (either concurrent or post-operative) are vital for improving overall survival [68][71]. Concurrent chemoradiotherapy (CRT) is also effective in cases where tumours are inoperable, often either due to tumour site or size [72][73]. In specific cases, complementary targeted strategies have been shown to improve treatment efficacy. As previously mentioned, a subgroup of HNCs possess elevated EGFR expression, which has been associated with poorer prognosis [39]. Monoclonal antibodies to EGFR (e.g. cetuximab) was demonstrated to enhance radiotherapy response and significantly improve survival as well as locoregional control in a large-scale study [74][75]. This landmark trial provided evidence for approval by the Food and Drug Administration (FDA) in 2006 of cetuximab for treatment of advanced locoregional HNC either as monotherapy or concomitant with radiation [76].

More work is needed with respect to targeted agents founded on the underlying molecular heterogeneity of HNC, as different subgroups of tumours respond differently to treatment [55]. Unfortunately, monoclonal antibodies for EGFR remains, to date, the only successful and impactful translation of a targeted agent for treatment of HNC [76]. For the most
part, current therapeutic approaches resort to extreme measures for cancer eradication. This entails employing high intensity radiation or highly toxic systemic therapies, which have serious morbidities. Doxeltaxel, cisplatin and 5-fluorouracil (5-FU) are commonly-prescribed chemotherapeutic agents, each possessing its own set of toxicities. Different types of radiotherapy have been known to result in side effects that severely impact a patient’s quality of life [77]. Complicating the matter, chemo- and radioresistance readily arise, posing a barrier to successful treatment and contributing to a high recurrence rate of 50% within two years [78–81]. Subjecting patients with favourable prognoses to these harsh treatments is unreasonable and better stratification needs to be in place for clinical decision making.

**Next Generation Sequencing**

Prior to the mid-2000s, Sanger sequencing was the most widely used approach for DNA sequencing [82, 83]. First introduced by Frederick Sanger and colleagues in 1977, the method featured “terminating” dideoxynucleotides, which prevented DNA polymerase from further synthesis during replication *in vitro* [84, 85]. Fragments (or “reads”) of varying lengths could then be collected; the size of these reads used to elucidate the sequence of DNA [84, 85]. Since fewer toxic and radioactive compounds were required in comparison to other methods of the time, Sanger sequencing was considered more practical and was later implemented in the first generation of DNA sequencers [82, 83]. Thus, automated Sanger sequencing is viewed as a “first-generation” sequencing technology whereby newer platforms developed in the last decade are considered next generation sequencing (NGS) technologies [82]. These newer technologies are founded on parallel sequencing, which vastly increase throughput, thus enabling large quantities of data to be generated in much shorter time and at much lower costs [82, 83, 86, 87].

The traditional Sanger approach requires bacterial cloning, followed by the isolation of DNA
fragments that are then passed through a separate detection process (i.e. electrophoresis) – where the organization by length allows for delineation of base ordering [84, 85, 88]. With NGS, sequencing and detection can occur simultaneously. A variety of NGS platforms exists – each with its own set of preparatory requirements, biochemistries and detection methods – but to outline how NGS achieves its high yield in general principles, first, a collection of smaller fragments is constructed based on the target DNA [82, 83, 88, 89]. This collection, also known as a “library”, contains fragments (or “templates”) that each has a piece of synthetic DNA (“adapter”) covalently joined at the end [82, 83, 88, 89]. Adapters are utilized in subsequent amplification steps, conducted on a solid surface (such as glass slide or a microbead), where a polymerase-based reaction or equivalent step produces many copies of each fragment [82, 83, 86, 88, 89]. NGS instruments orchestrate a series of these automated base addition, detection and wash steps in a massively parallel fashion to produce hundreds of thousands of sequenced reads in one experimental run [82, 83, 86, 88, 89]. This rapid production of primary data has revolutionized how genomes are studied across many disciplines.

Despite the efficiency of NGS methods, challenges are still present. First, while amplification of DNA is necessary to facilitate detection, this step is a major source of sequencing error that is propagated downstream [90]. Error rates have been quoted to range anywhere from 0.1-15% depending on platform type [89, 90]. More difficulties are also found in data analysis – which remains a bottleneck to understanding genomes [89]. In NGS, noise accumulates with each new base that is added – this ultimately reaches a point where true signals from subsequent bases later in a sequence will be in competition with error signals accrued since the first cycle, putting a technological restriction on how long a read can be [88, 89]. While this constraint varies by technology, almost all current NGS platforms produce shorter reads in comparison to Sanger sequencing [88, 89]. Although the assembly of short reads using
shared sequence context is conceivable, the shortened overlap limits the span of assembly across a contiguous region \[88\]. Coupled with the repetitive content, size and complexity of the human genome, rebuilding a genome from NGS data would not have been possible without additional innovation in long-read sequencing as well as a myriad of computational approaches. One of such fundamental improvements has been the introduction of “paired-end” sequencing, where sequence data from both ends of a library fragment are produced, enhancing the overall certainty of read placement \[88, 91\].

**Analysis**

Most importantly, high-quality reference genomes for model organisms, such as humans, already exist, thus a logical approach to the assembly of short reads in such cases would be to align them to the reference genome. A large number of computational approaches have since been devised for such purposes, these typically using score-based metrics for identifying the genomic position of best fit \[92, 93\]. The conventional output of these algorithms is one of two formats, Sequence Alignment/Map (SAM) or Binary Alignment/Map (BAM), where an indexed file in BAM format is typically passed down for further processing.

Prior to analysis of NGS data, quality control is the next important step with nascent alignments, as short reads can often be misaligned if an insertion/deletion (indel) is present \[94\]. If these errors are left uncorrected, rates of false positives for subsequent SNV discovery are unusually high \[95\]. Several tools are commonly used for this purpose and while a gold standard has yet to be established, the community at large has been using the Genome Analysis Toolkit (GATK) since its creation by the Broad Institute in 2011 \[95, 96\]. GATK supports the local realignment of reads around known indel sites as well as recalibrating base quality scores to reflect these changes. While it also possesses some capacity for identifying variants, the subsequent step of variant calling is often performed using one
of several available SNV callers based on an analyst’s preference [95–99]. Independent of the tool used, most export variants in Variant Call Format (VCF), a de facto standard file format in the field for the organization and listing variants and related information (such as quality score, coverage, estimated genotype, etc.) [96, 100].

Drug Screening and Discovery

The idea of dose-response pre-dates the Classical era, with Paracelsus first coining the notion that “the dose makes the poison” [101]. In the modern era, the dose-response relationship is utilized to assess how an organism responds to varying levels of a given compound [102, 103]. A sigmoidal curve is typically used to represent this relationship – this model then enabling inferences to be made on a compound’s biological activity [104, 105]. This is a key concept for target identification in drug discovery, and assays have been designed around this to evaluate drug performance and safety as clinically, pharmacological therapies tend to fail primarily due to inefficacy or toxicity [106]. Two well-established and widely-used metrics are half maximal effective concentration (EC$_{50}$) and half maximal inhibitory concentration (IC$_{50}$) [103]. These can be further classified as “absolute” or “relative”, with the “absolute” form portraying the response as a ratio of a known control and the “relative” definition used when such a standard does not exist [103]. Regardless of the unmet clinical need underlying a drug discovery initiative, generally all projects must commence with target identification prior to any clinical studies [107]. However, the search for lead compounds can be expensive and time-consuming, and often take years for drug companies to produce a handful of candidates [107]. Fortunately, technological innovations have vastly expedited this step.

High throughput screening (HTS) of compounds was first adopted by pharmaceutical industries by means of accelerating drug discovery but over the last two decades, has also gained widespread popularity in academia [108]. Previously, traditional methods – which
involved reactions conducted in individual test tubes – were labour-intensive and limited screening in academic laboratories to 20-50 compounds per week [108, 109]. With HTS, large compound libraries can be tested at rates that exceed several thousands per week [108, 109]. More importantly, HTS has enabled the exploration of numerous targets of interest that have emerged in the “post-genome” age [106]. As part of this effort, in vitro models have been and continue to be instrumental for both studying biology and for progressing research forward into in vivo models and ultimately, clinical studies [110]. This is perhaps demonstrated best by the critical role of cancer cell lines in anticancer initiatives [110, 115].

Cancer Cell Lines

Human cancer-derived cell lines have a long-established usage history for studying cancer biology and therapy [111, 116]. The HeLa cancer line – derived from cervical cancer – was the first human cancer-derived cell line to be cultured [117]. With the goals of increasing the scope of tumours covered, replacing the usage of murine models as well as facilitating reproducibility, the US National Cancer Institute 60 (NCI-60) cell line panel was launched in the early 1990s, composed of 60 cell lines from nine different tumour types [116]. Many similar panels have since been created but the implementation of these panels in research have been met with both successes and failures. Perhaps the most well-known achievement of the NCI-60 panel has been the development of proteasome inhibitor bortezomib for the treatment of multiple myeloma, but in general, the low rate of replicating many in vitro findings in in vivo or clinical settings has been the biggest shortcoming of these models [114]. The emergence of “omics” technologies have also enabled the genomic profiling of cell lines. Several studies have demonstrated that the processes of cell passage and immortalization promote cancer lines to resemble each other more so than their tumours of origin [118, 119]. However, others have noted that perhaps larger panels of cell lines, such as the panel of 1200 lines published by the Center for Molecular Therapeutics at the Massachusetts General Hospital Cancer
Biomarkers

The term “biomarker”, a portmanteau for the term “biological marker”, refers to quantifiable biological entities that can be used to characterize either biological processes, pathogenic processes or pharmacological responses to treatment, and distinguish normal ones from abnormal ones [121, 122]. In the history of clinical practice, medical signs have long been used for both diagnoses and prognoses of disease. Analogous to these medical signs, biomarkers are the latest of such implementations which can be more objectively measured in modern laboratories for assessing the disease status of a patient [122]. Three categories of biomarkers exist: diagnostic, prognostic and predictive.

Diagnostic Biomarkers

Diagnostic biomarkers predict disease onset, progression or recurrence in an individual. A good example of a diagnostic biomarker routinely employed in clinical oncology is testing for levels of prostate-specific antigen (PSA) in prostate cancer. PSA is often used for both initial diagnoses as well as identification of recurrence, if levels are found to rise post-treatment [123, 124]. For HNC, this is still an active area of research and while several studies propose different serum and salivary markers [64, 125], none have yet been implemented into routine clinical practice.

Prognostic Biomarkers

Prognostic biomarkers provide insight into the nature of the disease and predicts treatment-independent patterns of survival and recurrence. For example, the EpsteinBarr virus (EBV)
is prognostic in a subtype of HNCs – nasopharyngeal carcinomas – where higher levels of
viral DNA has been associated with more advanced disease, distance relapse as well as poorer
survival [126-129]. Contrastingly in HNSCCs, HPV is a prognostic biomarker (Figure 1.2)
where positive status is associated with more favourable outcomes [55, 130-133]. More
recently, an imaging-derived marker – the maximum standardized uptake value (SUV$_{\text{max}}$)
from functional imaging with $^{18}$F-fluorodeoxyglucose positron emission tomography (FDG-
PET) – has also been shown to be prognostic in HNC [134]. Although the use of prognostic
biomarkers in HNC is not necessarily routine, they are actively being incorporated into
clinical trials [1]. Here, more work is actually needed for the development of predictive
biomarkers that can guide how each patient is treated individually.

Predictive Biomarkers

Predictive biomarkers predict response to a specific therapeutic intervention. Since
prognostic biomarkers estimate response to treatment independent of outcome and predictive
biomarkers assess outcome with respect to a specific treatment, these two categories are
therefore not mutually-exclusive and thus, a biomarker can have both prognostic and
predictive capabilities [1]. A potential example of this is HPV, which in addition to being
prognostic, may also have predictive properties. There is evidence to suggest that HPV-
positive and -negative tumours respond differently to the same therapeutic agent – where
HPV-negative tumours appear more sensitive to the radiosensitizer nimorazole as well as
anti-EGFR therapy [135, 136]. Although some reports have noted a higher frequency of
EGFR mutations in HPV-negative tumours, other studies have shown that alterations in
EGFR alone is not predictive of sensitivity to anti-EGFR treatment [1, 136-140]. While the
mechanism remains unclear, the potential clinical implications are still enormous, as patients
may be given different therapies based on HPV status. However, the research on predictive
biomarkers in HNC is still an active area of investigation and while several molecular targets
have been examined, there is both a shortage of conclusive studies and studies in general to impact clinical care.

**Rationale and Aims**

**Goal**

The main goal of the project is to correlate genomic events with drug response and derive predictive biomarkers for HNC using a set of 33 HNC cell lines.

**Aim 1**

Characterize the genomic changes of HNC cell lines via single nucleotide variants (SNVs) and copy number aberrations (CNAs).

**Rationale**

Since HNC is essentially an umbrella term used to aggregate tumours originating from many sites, this clinical classification makes HNC an inherently heterogeneous disease. HNC is unique in the sense that within the realm of medical oncology, no other tumour type is regarded in quite the same manner – tumours are normally categorized and treated based on their tissue of origin whereas for HNC, tumours originating from several sites are given the same general clinical guidelines. This imprecise approach to therapy has likely contributed to high rates of treatment failure and recurrence for patients [79]. Large genomics studies, such as projects coordinated by the Cancer Genome Atlas (TCGA) network, have demonstrated the diversity of molecular subgroups that exist in HNC both in SNVs and CNAs [141]. Thus as a first step to the goal of the study, we sought comprehensively profile the genomic alterations in the cohort of 33 cell lines via SNVs and CNAs and identify recurrent features of interest that can be used for cohort stratification.
Aim 2

Correlate genomic features to drug response in a panel of drugs selected via high-throughput screening.

Rationale

Current treatment strategies for HNC take a broadly intensive approach. Unfortunately, given the molecular heterogeneity of the disease, a range of responses can arise from a single therapy, resulting in poor treatment success and often unnecessary reductions in a patient’s quality of life [77–81]. The only approved targeted agent to-date is the use of monoclonal antibodies for \textit{EGFR} [76]. Since drug development and translation into the clinic is both lengthy and expensive, it would be worth exploring the re-purposing of already-approved drugs to expedite the translation of bench-to-bedside [142]. Technological advances in the last decade have enabled drug screening to become highly-automated [108, 143]. Thus using a panel of drugs that were selected from the screening of over 1500 drugs, we aim to improve treatment precision by correlating the response of these drugs to molecular features identified in Aim 1 to derive molecular predictors of drug sensitivity.
Materials and Methods

Genomics

Copy Number Profiles

DNA was extracted from formalin-fixed, paraffin-embedded (FFPE) cell line samples and copy number aberrations (CNAs) were profiled using the Affymetrix OncoScan FFPE Assay kit. For 30 cell line samples, .OSCHP files were generated by Affymetrix using their in-house reference. BioDiscoverys Nexus Express$^{TM}$ (version 2.0) for OncoScan Software was used to call CNAs, using the SNP-FASST2 algorithm with default all parameters except for the minimum number of probes per segment, which was set to 10 from the default of 3. Computational estimates of aberrant cell fraction and tumour ploidy, as well as genomic segmentation were facilitated by the Allele-Specific Copy number Analysis of Tumours (ASCAT, v2.1) algorithm [144]. The Log R Ratio (LRR) and B allele frequency (BAF) of the SNP-array probes were computed using the two intensity values provided for each probe, which corresponded to the hybridization of the two alleles interrogated. The following equations were applied:

\[
LRR = \log_2(X + Y) \tag{2.1}
\]

\[
BAF = \frac{Y}{X + Y} \tag{2.2}
\]

where \(Y\) and \(X\) are intensity values corresponding to the minor and major alleles respectively.

The LLR and BAF values were visualized for each sample, ordered along the x-axis by genomic position and along the y-axis by the respective intensity measures. They were
inspected individually for evidence that a recentering was required – namely if a balanced copy number loss was observed (i.e. loss of a copy in a region but still maintained heterozygosity). For samples that required recentering, the ploidy estimated for the sample was set as the new baseline zero for the LLR values and all copy number (CN) calls were adjusted by this ploidy value. Following adjustments, gene-level CNAs for each cell line were identified by overlapping CN segments using RefGene (2014-07-15) and annotation by BEDTools (v2.21.0) [145]. These annotations were effectively used to collapse segment-level data into gene-level data.

GISTIC2.0 (v2.0.22) was applied to assess recurrence of gene copy number variations [146]. To do this, a profile for each cell line was created: each chromosome was segmented into regions of neutral, CN loss, and CN gain events. For each segmented region, the corresponding CN intensity was computed by averaging the copy number intensities of the probes in that particular segment (obtained from the SNP-array. GISTIC2.0 was run with largely default parameters, except for the following, which were changed from default (-genegistic 1 -smallmem 1 -broad 1 -brlen 0.5 -conf 0.99 -rx 0).

**Metrics**

Raw counts for the number of gains, losses and total alterations (losses and gains) were computed using a simple summation. The proportional equivalents of these raw metrics were generated by taking all genes available as the denominator. Lastly, the percent genome altered (PGA) was calculated for each sample by dividing the number of base-pairs that are involved in a CNA by the total length of the genome. The fraction of loss of heterozygosity (LOH) events was calculated at both the sample level (the fraction of genes with an LOH event) and the gene level (the fraction of samples with an LOH event at that gene).
Consensus Clustering

Copy number clustering was performed in the R statistical environment (v3.2.3) with the BioConductor package ConsensusClusterPlus (v1.8.1) using 1000 iterations of hierarchical clustering with 80% subsampling of the genes for the number of clusters ranging from 2 to 15 [147]. Clustering was performed using Ward’s method on Jaccard distances. The seed used was ‘3110’.

CNA Feature Selection

Genes that were identified as significantly altered through analysis with GISTIC2.0 ($q < 0.05$) were carried forward as CNA features for correlation with drug response. This resulted in a total of 201 genes. To decrease the burden of multiple testing, genes were further collapsed into regions of equivalent copy number status by grouping genes based on the Cytoband region.

Visualization

Visualizations were generated in the R statistical environment (v3.2.1) using the lattice (v0.20-33), latticeExtra (v0.6-28), BPG (v5.3.4) and VennDiagram (1.6.17) packages [148, 149]. Figures were compiled using LaTeX. Copy number changes were further split into five categories: “high gain” (more than 2 copies), “gain” (1-2 copies), “neutral” (no change), “loss” (loss of 1 copy) and “high loss” (2 copies or more).

Reference-free SNV Classifier: S22S

SNV identification without a reference sample significantly reduces detection accuracy [96, 150, 151]. As a result, many groups resort to using ad hoc metrics or repurposing tools originally designed for germline analysis [112, 151, 152]. While a small number of analytical
Figure 2.1: A schematic of the algorithm is shown. Per sample, features from the VCF file and from the tumour BAM file are extracted and annotated with population frequency data from publicly available germline and somatic databases to build a feature vector for that sample. From the total number of samples, 30% was set aside for validation as an independent test set and the feature matrices for the remaining 70% aggregated to generate the training set. A 10-fold cross validation scheme was employed using the training set to parameterize the model. The optimal parameters are used to train a random forest on the full training set. Model performance was assessed on the held-out test set.
tools can accommodate unmatched tumour samples, the results often contain large numbers of false positives, and no systematic benchmark yet exists [96]. The most popular approach to date is to generate a surrogate normal from a pool of normal samples [99, 153, 154]. This method can be computationally intensive and at best, is only able to remove a subset of false positive calls resulting from germline contamination, not other sources of sequencing error [99, 151, 152]. To address this challenge, S22S was developed, which is an ensemble machine learning approach to predict true somatic SNVs from single-sample tumour sequencing data when a matched normal sample is unavailable. The procedure is described in full in Figure 2.1.

Datasets

To build the training and test sets for the reference-free classifier (from here on referred to by its name, S22S), high quality exome sequencing samples from the head and neck study published by the The Cancer Genome Atlas (TCGA) network were used [141]. This composed of 376 exome sequences of primary tumours and their matched normal samples (total number of samples 752). BAM files were obtained from the Cancer Genomics Hub (CGHub, https://cghub.ucsc.edu/) and realigned and recalibrated prior to downstream analyses.

Re-Alignment of Exome Sequencing Data

For each BAM file obtained from CGHub, back-conversion to FASTQ files was done to allow realignment to the human reference genome for standardization, using the SamToFastq function from picard (v.1.92) (http://broadinstitute.github.io/picard). The realignment was performed using the human reference build GRCh37 with decoy (hs37d5) and bwa mem (v0.7.12). Merging of lane-level BAMs from the same library within each sample was facilitated via picard (v1.92), with duplicates marked, and was followed by library-level merging of BAMs, without marking of duplicates. The Genome Analysis Toolkit
(GATK, v3.4.0) was used to perform local realignment and base quality recalibration on each realigned tumour/normal BAM pair [95, 156, 157]. Separate tumour and normal sample-level BAM files were extracted, followed by header correction using samtools (v0.1.19) and indexing using picard (v1.107) to generate the final realigned and recalibrated BAM file per sample.

**Defining Classes**

In order to train S22S to accurately recognize true calls from false ones, a comprehensive approach was taken when defining the truth and false sets. To create an adequate feature space for machine learning, variant calling was conducted in four manners: two of which used the tumour/normal pair and two of which solely utilized the tumour samples. First, germline single nucleotide polymorphisms (SNPs) were identified using the standard approach with tumour/normal pairs using GATK while somatic SNVs were called using standard approaches employing tumour/normal pairs with SomaticSniper [97]. Subsequently, tumour-only variants were aggregated from running GATK on only the tumour samples and similarly with MuTect, by running the tool on only the tumour samples (with a panel of normals (PoN) generated from the normal samples). The true class was defined by variants identified using SomaticSniper on paired samples but which were absent in the GATK germline calls; the false class encompassed all calls that were present in the germline set as well as both sets derived from tumour-only variant detection. To eliminate ambiguity during model training, overlaps between the germline and somatic calls were omitted from downstream analysis.

**Germline SNP Calling**

GATK (v3.4.0) was used to call germline SNPs. For each sample, HaplotypeCaller was used followed by VariantFiltration to hard-filter the callset using the following filter expressions:

22 of 243
“QD > 10.0 || FS > 60.0 || MQ < 40.0 || DP < 50 || SOR > 4.0 || ReadPosRankSum < -8.0 || MQRankSum < -12.5” and “MQ0 >= 4 && ((MQ0 (1.0 * DP)) > 0.1)” to generate the final germline calls. The GATK Best Practices recommendations were consulted for the development of the processing pipeline [95, 157].

**Somatic SNV Calling**

Somatic SNVs were predicted using SomaticSniper (v1.0.4). First, candidate somatic calls were identified using bam-somaticsniper, with all default settings except for the -q option (mapping quality threshold) which was set to one instead of the default zero at the recommendation of the developer. Next, a series of Perl scripts provided by the SomaticSniper package were used to filter out possible false positives (http://gmt.genome.wustl.edu/packages/somatic-sniper/documentation.html). The candidate list was further filtered using several methods. First, standard and LOH filtering were applied using pileup INDEL files generated for both the normal and tumour BAM files using SAMtools (v1.0.6). Then, a bam-readcount filter (bam-readcount downloaded on Jan 10th, 2014) was applied, with a mapping quality filter -q 1 (using default settings otherwise). Additionally, the false positive filter was ran and a high confidence filter (with default parameters). The resulting VCF file with high confidence somatic SNV calls for each sample was subsequently filtered against several lists of known germlines to further reduce potential false positives.

In the filtering approach, SNVs that were found in any of the following databases (also referred to as “blacklists”) using tabix were removed to produce the final set of somatic calls [158]: dbSNP142 [159] (modified to remove somatic and clinical variants, with variants with the following flags excluded: SAO = 2/3, PM, CDA, TPA, MUT and OM), 1000 Genomes Project (v3) [100], Complete Genomics 69 whole genomes [160], duplicate gene database (v68) [161], ENCODE DAC and Duke Mapability Consensus Excludable databases [162] (comprising poorly mapping reads, repeat regions, and mitochondrial and ribosomal DNA),
invalidated somatic SNVs from 68 human colorectal cancer exomes (unpublished data) using the AccuSNP platform (Roche NimbleGen), germline SNPs from 477 samples used in our previous work in prostate cancer with an additional ten WGS from prostate cancer patients with higher Gleason score (Fraser et al., Nature, in press) and the Fuentes database of likely false positive variants [163]; SNVs were “whitelisted” or kept (independent of presence in other databases) if they were found in the Catalogue of Somatic Mutations in Cancer (COSMIC) database (v71) [164].

**Tumour-only SNV Calling**

To capture the error profiles of SNVs called in situations where a matched normal sample is absent, somatic SNV calling was performed using only the tumour samples. This was done in two ways, which were the most frequently employed methods in such a situation. The first approach was to run variant calling with GATK on each tumour sample without its matched normal, followed by hardfiltering and then the “VariantFiltration” function to generate the final list of variants for that sample. The second approach was to generate a panel of normal samples (PoN) to serve as the normal surrogate. This was done by pooling together a cohort of normal samples with MuTect. In brief, each normal BAM file was processed separately and passed to MuTect with the parameter label of “tumour”, and with the “artefact_detection_mode” set on. The output VCF per sample was merged together using the CombineVariants function from the GATK engine to create the final normal panel, which is a VCF file. This VCF file was then passed to MuTect as a normal panel when calling variants for each tumour sample and generating the final list of variants.

**Features**

The underlying hypothesis is that true variants will exhibit different underlying sequencing properties, genomic background characteristics and population distributions when compared
to germline variants or sequencing artifacts. Thus to characterize this for each variant
detected in the final VCF file, a set of features was extracted based on its genomic position.
These features consisted of metrics found within the VCF file, metrics that were derived
from the tumour BAM file as well as minor allele and population frequencies found in public
databases (the 1000 Genomes Project and NHLBI) and recurrence calculations at both
at the position- and gene-level based on results released from the original TCGA study
[100][141][165]. This resulted in a total of 61 features (Table 2.1).

<table>
<thead>
<tr>
<th>Feature Alias</th>
<th>Feature Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>GATK-derived Features</strong></td>
<td></td>
</tr>
<tr>
<td>AC</td>
<td>Allele count for ALT allele</td>
</tr>
<tr>
<td>AF</td>
<td>Allele frequency for ALT allele</td>
</tr>
<tr>
<td>AN</td>
<td>Total number of alleles called in genotype</td>
</tr>
<tr>
<td>BaseQRankSum</td>
<td>Z-score from Wilcoxon rank sum test of ALT Vs. REF base qualities</td>
</tr>
<tr>
<td>ClippingRankSum</td>
<td>Z-score From Wilcoxon rank sum test of Alt vs. Ref number of hard clipped bases</td>
</tr>
<tr>
<td>DP</td>
<td>Approximate read depth</td>
</tr>
<tr>
<td>FS</td>
<td>Phred-scaled p-value using Fisher’s exact test to detect strand bias</td>
</tr>
<tr>
<td>MLEAC</td>
<td>Maximum likelihood expectation (MLE) for the allele counts</td>
</tr>
<tr>
<td>MLEAF</td>
<td>Maximum likelihood expectation (MLE) for the allele frequency</td>
</tr>
<tr>
<td>MQ</td>
<td>Root Mean-squared (RMS) Mapping Quality</td>
</tr>
<tr>
<td>MQRankSum</td>
<td>Z-score From Wilcoxon rank sum test of ALT vs. REF read mapping qualities</td>
</tr>
<tr>
<td>QD</td>
<td>Variant Confidence/Quality by Depth</td>
</tr>
<tr>
<td>ReadPosRankSum</td>
<td>Z-score from Wilcoxon rank sum test of ALT vs. REF read position bias</td>
</tr>
<tr>
<td>SOR</td>
<td>Symmetric Odds Ratio of 2x2 contingency table to detect strand bias</td>
</tr>
<tr>
<td>REF.AD</td>
<td>Allelic depth for REF</td>
</tr>
<tr>
<td>ALT.AD</td>
<td>Allelic depth for ALT</td>
</tr>
<tr>
<td>REF.AF</td>
<td>Calculated allelic frequency of REF (REF.AD/DP)</td>
</tr>
<tr>
<td>ALT.AF</td>
<td>Calculated allelic frequency of ALT (ALT.AD/DP)</td>
</tr>
</tbody>
</table>

continued ...
### Feature Alias | Feature Description
---|---
mutect | Presence of call using MuTect with a panel of normals
gatk | Presence of call using GATK in tumour-only mode

#### Population Features

<table>
<thead>
<tr>
<th>Feature</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>NHLBI_EA_MAF</td>
<td>Minor allele frequency in percent for European population from NHLBI</td>
</tr>
<tr>
<td>NHLBI_AA_MAF</td>
<td>Minor allele frequency in percent for African American population from NHLBI</td>
</tr>
<tr>
<td>NHLBI_ALL_MAF</td>
<td>Overall minor allele frequency from NHLBI</td>
</tr>
<tr>
<td>AF.1000Genomes</td>
<td>Global allele frequency based on AC/AN from 1000 Genomes Project</td>
</tr>
<tr>
<td>AFR_AF</td>
<td>Allele frequency for the Afiican population based on AC/AN from 1000 Genomes Project</td>
</tr>
<tr>
<td>AMR_AF</td>
<td>Allele frequency for the American population based on AC/AN from 1000 Genomes Project</td>
</tr>
<tr>
<td>ASN_AF</td>
<td>Allele frequency for the Asian population based on AC/AN from 1000 Genomes Project</td>
</tr>
<tr>
<td>EUR_AF</td>
<td>Allele frequency for the European population based on AC/AN from 1000 Genomes Project</td>
</tr>
<tr>
<td>LDAF</td>
<td>Allele frequency as inferred from haplotype estimation</td>
</tr>
<tr>
<td>TCGA_fraction_MAF</td>
<td>Proportion of recurrence in the TCGA dataset, position-based</td>
</tr>
<tr>
<td>TCGA_fraction_RecSNV</td>
<td>Proportion of recurrence in the TCGA dataset, gene-based</td>
</tr>
</tbody>
</table>

#### Sequencing Features

<table>
<thead>
<tr>
<th>Feature</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>MappingQual</td>
<td>Mapping quality scores (in phred-scale) for probability of misplaced read</td>
</tr>
<tr>
<td>ReadPosition</td>
<td>Position within the read</td>
</tr>
<tr>
<td>TumourCoverage</td>
<td>Read depth from tumour bam file</td>
</tr>
<tr>
<td>DistanceSNP</td>
<td>Distance (in bases) to closest known germline SNP</td>
</tr>
<tr>
<td>DistanceIndel</td>
<td>Distance (in bases) to closest known germline Indel</td>
</tr>
<tr>
<td>StrandBias</td>
<td>Proportion of reads mapped to forward strand</td>
</tr>
<tr>
<td>RefForward</td>
<td>Quality of at least 13 REF bases on forward strand</td>
</tr>
<tr>
<td>RefReverse</td>
<td>Quality of at least 13 REF bases on reverse strand</td>
</tr>
<tr>
<td>NonRefForward</td>
<td>Quality of at least 13 non-REF bases on forward strand</td>
</tr>
<tr>
<td>NonRefReverse</td>
<td>Quality of at least 13 non-REF bases on reverse strand</td>
</tr>
<tr>
<td>SumRefBase</td>
<td>Sum of reference base qualities</td>
</tr>
<tr>
<td>SumSqRefBase</td>
<td>Sum of squares of reference base qualities</td>
</tr>
<tr>
<td>SumNonRefBase</td>
<td>Sum of non-ref base qualities</td>
</tr>
</tbody>
</table>

continued . . .
...continued

<table>
<thead>
<tr>
<th>Feature Alias</th>
<th>Feature Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>SumSqNonRefBase</td>
<td>Sum of squares of non-ref base qualities</td>
</tr>
<tr>
<td>SumRefMap</td>
<td>Sum of ref mapping qualities</td>
</tr>
<tr>
<td>SumSqRefMap</td>
<td>Sum of squares of ref mapping qualities</td>
</tr>
<tr>
<td>SumNonRefMap</td>
<td>Sum of non-ref mapping qualities</td>
</tr>
<tr>
<td>SumSqNonRefMap</td>
<td>Sum of squares of non-ref mapping qualities</td>
</tr>
<tr>
<td>SumTailRef</td>
<td>Sum of tail distance for ref bases</td>
</tr>
<tr>
<td>SumSqTailRef</td>
<td>Sum of squares of tail distance for ref bases</td>
</tr>
<tr>
<td>SumTailNonRef</td>
<td>Sum of tail distance for non-ref bases</td>
</tr>
<tr>
<td>SumSqTailNonRef</td>
<td>Sum of squares of tail distance for non-ref</td>
</tr>
<tr>
<td>RefAllele</td>
<td>Forward ref allele</td>
</tr>
<tr>
<td>NonRefAllele</td>
<td>Forward non-ref allele</td>
</tr>
<tr>
<td>BaseQual</td>
<td>Mapping quality scores (in phred-scale) for probability of</td>
</tr>
<tr>
<td></td>
<td>misidentified base</td>
</tr>
<tr>
<td>DP.pileup</td>
<td>Number of reads covering or bridging position</td>
</tr>
<tr>
<td>VDB</td>
<td>Variant distance bias for indication of potential</td>
</tr>
<tr>
<td></td>
<td>misalignment due to nearby SNP</td>
</tr>
</tbody>
</table>

**Genomic Context**

<table>
<thead>
<tr>
<th>Feature</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>homopolymer.rate</td>
<td>Frequency of homopolymers within 100 bases up and</td>
</tr>
<tr>
<td></td>
<td>downstream of position</td>
</tr>
<tr>
<td>GC</td>
<td>Proportion of bases that are GC within 100 bases up and</td>
</tr>
<tr>
<td></td>
<td>downstream of position</td>
</tr>
<tr>
<td>trinucleotide</td>
<td>Trinucleotide context of position</td>
</tr>
</tbody>
</table>

**Table 2.1:** The full set of 61 features used, categorized based on origin and purpose, with a brief description of each.

**Model Training**

Random forest (RF) was selected as the machine learning method for the classifier. An RF is an ensemble approach that is resistant to outliers and can effectively handle highly-correlated variables [166]. The model training was done using the randomForest package (v4.6-10) in the R statistical environment (v3.2.3). In order to properly validate model performance, 30% of the samples was set aside as a held-out test set. With the 70% remaining – the training set – a grid search was performed to fine-tune model parameters as some parameters of random
forest are more sensitive to tuning [167]. Samples were randomized to training and testing groups by first finding the nearest integer count corresponding to 30% of the total samples available and using the base function sample from the stats package in the R statistical environment (v3.2.3) to randomly sample the respective sizes. A seed of ‘333 was used for reproducibility.

**Random Forest Parameterization**

As mentioned previously, it has been shown that some parameters of random forest are more sensitive to tuning than others [167]. In addition, due to the nature of our dataset, a large class imbalance exists between the negative cases and positive cases (the number of germline SNPs called greatly outweigh the number of somatic SNVs called). Thus to improve predictive power, the majority class was down-sampled as a function of the minor class and grid search was performed to tune the parameters of random forest using the 70% training set. The performance was assessed using a 10-fold cross-validation scheme with the metric being the area under the precision recall curve (AUPRC). The most optimal values to use for the full model were selected based on the highest overall AUPRC. Three parameters were tuned for the model: mtry, nodesize and ntree.

For the parameter mtry, factor levels of the default value were tested. Since the nature of the problem is classification and the default mtry for classification is the square root of the number of features, thus in addition to the default value of 7, half of the default (4), twice the default (14) and three times the default (21) were used. For ntree, 1000, 5000 and 10,000 trees were tested while values of 5, 25, 50 and 100 were tested for the parameter nodesize. To split the 70% training set into 10 reasonably equal sets, an approach similar to the train-test splitting was taken. First, the nearest integer count corresponding to 10% of the number of samples in the training set was calculated. The base function sample from the stats package
in the R statistical environment (v3.2.3) was then used to randomly sample from the training samples 9 times, with the remaining samples all assigned to the 10th fold. A seed of ‘99’ was set for reproducibility. The last fold on average contained 1-2 more samples than the rest of the folds. During grid search parameterization, each fold was kept constant across all parameter combinations tested. Parameter combinations were ranked based on the AUPRC across ten folds (ties were broken using the standard deviations where a higher rank was attributed to the lower standard deviation) and the set of parameters that generated the highest overall AUPRC was selected for the full model.

**Performance Assessment**

By varying the vote threshold from zero to one, the number of false positive (FP), false negative (FN), true positive (TP) and true negative (TN) calls generated using different cutoffs can be obtained across a continuous range. This was facilitated by the pROC package (v1.8) in the R statistical environment (v3.2.3) [168]. These were then used to calculate metrics for assessing model performance such as sensitivity, specificity and precision. A curve was also constructed using the continuous range of precision and recall values and the AUPRC approximated using the trapezoid method:

$$AUPRC = \frac{1}{2} \sum_{k=1}^{N} (x_{k+1} - x_k)(y_{k+1} - y_k)$$

where \( x \) is the the recall and \( y \) is the precision at cutoff \( k \). The set of parameters with the highest average rank of AUPRC across the ten folds was selected for the full model. When selecting an operating point (the vote threshold cutoff point) for the model based on the AUPRC, the threshold that maximized the harmonic mean of precision and recall, also
known as the $F_1$ score, was selected. This was calculated by:

$$F_1 = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$  \hspace{1cm} (2.4)$$

where precision is defined as:

$$\text{precision} = \frac{TP}{TP + FP}$$  \hspace{1cm} (2.5)$$

and recall is defined as:

$$\text{recall} = \frac{TP}{TP + FN}$$  \hspace{1cm} (2.6)$$

**Cross-tumour Performance and Model Convergence**

Two approaches were taken to gauge the generalizability of the model and assessed the effects of tumour type and cohort size on overall performance. First, the model was applied a set of prostate samples that were processed in a similar fashion and its performance evaluated using the same AUPRC metric. Second, to determine the effects of training size on model performance, a convergence experiment was conducted where the full training set was subset into smaller, reasonable sample sizes of 5, 10, 25, 50, 100, 150 and 200. Performance was assessed using the AUPRC metric and compared with the full model ($n = 263$).

**Data Visualization**

Visualizations were generated in the R statistical environment (v3.2.3) using the lattice (v0.20-33), latticeExtra (v0.6-28), BPG (v5.6.8) and VennDiagram (1.6.17) packages [148, 149]. Figures were compiled using LaTeX.

**Benchmarking**

To assess how model performance compares to available methods developed under similar pretenses, three published tools (EBFilter, Virtual Normal Correction (VNC) and VarDict)
were applied to the same test sets of 113 samples as the S22S model [153, 154, 169]. These tools were selected based on their proposed functionality and the fact that they accepted similar inputs and generated comparable outputs to the S22S method. The AUPRC metric was used to assess performance across the different algorithms. Since each tool had its own specific method or score for annotating potential true variants or variants of interest, to generate a curve for each, the threshold for each tool was varied across its full range of values for the entire cohort. All tools were ran with default options or with those suggested in the user guidelines for the tool. VarDict was executable given a VCF file, BAM file and the human genome reference fasta as inputs, however, some of the other tools required additional input files. Both EBFilter and VNC took an approach that was similar to MuTect in that both required an additional list of normal samples to create a surrogate normal. For EBFilter, a list of high quality prostate normal samples from the TCGA were used to generate this panel in all of the EBFilter runs. For VNC, all available (n = 427) variant files of normal samples that were processed by Complete Genomics (CG) for the 1000Genomes Project were downloaded from the FTP site (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase3/data/). These samples were used instead of TCGA normal samples because the tool was designed to work with CG outputs and only accepted the varfile format as input to generate the virtual normal. To generate a range of precision and recall values for EBFilter, the score that was outputted by the tool was varied from 0 to the maximum score of the cohort. Since VNC implemented two fields for filtering calls both relating to the number of samples out of the total pool of samples used in the construction of the virtual normal in which a variant was found both thresholds were varied from 0 to 427 and precision and recall at each combination of thresholds calculated. In single-sample mode, VarDict implements a single an allele frequency (AF) filter during variant identification, so to calculate precision and recall for the PR-curve, the AF was varied from zero to one. When selecting operating points for each tool, the same approach as with S22S was taken where the operating point
was selected based on maximum F₁ score.

Mutational Profiles

Re-Alignment of Sequencing Data

The Binary Alignment/Mapping (BAM) file for each cell line exome sequence was back-converted to FASTQ format to allow for realignment to the human reference genome for standardization purposes. This was done using the SamToFastq function from picard (v.1.92) [http://broadinstitute.github.io/picard](http://broadinstitute.github.io/picard). The realignment was performed using the human reference build GRCh37 with decoy (hs37d5) and bwa mem (v0.7.11). Merging of lane-level BAMs from the same library within each sample was facilitated via picard (v1.92), with duplicates marked, and was followed by library-level merging of BAMs, without marking of duplicates. The Genome Analysis Toolkit (GATK, v3.4.0) was used to perform local realignment and base quality recalibration on each realigned tumour/normal BAM pair [95, 156, 157]. Separate tumour and normal sample-level BAM files were extracted, followed by header correction using samtools (v0.1.19) and indexing using picard (v1.107) to generate the final realigned and recalibrated BAM file per sample.

Variant Calling and Germline SNP Removal

GATK (v3.4.0) was used to filter out potential germline SNPs from the re-aligned samples. For each sample, HaplotypeCaller was used followed by VariantFiltration to hard-filter the callset using the following filter expressions: “QD > 10.0 || FS > 60.0 || MQ < 40.0 || DP < 50 || SOR > 4.0 || ReadPosRankSum < -8.0 || MQRankSum < -12.5” and “MQ0 >= 4 && ((MQ0 (1.0 * DP)) > 0.1)” to generate the final germline calls in Variant Call Format (VCF). The GATK Best Practices recommendations were incorporated into the development of the processing pipeline [95, 157].

32 of 243
Application of S22S to Cell Lines

The VCFs following the removal of germline SNPs in section 2.1.3 were passed through the S22S pipeline. First, a variant-by-feature matrix was constructed for each sample, with the features extracted from the VCF and BAM files and annotated with SNV frequencies (see Figure 2.1). Missing values from the VCF or BAM files were imputed with the sample median for that feature while missing allele frequency values were set to 0. The trained RF model was applied using default prediction cutoff of 0.5 to generate a list of predicted classes for each sample. The input VCF was then subset into two files: predicted somatic and non-somatic variants.

Predicted somatic SNVs were then functionally annotated by ANNOVAR (v2015-06-17) using the RefGene database where SNVs that were annotated as “nonsynonymous”, “stop-loss”, “stop-gain” and “splice-site” were considered functional [170]. If more than one mutation is found in a sample for a gene, then the mutation of the higher priority functional class was selected for visualization. SNVs were further filtered using tabix based on presence of the SNV in a set of databases (termed “blacklists”); any SNVs found were removed to generate the final set of somatic calls [158]. The blacklists used were:

- dbSNP142 (modified to remove somatic and clinical variants, with variants with the following flags excluded: SAO = 2/3, PM, CDA, TPA, MUT and OM) [159]
- 1000 Genomes Project (v3) [100]
- Complete Genomics 69 whole genomes [160]
- duplicate gene database (v68) [161]
- ENCODE DAC and Duke Mapability Consensus Excludable databases (comprising poorly mapping reads, repeat regions, and mitochondrial and ribosomal DNA) [162]
• invalidated somatic SNVs from 68 human colorectal cancer exomes (unpublished data) using the AccuSNP platform (Roche NimbleGen)

• germline SNPs from 477 samples used in our previous work in prostate cancer with an additional 10 WGS from prostate cancer patients with higher Gleason score (Fraser et al., Nature, in press)

• Fuentes database of likely false positive variants [163]

SNVs were kept or “whitelisted” (independent of presence in other databases) if they were found in the Catalogue of Somatic Mutations in Cancer (COSMIC) database (v71) [164].

SNV Feature Selection

Following additional blacklist filtering of S22S predictions, the remaining somatic SNVs were tabulated across the cohort of 33 cell lines and reorganized according to recurrence. If more than one mutation is found in a sample for a gene, then the mutation of the higher functional priority was carried forward – where “nonsynonymous” takes priority over “stopgain” and “stoploss”, which are mutually-exclusive and thus tied, and the three over “splice-site”. In an effort to have reasonable-sized groups for statistical testing, a recurrence threshold of presence in at least three samples was applied to the gene list to select SNV features for correlation with drug response.

Drug Screening

Drug Selection

The high-throughput drug screening and selection of the panel of drugs in HNC is the subject of another thesis by a colleague that has already been published [171]. Briefly, a total of 1,505
drugs were screened at the S.M.A.R.T Facility in the Samuel Lunenfeld Research Institute at Mount Sinai Hospital in Toronto, Ontario, Canada. Initial screens were conducted at a single concentration of 4 µM, with cell viability quantified using fluorescence measurements following the addition of alamarBlue – metabolic agent that undergoes a colour change when it is reduced by mitochondrial activity in live cells. The top 5% of compounds (n = 78), as ranked by the greatest average percent reduction in cellular growth across all cell lines, were further filtered based on three criteria: universal potency (defined as broad activity across all cell lines), differential activity (defined as compounds that were extremely potent in some cell lines compared to others) and compounds affecting the PI3K/Akt/mTOR pathway. This narrowed the 78 hits down to a panel of 23 drugs, which underwent further screening with the same procedure but at ten different doses [171]. At a later time point, 14 of the initial 23 drugs were repeated for a subset of cell lines and an additional 10 new drugs of interest were added to the panel. Screening procedures were as outlined previously. These data were the input into my thesis and were used for all subsequent analyses.

**Drug Sensitivity**

**Dose Response Curve**

I performed model fitting in the R statistical environment (v 3.2.3) using the nplr (v0.1-1) package. All response measurements were rescaled as a fraction of response at zero drug concentration. Subsequently, a five parameter logistic (5PL) regression model was used to fit the percent inhibition and drug concentrations in log\(_{10}\) space for each cell-line-drug pair to generate log-dose-response (LDR) curves. The 5PL model is a more general version of the well-known four-parameter logistic (4PL) model, where the symmetry of the curve is
changeable instead of set to one \[103\]. For this, the following formula was implemented:

\[
f(x) = A + \frac{(D - A)}{[1 + (\frac{x}{C})^B]^E}
\]  

(2.7)

where:

- \(A\) is the maximum asymptote or maximum affect level (typically denoted \(A_{\text{max}}\))
- \(B\) is the Hill’s coefficient – slope of the curve
- \(C\) is the inflection point
- \(D\) is the bottom asymptote or response at standard concentration
- \(E\) is the asymmetry factor (where \(E = 1\) results in a 4PL model)

**Sensitivity Metrics**

Two continuous metrics and one categorical one were used to assess sensitivity of a cell line to a compound. The two continuous metrics were the relative IC\(_{50}\) and area under the LDR curve – termed modified activity area (AA), an adaptation of a previously-described and well-known area-based metric \[112\]. The relative IC\(_{50}\) was estimated as the anti-log of the concentration at \(f(x) = 0.5\). Since this value represented half of the maximal tested for a particular drug – without the use of a known true antagonist – the relative definition was employed. From here onwards, the term IC\(_{50}\) will be used exclusively for the relative definition unless otherwise specified. For cases where an IC\(_{50}\) could not be estimated, the IC\(_{50}\) was set to N/A. A 95% confidence interval was also generated for calculable IC\(_{50}\)s based on the LDR curve. The AA was a concept first introduced by the Cancer Cell Line Encyclopedia (CCLE) and refers to the area between the response curve and either a fixed maximal response or the maximum that was tested \[112\]. Due to differences in normalization procedures between this project and the CCLE project, a modified metric was utilized instead. This metric was calculated using Simpson’s rule – by taking the integral of the
fitted model for calculating the AUC of the LDR curve.

\[
Modified \text{ Activity Area} = \int_{a}^{b} f(x)dx
\]  \hspace{1cm} (2.8)

where \( a \) and \( b \) represent the range of \( x \) and \( f(x) \) is the fitted model.

Subsequently, response was also further classified into one of five categories – resistant, partially sensitive, partial response, poor model fit, or sensitive – based on characteristics of the LDR curve.

1. A resistant response was defined by a maximal inhibition of under 80% and either a coefficient of variation (CV) across the range of responses under 30% or the different between the top and bottom asymptotes under 30%.

2. A partial response was defined as a maximal inhibition of greater than or equal to 80% despite having a CV or the different between top and bottom asymptotes of under 30%.

3. Partial sensitivity was defined as having a different between the asymptotes of between 30-60% and less than 80% maximal inhibition as well as a goodness of fit greater than 0.9.

4. Any potentially “sensitive” models with a goodness of fit under 0.9 was categorized as having poor model fit.

5. Responses that had more than 60% difference between top and bottom asymptotes and more than 80% maximal inhibition were deemed to be sensitive.
Statistical Testing

SNV and CNA features were first binarized and tested for whether the presence of any single feature resulted in differential sensitivity (evaluated using IC\textsubscript{50} and AA) that was statistically significant, with a two-sided Wilcoxon rank sum test \[172\]. Prior to statistical testing, a variance and sample size filter was applied to reduce multiple testing concerns. Drugs that exhibited low variance with respect to either response metrics were filtered: variances were calculated separately for IC\textsubscript{50} values (in log\textsubscript{10} space) and AAs where a drug with either variance\textsubscript{IC50} < 1 (in log\textsubscript{10} space) or variance\textsubscript{AA} < 0.5 was not assessed using the Wilcoxon test. Furthermore, any drug with fewer than 10 samples was also omitted from testing. IC\textsubscript{50}s that could not be estimated were omitted as well and the resulting \(p\)-values from testing were adjusted for multiple testing using a false discovery rate (FDR) of 5\% \[173\].

To assess whether there was a significant difference in the proportions of cell lines categorized as “sensitive” vs. any of the other four categories for each drug, given the status of a genomic feature, a two-sample proportions test (also known as a two-sample z-test with \(H_0 : p_1 = p_2\) and \(H_a : p_1 \neq p_2\)) was carried out on the two groups defined by each genomic features and the proportion of each group identified as “sensitive”. Similarly, the resulting \(p\)-values were adjusted for multiple testing using a false discovery rate (FDR) of 5\% \[173\].

Sensitivity Modelling

The model fitting process was conducted in the R statistical environment (v3.3.1) using the MASS (v7.3-45), caret (v6.0-71) and glmnet (v2.0-5) packages. First to examine individual regression parameters, a generalized linear model (GLM) was fit for each drug using all
genomic features in the following manner \[174\]:

\[ \text{Response}_{\text{drug}} \sim g_1 + g_2 + \ldots + g_n \] (2.9)

where:

\( \text{Response}_{\text{drug}} \) corresponds to modified AA, IC\(_{50}\) or response category

\( g_1 \ldots g_n \) correspond to genomic features

Next, potential interactions between genomic variables were examined. Due to the exhaustive number of potential interactions with the current set of features, the number of features was first reduced prior to expedite model fitting, facilitate interpretation as well as enhance generalizability by reducing the potential for overfitting \[175\]. To do this, a LASSO (least absolute shrinkage and selection operator) – a penalized regression method based on least squares – was used to perform feature selection for each drug \[176\]. This method penalizes the \( \beta \) coefficients for each variable and shrinks those that are not important for predicting the outcome to zero, thus only retaining important variables for predicting the response variable. The standard definition was used:

\[
\min_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{N} \| y - X\beta \|_2^2 + \lambda \| \beta \|_1 \right\}
\] (2.10)

where:

\( p \) is the number of features

\( X \) is the covariate matrix

A three-fold cross validation scheme was employed to estimate error and optimize \( \lambda \). In this case, \( \lambda \) was selected as the value resulting in the simplest model out of all models that were equivalent or nearly equivalent in terms of error with the best model. Subsequently,
$X, y$ as well as the optimized $\lambda$ were passed to equation 2.2.4 to obtain a vector of $\beta$ values. This process was repeated 100 times for stability and variables with non-zero coefficients in at least 10% of iterations were selected for the final model for each drug. The final model for each drug employed a similar GLM as equation 2.2.4 and used variables selected using LASSO and also assessed all possible interactions between these variables.

\[ \text{Response}_{\text{drug}} \sim g_1 + g_2 + g_1 : g_2 + g_3 + g_1 : g_3 + g_1 : g_2 : g_3 + \ldots + g_{n-1} : g_n \] (2.11)

where:

- $\text{Response}_{\text{drug}}$ is modified AA, IC$_{50}$ or response category
- $g_1...g_n$ correspond to genomic features
- $g_1 : g_2...g_{n-1} : g_n$ correspond to all potential feature interactions

A stepwise backward and forward selection approach was taken to capture the best combination of variables and interactions for predicting outcome. Here, a term was iteratively added or removed from the model and the quality of the new model assessed using the Akaike information criterion (AIC), a measure of relative model quality [177]. To gauge model performance and estimate error profiles of the model, this process was repeated using a leave one out cross-validation scheme (LOOCV), where a model was fitted per drug using the same stepwise forward and backwards selection approach using all but one sample. The model selected from this stepwise approach was then used to predict on the response for the sample that was left out. This was repeated for all samples and the root mean squared error (RMSE) calculated to quantify model performance. The final model for each drug was fit using all samples and the combination of variables with the lowest AIC was selected as the final model of drug sensitivity.
Power Analysis

Wilcoxon Test

To calculate power of the Wilcoxon test applied – a statistical test that is distribution-free under the null hypothesis – a test distribution was simulated in order to approximate statistical power. For a given drug, the standard deviations (SDs) in the response variable for each of two groups (defined by the presence/absence of a genomic feature) were used to randomly generate sample distributions around either a mean of 0 (no effect), or the desired effect size (effect sizes from 0.5 to 2 in increments of 0.5 were tested). A Wilcoxon rank-sum test was applied to the simulated responses to obtain a \( p \)-value. This process was simulated 1000 times and the simulation (of 1000 iterations) repeated three times where the median \( p \)-values were taken as the estimated \( p \)-values. The percent of \( p \)-values that were lower than the \( \alpha \) value (set to 0.05) was the estimate of statistical power for a given feature. This approach of simulating power was done for all features for each drug.

Proportions Test

Since the underlying distribution of the two-sample proportions test is the normal distribution, a simulation was not necessary to calculate power. Here, power was calculated in the R statistical environment (v3.3.1) using the base \texttt{stats} package. Effect sizes from 0.5 to 2 in increments of 0.5 were tested.

Generalized Linear Model

To calculate power for each GLM model, a power simulation similar to the approach described with the Wilcoxon test was undertaken. Here, the final models selected were summarized using a single coefficient by predicting on the original model.
\[ Y \sim \text{feature}_1 + \text{feature}_2 + ... + \text{feature}_n(fit_1) \quad (2.12) \]

\[ Z = \text{predict}(fit_1) \quad (2.13) \]

For continuous response variables (AA and IC\textsubscript{50}), a simulated response variable (\(Y'\)) was drawn randomly with mean and standard deviation defined by the original \(Y\). For the categorical variable, random response labels (“sensitive” vs. “not sensitive”) were generated using the same proportions seen in the original \(Y\). Then, the newly sampled response values, \(Y'\), were then fitted using the summarized version of the original model (equation 2.2.5) and the \(p\)-value for the \(\beta\) term calculated.

\[ Y' \sim \beta Z \quad (2.14) \]

This simulation was repeated 1000 times for each model and the percent of \(p\)-values below the \(\alpha\) value (defined as 0.05) was the estimated power of that GLM model.

**Visualization**

Visualizations were generated in the R statistical environment (v3.2.3) using the lattice (v0.20-33), latticeExtra (v0.6-28) and BPG (v5.6.12) packages \cite{L48}. Figures were compiled using LaTeX.
Results

Overview

<table>
<thead>
<tr>
<th>Cell Lines (n)</th>
<th>Source</th>
<th>HPV-status</th>
</tr>
</thead>
<tbody>
<tr>
<td>CNA 30</td>
<td>Affymetrix OncoScan</td>
<td>6 HPV+, 24 HPV-</td>
</tr>
<tr>
<td>SNV 33</td>
<td>Illumina Exome Sequencing</td>
<td>6 HPV+, 27 HPV-</td>
</tr>
<tr>
<td>Drug 28</td>
<td>Screening (10 points)</td>
<td>5 HPV+, 23 HPV-</td>
</tr>
</tbody>
</table>

Table 3.1: Outline of all data sources

The majority of HNC cell lines (28 out of 33) had complete data to generate all three types of profiles (mutational, copy number and drug response) (Table 3.1, Figures 3.1 and 3.2). Samples with incomplete data were still analyzed with respect to its available data types but not used in the final correlation of genomics with drug response. The final set of 28 cell lines was composed of 23 HPV-negative cell lines and 5 HPV-positive cell lines.

Copy Number Analysis

Experimental Design

For most genomic assays, CN calls made from FFPE samples are of poorer quality in comparison to fresh frozen samples due to the sample preservation procedure – which uses formaldehyde in a highly acidic environment – typically causing damages to the DNA [178-181]. To mitigate this, the molecular inversion probe (MIP) technology was developed, which requires very little DNA input (< 100 ng), works well with degraded FFPE DNA and has
been shown to produce high quality CN data with both FFPE samples and cell lines \[^{182-185}\]. The OncoScan\textsuperscript{TM} assay from Affymetrix makes use of this technology and was thus selected for both its low sample requirement and its capacity for dealing with lower quality DNA.

After initial processing, it was determined that 17/30 samples did not pass Affymetrix’s internal quality control (QC) metrics (Table 3.2). Upon closer inspection, it was clear that a large fraction of these samples had extremely noisy genomes and the majority of those which had failed the QC step also had evidence for aberrant ploidy as well as large-scale genomic events, such as chromothripsis. (Figures 3.3, 3.4, Supplementary Figures S1-S31 \[^{186}\]. To better discriminate important gains and losses from genomic noise, I developed a protocol to adjust copy number calls for each sample for their estimated ploidy.
Table 3.2: The quality check status of all 30 cell line samples following processing. A subset of cell lines were observed to have atypical ploidy whereas other cell lines appeared to be near diploid.

Ploidy Estimation and Adjustments

Polyploidy and aneuploidy are commonly observed events in cancer that alter CN profiles and as mentioned previously, many HNCs are aneuploid [42, 43, 187, 188]. The vast majority of HNC cell lines had evidence of increased ploidy relative to the diploid state (Table 3.2, Supplementary Figures S32-S62). While changes to CN states resulting from atypical ploidies are artifacts of cancer biology, massive alterations across a genome make the identification of cancer driver genes difficult. Thus to facilitate the detection of regions of importance, the CN measures across a cell line’s genome were adjusted by the estimated ploidy of the cell line as a way to remove genomic noise in order to determine regions of relative gain or loss. Since ASCAT estimates of ploidy are non-integer values, in the attempt to prevent
Figure 3.3: The copy number profile following standard processing with standard processing with Biodiscovery’s Nexus Express\textsuperscript{TM} for OncoScan for the cell line RF15B with the log R ratio on the top panel and the B-allele frequency on the bottom panel. The CN profile is ordered along the x-axis by genomic position. There appears to be evidence of chromothripsis on chromosomes 7 and 8.

errors from being introduced through rounding as well as information loss, CN estimates were adjusted by the raw, non-integer ploidy estimate, opposed to a rounded whole number. The final adjusted CN values for each sample were used for CNA profiling.

**Consensus Clustering**

CNAs – and noisy genomes in general – have been associated with poorer prognosis as well as differential response to treatment based on the CN change \cite{189,192}. Previous studies have reported patterns of copy number gains and losses for sub-populations of HNC
patients within a cohort [141, 192, 193]. Thus to identify potential patterns that may exist across copy number profiles of the HNC cell lines, consensus clustering was performed for subtype classification, which is an unsupervised machine learning approach often employed for class discovery [147]. The optimal number of clusters, $k$, was selected based on the highest average consensus values for each $k$ number of clusters as well as visual inspections of the consensus cumulative distribution function (CDF) curve and the relative again in the area under the CDF curve with each new cluster (the rule of thumb is to select the $k$ value before the gain starts to plateau) to further support the selection [194] (Figures 3.5, 3.6, 3.7).
Figure 3.5: Consensus Matrix

and Supplementary Figure S63-S75. The optimal number of clusters for this set of CNA profiles was thus determined to be seven clusters.

Copy Number Profiles

The CN subgroups defined by consensus clustering did not separate cell lines based on HPV status, however, it did consistently clustered two cell lines together (UPCI:SCC090 and UPCI:SCC152) that were derived from the same patient – where one cell line (UPCI:SCC090) was derived from a sample taken at initial diagnosis and the other (UPCI:SCC152) was derived from a sample taken at recurrence – confirming that proper and accurate clustering
Figure 3.6: CDF Curves

Figure 3.7: Change in area under CDF curves
Figure 3.8: The ploidy-adjusted copy number (CN) calls for all 30 assayed cell lines, arranged along the x-axis in chromosomal order and along the y-axis by clusters determined through consensus clustering. CNs are binned into “high gain” (> 2 copies), “gain” (1-2 copies), “neutral” (at ploidy), “loss” (1 copy) and “high loss” (2 or more copies). Loss of heterozygosity (LOH) is calculated at both the sample-level (proportion of segments with LOH event in a sample) and at the cohort level (proportion of samples with an LOH event at a particular locus).
took place (Figure 3.8, Supplementary Table S1). Following correction for ploidy, several previously-reported CNAs were detectable upon visual inspection, the most pronounced of these including loss of the short arm of chromosome 3 (3p), gain of the long arm of chromosome 3 (3q) and several aberrations along chromosomes 8 and 9 [195][202].

**Significant Regions**

![Figure 3.9: The distribution of significant copy number altered genes across the genomic regions identified by GISTIC.](image)

Figure 3.9: The distribution of significant copy number altered genes across the genomic regions identified by GISTIC.
<table>
<thead>
<tr>
<th>CytoBand</th>
<th>Type</th>
<th>Genes</th>
<th>q-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>9p21.3</td>
<td>Loss</td>
<td>CDKN2A, C9orf53</td>
<td>3.92 × 10^{-30}</td>
</tr>
<tr>
<td>3p14.2</td>
<td>Loss</td>
<td>FHIT</td>
<td>2.16 × 10^{-9}</td>
</tr>
<tr>
<td>14q32.33</td>
<td>Gain</td>
<td>ADAM6, KIAA0125, LINC00226</td>
<td>3.19 × 10^{-7}</td>
</tr>
<tr>
<td>3q26.1</td>
<td>Gain</td>
<td>hsa-mir-569, hsa-mir-551b, hsa-mir-720, hsa-mir-1263, hsa-mir-16-2, BCHE, ECT2, MECOM, GHSR, IL12A, KPN4A, MLF1, MME, CLDN11, SERPINI1, SERPIN2, PLD1, PRKCI, PTX3, RARRES1, SHOX2, SI, SKIL, SLC2A2, SSR3, TERC, SEC62, KCNAB1, B3GALNT1, TNFSF10, GMP5, SLC33A1, SMC4, PDCD10, SLITRK3, NLGN1, PLCH1, TNK, TIPARP, GPR160, GOLIM4, SCHIP1, NMD3, RSRC1, MYNN, EIF5A2, LIXN, CCNL1, NCEH1, IFT80, SLC7A14, MFSD1, FNDC3B, VEPH1, ZBBX, LRR3C31, PHC3, ARL14, SPATA16, ARPM1, GFM1, EGFEM1P, OTOL1, PPM1L, WDR49, LRR3C34, C3orf55, SPTSSB, RPL22L1, NAALADL2, C3orf33, TRIM59, LOC339894, LRR3Q4, SAMD7, LEKR1, TMEM212, C3orf80, MIR15B, MIR16-2, LOC646168, PA2G4P4, LOC647107, IQCJ, SCARNA7, MIR551B, MIR569, LOC730091, LOC100128164, TIPARP-AS1, LOC100498859, MIR3919, IQCJ-SCHIP1, LOC100507537, MIR4789</td>
<td>6.51 × 10^{-6}</td>
</tr>
<tr>
<td>11q13.3</td>
<td>Gain</td>
<td>hsa-mir-548k, CTTN, FGF3, FGF4, PPFIA1, FADD, FGF19, SHANK2, ANO1, ORAOV1, MIR548K</td>
<td>1.61 × 10^{-4}</td>
</tr>
<tr>
<td>18q21.33</td>
<td>Loss</td>
<td>SERPINB5, SERPINB13, VPS4B, SERPINB12</td>
<td>6.18 × 10^{-4}</td>
</tr>
<tr>
<td>11q22.1</td>
<td>Gain</td>
<td>BIRC2, BIRC3, MMP1, MMP3, MMP7, MMP8, MMP10, MMP12, MMP13, TRPC6, MMP20, YAP1, KIAA1377, MMP27, DYNCH2H1, DCUN1D5, C11orf70, TMEM123, ANGPTL5, LOC100288077, MIR3920</td>
<td>8.32 × 10^{-4}</td>
</tr>
<tr>
<td>3p25.3</td>
<td>Loss</td>
<td>FANCD2, IRAK4, VHL, BRK1, C3orf24, CIDECP</td>
<td>2.53 × 10^{-3}</td>
</tr>
<tr>
<td>9p24.3</td>
<td>Gain</td>
<td>hsa-mir-1302-9, FOXD4, CBWD1, DOCK8, C9orf66, FAM138C, WASH1</td>
<td>5.52 × 10^{-3}</td>
</tr>
<tr>
<td>8p11.22</td>
<td>Loss</td>
<td>ADAM3A, ADAM5P, LOC100130964</td>
<td>7.13 × 10^{-3}</td>
</tr>
<tr>
<td>9p24.1</td>
<td>Loss</td>
<td>PTPRD, C9orf123</td>
<td>1.09 × 10^{-2}</td>
</tr>
<tr>
<td>3p26.3</td>
<td>Loss</td>
<td>CHL1</td>
<td>1.56 × 10^{-2}</td>
</tr>
<tr>
<td>21q22.2</td>
<td>Loss</td>
<td>hsa-mir-3197, DSCAM, MX1, MX2, TFF2, TFF3, TMRSS2, ABCG1, BACE2, C2CD2, ZNF295, LINC00112, LINC00111, FAM3B, RIPK4, PRDM15, UMODL1, LINC00479, ZNF295-AS1, C21orf128, PLAC4, LINC00323, MIR3197, DSCAM-AS1, MIR4760</td>
<td>1.56 × 10^{-2}</td>
</tr>
<tr>
<td>7q21.11</td>
<td>Gain</td>
<td>GRM3, ABCB1, ABCB4, DMTF1, DBF4, TP53TG1, ADAM22, CROT, SLC25A40, C7orf23, RUNDPC3, KIAA1324L, SEMA3D</td>
<td>3.07 × 10^{-2}</td>
</tr>
<tr>
<td>22q11.23</td>
<td>Loss</td>
<td>GSTT1, GSTT2, GSTTP1, LOC391332, GSTTP2</td>
<td>3.31 × 10^{-2}</td>
</tr>
<tr>
<td>2q22.1</td>
<td>Loss</td>
<td>LRP1B, LOC647012</td>
<td>4.98 × 10^{-2}</td>
</tr>
</tbody>
</table>

Table 3.3: Significant copy number gains and losses collapsed by region.
In addition to visually-detectable CNAs, several regions were found to have significant alterations in copy number state following analysis using GISTIC2.0 (Table 3.3, Figure 3.9). Amongst these, several CNAs were reported along the short arm of chromosome 9 (9p) where the most significant change was a loss observed at the 9p21.3 locus, which houses the gene CDKN2A. Previous studies have shown that selective loss of 9p is an early event in HNSCC carcinogenesis and because 9p21-22 was one of the most common aberrations found in HNSCC, it was long speculated to harbour a putative tumour suppressor gene (TSG) [195, 201]. This of course, was later identified to be CDKN2A [195, 201]. In total, 201 statistically-significant (q < 0.05) gene-level CNAs were found. Since CNAs are not necessarily bound by a single gene and typically span across many chromosomal regions, adjacent CNAs of the same kind (i.e. gain or loss) were binned together into the same Cytoband region. This approach was both logical from a biological perspective and was able to reduce the number of genomic candidates to a sensible size (16 CNA-derived features). Consistent with previous studies as well as visual observations, amongst these 16 significant CNA-derived features were regions of loss on chromosome 3p, a large number of gains within a specific region of chromosome 3q as well as several aberrations along chromosomes 8 and 9 that were also statistically significant (Table 3.3 [195, 202]).

**SNV Profile**

**Reference free SNV Classifier: S22S**

The gold standard for identification of somatic SNVs is to simultaneously analyze the sequencing data of primary tumour tissue and that of a patient-matched normal tissue [96]. This facilitates removal of germline mutations and helps control for sequencing errors and analysis artifacts and SNV identification without a reference sample significantly reduces detection accuracy [96, 150, 151]. Many groups resort to ad hoc metrics or repurposing...
tools originally designed for germline analysis, since only a small number of analytical tools accommodate unmatched tumour samples. The most popular approach is to generate a surrogate normal from a pool of normal samples, which only removes false positives resulting from germline contamination, not other types of errors. The resulting datasets anecdotally have high false-positive rates, but no systematic benchmark yet exists. Unfortunately the samples used in the study are cell lines and matched normal references are unavailable. To address this challenge, a random forest classifier was developed, which acts to identify true somatic SNVs from single-sample tumour sequencing data. The algorithm, termed Single-sample Somatic SNV Selector (S22S), classifies SNVs as somatic or not by integrating sequencing characteristics, background genomic context (such as GC content, homopolymer rate and trinucleotide context) and population prevalence of both germline and somatic mutations.

**Experimental Design**

To generate a set of gold-standard SNVs for model training, samples with matched-normal references from the TCGA head and neck squamous cell carcinoma project were used. S22S operates under the hypothesis that true somatic calls possess distinct profiles of sequencing properties, genomic contexts and population frequencies and thus can be discriminated from false calls. To build the feature set, SNVs were called for each sample in four ways: (1) the normal sample was used to identify germline variants using GATK, (2) somatic mutations were identified by SomaticSniper using the T/N pair (3) GATK (HaplotypeCaller) and (4) MuTect with a panel of normal samples (PoN) were used to call SNVs from tumour samples alone to capture characteristics of unmatched analyses (see Materials and Methods). A PoN was generated by running MuTect on each normal sample individually and subsequently aggregating the results. SNVs were then assigned true or false class labels according to the overlap between the four call-sets
To optimize signal for machine-learning, in the training set those (rare) calls that were both somatic (SomaticSniper on T/N pair) and germline (GATK on normals) were removed. For validation, 30% of the samples from each tumour type were set aside. For these, all somatic SNVs detected from T/N analysis were deemed as true positives and all other genomic positions as true negatives. The overall process is outlined in Figure 2.1.

Figure 3.10: Venn diagrams showing the overlap of variants called across the four methods employed germline SNP calling using GATK on tumour/normal pairs (labelled Germline), somatic SNV calling using SomaticSniper on tumour/normal pairs (labelled Somatic) and the two tumour-only SNV calling approaches, GATK on tumour samples alone and MuTect with a panel of normals (PoN) (labelled GATK and MuTect respectively). The overlap between the germline and somatic sets were removed from downstream analyses.

To accurately identify somatic SNVs in the absence of a normal sample, four largely orthogonal data types were leveraged. First, the frequency of known mutations both at the
gene and variant levels in both normal populations and tumour cohorts (where available) was incorporated. Second, metrics indicating sequence quality, such as mapping quality, coverage and position along the read were included. Third, the genomic context of each variant was considered, including elements like homopolymer rate and GC content, which are correlated with errors in somatic SNV prediction [203], and trinucleotide context, to adjust for trinucleotide mutational signatures [204]. Fourth, direct evidence supporting the mutation was measured, for example with the number of reads supporting each allele. As mentioned previously, all features are listed in Table 2.1 These were used for hyper-parameter tuning via 10-fold cross-validation and model fitting in the training cohort (Figure 2.1).

Model Performance

S22S was applied to the held-out test sets (n = 113) and the Area Under the Precision Recall Curve (AUPRC) calculated. In both cases, S22S showed high precision and recall across a wide range of thresholds (Figure 3.1). We selected the operating point (i.e. the threshold used for binary classification, representing a specific precision-recall tradeoff) to be that with the maximum F<sub>1</sub> score. This resulted in a maximum F<sub>1</sub> score of 0.92 presenting 99.96% specificity, 90.8% sensitivity, 93.49% precision and an overall accuracy of 99.91%.

To compare the performance of S22S with alternative approaches, the precision of recall of five different tumour-only approaches were examined. First, the use of a germline caller (GATK) on the tumour sample was considered. Second, the use of MuTect on the tumour sample with a reference normal sample derived from the training dataset was assessed. Subsequently, three additional reference-free SNV prediction tools: Virtual Normal Correction (VNC), EBFilter and VarDict were considered [153, 154, 169]. None of these exceeded a precision of 0.10 for any operating point in either of the two tumours assessed and S22S was Pareto superior to all of them (Figure 3.1).
Figure 3.11: Overall model performance is shown. The model showed great performance on the test set with an AUPRC of 0.98. This is superior to existing reference-free SNV calling methods. Alternatively, cross-tumour application using prostate cancer samples processed in the same manner resulted in reduced AUPRC, yet still above alternative strategies in overall performance.

To demonstrate the impact of improved reference-free somatic classification on applications evaluating mutational landscapes, the recurrently-mutated genes reported by the original
Figure 3.12: S22S improves variant calling for the most recurrently mutated driver genes in the HNSC dataset compared to MuTect.

TCGA HNSC study was investigated. For each recurrent gene, the proportion of correctly predicted mutations (true positive rate) and the proportion of correctly predicted non-mutated genes (true negative rate) were calculated. S22S dramatically improved true negative rates for all recurrent genes (Figure 3.12-3.13).
Figure 3.13: S22S improves variant calling for the most recurrently mutated driver genes in the HNSC dataset compared to GATK.

Model Robustness

While the use of TCGA data provided large cohorts for method benchmarking, smaller studies (e.g. rare tumour types) will have fewer training data available. To assess the sensitivity of S22S performance to training set size, a subset analysis was performed. S22S hyper-parameter optimization and modeling was repeated in a titration series with subsets of
the training dataset and validated on the same independent held-out dataset used previously. Sample sizes of 5, 10, 25, 50, 100, 150 and 200 were used. AUPRC was observed to increase in large steps up to a 50-sample training cohort, with smaller gains from additional samples beyond this threshold (Figure 3.14).

Figure 3.14: S22S is robust to decreased sample-size. PR-curves were constructed from models trained on subsets of the full training data. The lower AUPRC seen with smaller sample sizes demonstrates that adding additional samples will increase the predictive power of the model.
Figure 3.15: Further investigation of how model decisions are made reveals an overlap in important features between the two models. Each model also relies on some features more than others. Several of the most important features are derived from the INFO field of a post-GATK VCF or from population-based allele frequency features.

S22S has been shown to improve prediction accuracy in the presence of a training dataset related to the given experiment. However, there are many situations where disease-specific training data is unavailable. In this scenario, a surrogate dataset may prove sufficient. To assess the performance of S22S in such a case, the model training procedure was repeated on a set of samples from the TCGA prostate adenocarcinoma (PRAD) project [202]. Subsequently, the original HNSC model was applied to PRAD test set and vice
versa. As expected, performance was decreased by training on a different tumour type but the AUPRC remained higher than any other method, reaching an AUPRC of 0.81 for the PRAD-trained model evaluated on the HNSC testing dataset (Figure 3.11). This suggests that a similar set of features is important for prediction performance across tumour types. To assess this, the most discriminative features in each model by quantifying the mean decrease in the Gini coefficient. This is a metric that describes node purity and quantifies a variables usefulness in partitioning observations [205]. The same set of features was important in the two tumour types, although some variability in ranking that accounts for the models differential performance on other tumour types (Figure 3.15).

**Recurrent SNVs of HNC Cell Lines**

Since S22S displayed superior performance to existing methods of SNV calling, it was selected over the other methods to identify somatic SNVs from the HNC cell lines. A default vote threshold of 0.5 was implemented for the S22S random forest model, opposed to the operating point determined from TCGA HNSC data training, as a result of differing data types (cell lines vs. primary tumour samples). After additional germline SNP filtering (see section 2.1.3 for detailed methods) and gene annotation, several recurrently mutated genes were identified across the cohort of 33 cell lines (Figures 3.16 and 3.17). Some (notably TP53 and PIK3CA) were also previously reported in the TCGA HNSC study [141]. These gene mutations were statistically significant following correcting for the basal mutation rate across the genome of a cell line as well as adjustments for multiple testing correction using a false discovery rate of 5%. To ensure sufficiently balanced and powered group sizes for statistical testing, a recurrence level of presence in three samples was imposed. This resulted in a final gene list of 8 SNV-derived features for correlations with drug response (Figure 3.17).
Figure 3.16: Mutations predicted to be recurrent in at least two samples across the cohort of 33 cell lines.
<table>
<thead>
<tr>
<th>Gene</th>
<th>q-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>TTN</td>
<td>$2.78 \times 10^{-11}$</td>
</tr>
<tr>
<td>TP53</td>
<td>$7.63 \times 10^{-15}$</td>
</tr>
<tr>
<td>SYNE1</td>
<td>$2.71 \times 10^{-10}$</td>
</tr>
<tr>
<td>CLDN16</td>
<td>$7.63 \times 10^{-15}$</td>
</tr>
<tr>
<td>PIK3CA</td>
<td>$2.44 \times 10^{-15}$</td>
</tr>
<tr>
<td>AHNAK2</td>
<td>$2.79 \times 10^{-11}$</td>
</tr>
<tr>
<td>TPSD1</td>
<td>$4.97 \times 10^{-12}$</td>
</tr>
<tr>
<td>OBSCN</td>
<td>$7.75 \times 10^{-8}$</td>
</tr>
</tbody>
</table>

Table 3.4: Genes that were recurrent (found in at least three samples) across cell lines and the statistical significance for each, assessed using a Poisson distribution against the basal mutation rate of the sample. $p$-values were adjusted for multiple testing correction at a false discovery rate of 5%.

Figure 3.17: Mutations predicted to be recurrent in at least three samples across the cohort of 33 cell lines and used for correlation with drug response.

**Drug Response**

A large number of compounds have been previously screened via a high-throughput system to select a panel of drugs for further screening at ten dose points [171]. The drugs and
their known targets of action are described in Table S2. The appeal of this panel lies in the fact that a large fraction of drugs screened were either of natural origin or had previous clinical usage and history, making any potential translation into the clinic faster and more...
cost-effective [132]. In addition to the standard approach of assessing drug response using LDR curves, two additional metrics (AA and response classification) were implemented to supplement sensitivity profiling and enhance overall generalizability.

**Sensitivity Assessment**

<table>
<thead>
<tr>
<th></th>
<th>Drugs</th>
<th>Cell Lines</th>
<th>Total Models</th>
<th>IC50</th>
<th>N/A</th>
</tr>
</thead>
<tbody>
<tr>
<td>Count</td>
<td>33</td>
<td>28</td>
<td>1067</td>
<td>300</td>
<td></td>
</tr>
</tbody>
</table>

Table 3.6: Summary of drugs, cell lines and outcomes of model fitting.

Figure 3.18: The distribution of goodness of fit from logistic regression modelling shows that the vast majority (> 95%) of fitted models were between 0.8 and 1.

LDR curves were first generated for each cell line and drug combination – where the IC50 was estimated from this fitted model. In total, there were 33 unique drugs across 28 unique
Figure 3.19: The distribution of the number of cell lines screened per drug.

cell lines (Table 3.6, see Appendix 3 for all LDR curves generated per cell line). Since a few cell lines had replicate drug screens, the total number of models fit for all cell line-drug combinations were 1067, of which, 767 models had an IC$_{50}$ that could be approximated. Of these 767, > 95% had a high goodness of fit (GOF) (Figure 3.18). Although not all cell lines were screened for each drug, there was still a minimum of at least 15 cell lines for any given drug (Figure 3.19). This allowed statistical methods to be applied to most drugs in the panel, although a few were inevitably omitted due to low sample size following quality filters (for goodness of model fit) and sample size filters (to maintain sufficient variability of
Although IC$_{50}$s are a common way of assessing and comparing drug sensitivity, these estimates only capture one aspect of the LDR curve, not to mention, they are not always calculable. Cases where an IC$_{50}$ value cannot be derived are often omitted from downstream analysis. Thus, a second continuous measure of sensitivity – AAs – was included for each cell line and drug pair. In addition to being more robust in sensitivity comparison – both within and across cell lines – as previously demonstrated, AAs also eliminate the issue of missing data and maximize sample size for statistical methods [112]. The range of AAs across the panel of drugs visually demonstrate that while some drugs have a narrow range of response (such as cisplatin), others exhibit differential sensitivity across the cell lines tested (Figure 3.20). Visual inspection also revealed signs of a potential batch effect; however due to the small sample overlap between the two batches, this suspicion could not be confirmed.
A response category was also outlined based on different graphical parameters of the LDR curve (see section 2.2.2 for detailed definitions). This classification was designed with the intent of supplementing either continuous measures of sensitivity (AA or IC\textsubscript{50}) by more broadly summarizing the shape of fitted curve. Profiles of IC\textsubscript{50} values for each drug as well as breakdowns of the response categories in each drug are available in Supplementary Figures S76-S79.

**Statistical Testing**

Figure 3.21: The distribution of \( p \)-values from two-sided Wilcoxon rank-sum tests on AAs based on status of 24 genomic features.
Figure 3.22: The distribution of \( p \)-values from two-sided Wilcoxon rank-sum tests on IC\(_{50}\)s based on status of 24 genomic features.

Since the distributions of both AAs and IC\(_{50}\)s are unknown, to avoid making erroneous assumptions, a non-parametric test – the Wilcoxon rank-sum test – was used to test for statistical significance. A handful of \( p \)-values were noted to be less than the predefined \( \alpha \) of 0.05, however, none of these were statistically significant as these \( p \)-values do not survive correction for multiple testing (Figures 3.21-3.23). Moving beyond a single-feature approach, a more comprehensive method using all of the SNV and CNA features was carried out. This involved modelling the response as a function of each genomic features independently as well as assessing potential synergistic interactions. To mitigate the amount of noise that might propagate through to sensitivity modelling, the drug response data was first filtered for
model GOF greater than 0.9 and subsequently, only drugs with at least 10 samples following the quality filter were analyzed.

**Sensitivity Modelling and Model Power**

A separate GLM was constructed to describe each of three response variables – AA, IC_{50} and response category – for a given drug, as a function of the genomic features. Following feature selection with LASSO, a handful of drugs were found to have associated features that were predictive of a particular response (n_{AA} = 7, n_{IC_{50}} = 2 and n_{response} = 9). Using the stepwise model fitting approach, a final set of features was selected to model each drug,
Figure 3.24: Seven drugs had significant features following feature reduction and several of these were significant at an $\alpha$ of 0.05.

amongst which several were significant at $p$-value $< 0.05$ (AA model profile is shown in Figure 3.24, IC$_{50}$ and response category model profiles are shown in Supplementary Figures S80 and S81). In particular when modelling AA, four genomic interactions were ultimately selected from the stepwise method. Interestingly, two of these four interactions were also statistically significant: chr3p25.3 and AHNAK2 for the drug Pyrrolidine and chr8p11.22
and chr9p24.1 for the drug Rapamycin (Figure 3.24).

To assess how well each model performed at predicting drug sensitivity, the error profiles were estimated using a LOOCV scheme. Here, the RMSE was calculated using the predictions made on the sample that was left out and used to quantify model performance when assessing continuous response variables: AA and IC\textsubscript{50}. The area under the receiver operating curve (AUC) was used for assessing classification performance when predicting categorical response. In general, model performance is high across the response variables modelled – with all RMSEs being relatively small and particularly small for the cisplatin and PRIMA-1 models. Model power is generally high, with most models having a power greater than 80%, estimated from 1000 simulations, and the lowest power being around 70% (for Everolimus, Figure 3.24).

**Power Analysis: Wilcoxon Test**

The power attributed to any statistical test of significance is related to type II errors ($\beta$) or the probability of accepting the null hypothesis when it is false ($1 - \beta$) [239]. In other words, statistical power measures the probability of finding an effect when one exists and governs the quality of inferences drawn from samples – where low power will lead to higher proportions of incorrect or inconclusive results [239]. To assess whether the Wilcoxon tests applied earlier had sufficient power for detecting a potential real effect, a power simulation was conducted to estimate test power. It was observed that the power of each feature differed widely depending on the drug where for some drugs, it was shown that all of the features were powered to detect an effect size of 1 whereas for others, some features were underpowered to detect even large effect sizes (Figures 3.25 and 3.26). However in general, it was determined that the test was well-powered in most cases for detecting large effect sizes and that additional samples may help increase power for identifying the smaller effects. All
Figure 3.25: The power of each feature to detect a given effect size for the drug AZD4547.

Power simulation results are shown in Supplementary Figures S82-S135.
Figure 3.26: The power of each feature to detect a given effect size for the drug SCS.
Discussion

In addition to comprehensively profiling the genomes of 33 HNC cell lines using CNAs and SNVs, we generated several models of drug sensitivity that were well-powered and exhibited high predictive performance. In our CNA analysis, we demonstrated that by using a rigorous approach and adjusting CNA calls for the baseline ploidies for each cell lines, a large fraction of noise can be removed to reveal important chromosomal regions of gains and losses. Notably, we detected a significant loss of the p-arm of chromosome 9, which contains the TSG CDKN2A. The inactivation of CDKN2A is thought to initiate G₁ arrest and prolong cell life via released inhibition of the cyclin D-dependent kinases Cdk4 and Cdk6, thus inactivating apoptosis and other cell regulatory pathways that are generally in place and controlled by the Rb gene. The INK4a-AR gene, which is also found at the 9p21 locus, also arrests proliferation by activating both Rb and TP53. Hence, the loss of this region is thought to prevent initiation of a critical apoptotic pathway. Recent studies suggest that the loss of 9p, along with a loss on the long arm of chromosome 18 – which was also observed in our CNA analysis – is indicative of higher stage tumours [240, 241].

From our SNV analysis, we detected single-base alterations both for genes that were previously characterized in HNC – such as TP53 and PIK3CA – as well as others that are less well-known. Interestingly, SYNE1, a gene that encodes a spectrin repeat containing protein expressed in skeletal and smooth muscle, has been recently described in the context of HNC where a mutation (in addition to several others) was associated with poorer response to induction chemotherapy [242]. While TPSD1 – encoding a tryptase – and CLDN16 – a component of tight junctions primarily found in renal cells – were identified to interact with CNAs for predicting sensitivity to Pyrrolidine, their exact roles have yet to be elucidated.
While HPV status is an important variable governing differential survival as well as response to therapy, we did not find major differences between our HPV-negative and -positive cell lines in terms of drug sensitivity. This may be due to the fact that some of the HPV-positive cell lines were derived from patients who were not characteristic of a case of HNC initiated by HPV. Namely, these were individuals who were smokers and who were also slightly older (opposed to a younger patient who is a non-smoker). This may have lend to similarities between the genomes of the HPV-positive and -negative cell lines, which we did see some evidence for in the genomic analyses. Due to the sparsity of SNVs, statistical tests were not applicable in each case, however, we did not observe striking differences between the genes altered based on HPV status (Figure 3.17). Similarly, consensus clustering of CNA profiles also did not reveal a cluster of HPV-positive cell lines (Figure 3.8). As the resource of HNC cell lines grow – and especially HPV-positive lines – this would be a question that is worth further investigation.

Our approach is intentionally stringent and statistically conservative. The resulting number of false negatives, however, are acceptable to allow us to identify a small set of features that are of high confidence, which is important for both the statistical modelling and the end goal of clinical translation. When it comes to statistical modelling, a common event that occurs when there are too many parameters given a set of observations is “overfitting”. This is when a model learns to predict noise rather than the underlying relationship [243]. Maintaining a sensible number of features also minimizes overfitting – this evident in the low RMSEs from LOOCV – and allow our models to perform well across the cohort of cell lines. But most importantly, we explicitly prefer small and easily-translated biomarkers with high probabilities of validation in this study for clinical utility.

With respect to clinical adoption, since only a very small fraction of the many biomarkers that are currently being researched and developed will eventually be implemented as a part of
routine patient care, a successful biomarker must be robust, with high specificity, sensitivity – amongst other metrics of performance (such as precision and accuracy) – as well as having a short turnaround time [1, 244, 245]. Thus, a simple biomarker will have an advantage over a complicated one in terms of time, cost and ease of communication with physicians and patients.

**Importance**

**Resource**

Although five cell lines are removed from the final correlation analysis due to incomplete data, in general, this is still considered a large cohort with respect to the number HNC cell lines, even when compared to two of the largest cell line drug sensitivity studies published to-date. The Cancer Cell Line Encyclopedia (CCLE), despite being a massive resource composed of 946 cell lines, only had 31 cell lines derived from cancers of the “upper aerodigestive tract”, none of which were known to be HPV-positive lines, and only seven of the 31 HNC cell lines underwent drug sensitivity profiling [112]. Similarly, the Genomics of Drug Sensitivity in Cancer (GDSC) screened 639 cell lines in total, out of which 22 cell lines were derived from the HNC [113]. In contrast, our dataset has the largest set of HPV-positive HNC cell lines, in addition to being comparable in overall number of HNC cell lines to both the CCLE and GDSC datasets, and has genomic information (CNAs and SNVs) on all of them.

**S22S**

Technological advances in DNA sequencing have enabled routine analysis of cancer genomes where identification of SNVs is commonly used to catalog mutational landscapes, to link genomics with drug efficacy and to create clinically useful diagnostics [112, 113, 114, 246–250]. In most of these studies, predictions are made from matched tumour/normal pairs.
and large benchmarks have evaluated their accuracy [203, 251–253]. There is, however, an increasing number of applications where matched normal reference samples are not readily available. Normal samples comprise about half of sequencing costs in clinical studies, and their collection is not always routine or even possible for retrospective cohorts that include deceased patients. Similarly, in cell-line studies, such as those used for drug-screening, rarely have matched normal samples.

We demonstrated accurate prediction of somatic variants without matched reference samples by integrating sequencing data with genomic and population prior knowledge. By using these effectively as a prior, our method – S22S – outperformed current standards of tumour-only SNV calling in all performance metrics, even when little or no training data are available. S22S is agnostic to sequencing platform, and can work with WGS, exome or targeted sequencing targeted panels, and its implementation is freely available. Many somatic SNV callers (both tumour/normal and reference-free) incorporate panels of reference samples, but the distribution of such panels can be restricted by their incorporation of germline SNP data. By contrast, S22S models can be readily shared, and indeed those generated here are available [https://labs.oicr.on.ca/boutros-lab/software/s22s](https://labs.oicr.on.ca/boutros-lab/software/s22s). Large consortia like ExAC, TCGA, ICGC and ICGCmed are ideally placed to generate, in a centralized way, robust models for somatic SNV detection that can be shared with the community as a whole.

**Limitations**

Several issues impacted the quality of our final models. First, since the drug screens were conducted over two separate occasions, it is highly possible that a batch effect exists within the response data. Although these potential effects were visually detectable in a few drugs across the panel (Figure 3.20), they could not be confirmed statistically due to insufficient
sample size. However to minimize noise as a whole, a model GOF filter was applied prior to any statistical testing and modelling.

Secondly, due to technical and budgetary constraints, not all cell lines were screened with all drugs in the panel. This lead to incomplete data and small sample sizes for some drugs, which inevitably forced some to be excluded from downstream modelling. This of course also directly impacted the power of our study. To ensure high power for the models that were fitted, a sample size threshold was implemented. However, additional drug screens – as well as more replicates – would definitely benefit the project as a whole. Nevertheless, this goes towards increasing our false-negative rate, rather than our false positive one.

Lastly, in this project, three metrics were used to quantify drug response. This was done in order to make up for the inability of a single metric to accurately describe the shape of the LDR curve. While the IC$_{50}$s (as well as other inhibition concentrations) are widely used in the field, as a single metric, it is insufficient to characterize the five different parameters of the LDR curve. The AA is slightly more robust in this regard but it can also be improved. Currently, this is a gap in the field and perhaps a novel metric will be needed to resolve this issue.

**Future Directions**

Several tasks can still be done with the existing data to improve the project overall. Firstly, a new method for assessing drug sensitivity that takes into account the slope as well as the top and bottom asymptotes could be developed. This could be trained and validated using either the existing data at-hand or publicly available data (such as that provided by the CCLE). This would reduce the number of response variables assessed and facilitate ease of interpretation. Additional algorithms could also be explored for modelling sensitivity and the
performance of those models compared with the GLM models described here. Additional experiments could also improve the project. For example, mRNA abundance data from microarrays or gene expression from RNA sequencing would enable both the association of genomic alterations to functional alterations as well as pathway analysis, which may unveil additional insights into how changes on the molecular level can affect response to therapeutic agents.

However moving forward, the next step towards clinical translation would be to test the models selected here in an *in vivo* setting. A logical approach would be to first validate all genomic alterations identified and used in drug sensitivity modelling and if the validation is a success, to proceed forward using cell-line derived xenografts models. Here, specific cell lines can be selected based on their known sensitivities and genomic alterations and injected subcutaneously into immune-compromised mice. The efficacy of a treatment and accuracy of a model can then be assessed based on reduction of the engraftment.
Appendix 1: Accurate Reference-Free Somatic Variant-Calling by Integrating Genomic, Sequencing and Population Data

A version of this appendix has been submitted to Nature Methods.

Abstract

The detection of somatic single nucleotide variants (SNVs) is critical in both research and clinical applications. Studies of human disease typically use a matched normal (reference) samples from a distant tissue to increase SNV prediction accuracy. This process both doubles sequencing costs and poses challenges in where reference samples are not readily available, such as cell-lines. To address these challenges, we created S22S: an approach for the prediction of somatic mutations without need for matched references. S22S takes underlying sequence data, augments them with genomic background context and population frequency information, and classifies SNVs as somatic or non-somatic. We validated S22S using 376 primary head and neck tumor/normal pairs and 188 primary prostate tumor/normal pairs. S22S robustly identifies somatic SNVs, with the area under the precision recall curve reaching 0.98 in head and neck cancer, comparable to the best tumor/normal analysis pipelines. S22S is freely available at [http://labs.oicr.on.ca/Boutros-lab/software/s22s](http://labs.oicr.on.ca/Boutros-lab/software/s22s)
Introduction

Technological advances in DNA sequencing have enabled routine analysis of cancer genomes. The identification of somatic single nucleotide variants (SNVs) is commonly used to catalog mutational landscapes, to link genomics with drug efficacy and to create clinically useful diagnostics [112, 113, 115, 246–250]. In most of these studies, predictions are made from matched tumor/normal pairs, and large benchmarks have evaluated their accuracy [203, 251–253]. There is, however, an increasing number of applications where matched normal reference samples are not readily available. Normal samples comprise about half of sequencing costs in clinical studies, and their collection is not always routine or even possible for retrospective cohorts that include deceased patients. Similarly, in cell-line studies, such as those used for drug-screening, rarely have matched normal samples.

However SNV identification without a reference sample significantly reduces detection accuracy [96, 150, 151]. Many groups resort to ad hoc metrics or repurposing tools originally designed for germline analysis since only a small number of analytical tools accommodate unmatched tumor samples [112, 151, 152]. The most popular approach is to generate a surrogate normal from a pool of normal samples, which only removes false positives resulting from germline contamination, not other types of errors [99, 151, 154]. The resulting datasets anecdotally have high false-positive rates, but no systematic benchmark yet exists [96].

To address this challenge, we created Single-sample Somatic SNV Selector (S22S): a random forest classifier that acts to identify true somatic SNVs from single-sample tumor sequencing data. S22S classifies SNVs as somatic or not by integrating sequencing characteristics, background genomic context (such as GC content, homopolymer rate and trinucleotide context) and population prevalence of both germline and somatic mutations (see Online Methods).
Figure 1. (Caption next page.)
Figure 1. (Previous page.) S22S predicts somatic variants without reference samples — A schematic of the algorithm is shown in (A). Per sample, S22S extracts features from the VCF file and from the tumor BAM file and annotates these with population frequency data from publicly available germline and somatic databases to build a feature vector for that sample. We set aside 30% of the samples for validation as an independent test set and aggregate the feature matrices for the remaining 70% of the samples to generate our training set. We used 10-fold cross validation on the training set to parameterize our model. The optimal parameters are used to train a random forest on the full training set. Finally, we assessed model performance on the held-out test set. This procedure was repeated for both tumor types. Overall model performance is shown in (B). Each model performed well in cis-tumor application on the test dataset with an AUPRC of 0.86 for the PRAD model and an AUPRC of 0.98 for the HNSC model. This is superior to existing reference-free SNV calling methods. Alternatively, cross-tumor application results in reduced AUPRC, yet still above alternative strategies in overall performance. Our classifier improves variant calling for the ten most recurrently mutated driver genes in the HNSC dataset (C). S22S is robust to decreased sample-size (D). PR-curves are constructed from models trained on subsets of the full training data for HNSC. The lower AUPRC seen with smaller sample sizes demonstrates that adding additional samples will increase the predictive power of our models. Further investigation of how model decisions are made reveals an overlap in important features between the two models (E). Each model also relies on some features more than others. Several of the most important features are derived from the INFO field of a post-GATK VCF or from population-based allele frequency features.

Results

To generate a set of gold-standard for model training, samples with matched-normal references from two tumor types from the Cancer Genome Atlas (TCGA) network, prostate adenocarcinoma (PRAD) and head and neck squamous cell carcinoma (HNSC), were used [141, 202]. S22S operates under the hypothesis that true somatic calls possess distinct profiles of sequencing properties, genomic contexts and population frequencies and thus can be discriminated from false calls. To build the feature set, SNVs were called for each sample in four ways: (1) the normal sample was used to identify germline variants using GATK, (2) somatic mutations were identified by SomaticSniper using the T/N pair (3)
GATK (HaplotypeCaller) and (4) MuTect with a panel of normal samples (PoN) were used to call SNVs from tumor samples alone to capture characteristics of unmatched analyses (see Online Methods) [95, 97, 99, 156, 157]. A separate PoN was generated for PRAD and HNSC by running MuTect on each normal sample individually and subsequently aggregating the results. SNVs were then assigned true or false class labels according to the overlap between the four call-sets (Supplementary Figure 1). To optimize signal for machine-learning, in the training set those (rare) calls that were both somatic (SomaticSniper on T/N pair) and germline (GATK on normals) were removed (Supplementary Tables 1-2, Supplementary Figure 1). For validation, 30% of the samples from each tumor type were set aside. For these, all somatic SNVs detected from T/N analysis were deemed as true positives and all other genomic positions as true negatives. The overall process is outlined in Figure 1A.

To accurately identify somatic SNVs in the absence of a normal sample, we leveraged four largely orthogonal data types. First, we incorporated the frequency of known mutations – both at the gene and variant levels – in both normal populations and tumor cohorts (where available). Second, we included metrics indicating sequence quality, such as mapping quality, coverage and position along the read. Third, the genomic context of each variant was considered, including elements like homopolymer rate and GC content, which are correlated with errors in somatic SNV prediction, and trinucleotide context, to adjust for trinucleotide mutational signatures [203, 204]. Fourth, direct evidence supporting the mutation was measured, for example with the number of reads supporting each allele. Table 1 lists all features. These were used for hyper-parameter tuning via 10-fold cross-validation and model fitting in the training cohort (Figure 1A).

The evaluate the performance of S22S we used two held-out test datasets (n_{HNSC} = 113, n_{PRAD} = 56) and assessed the Area Under the Precision Recall Curve (AUPRC). In both test datasets S22S showed high precision and recall (Figure 1B). We selected our operating
<table>
<thead>
<tr>
<th>FEATURE ALIAS</th>
<th>FEATURE DESCRIPTION</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>GATK-derived Features</strong></td>
<td></td>
</tr>
<tr>
<td>AC</td>
<td>Allele count for ALT allele</td>
</tr>
<tr>
<td>AF</td>
<td>Allele frequency for ALT allele</td>
</tr>
<tr>
<td>AN</td>
<td>Total number of alleles called in genotype</td>
</tr>
<tr>
<td>BaseQRankSum</td>
<td>Z-score from Wilcoxon rank sum test of ALT Vs. REF base qualities</td>
</tr>
<tr>
<td>ClippingRankSum</td>
<td>Z-score From Wilcoxon rank sum test of Alt vs. Ref number of hard clipped bases</td>
</tr>
<tr>
<td>DP</td>
<td>Approximate read depth</td>
</tr>
<tr>
<td>FS</td>
<td>Phred-scaled p-value using Fisher’s exact test to detect strand bias</td>
</tr>
<tr>
<td>MLEAC</td>
<td>Maximum likelihood expectation (MLE) for the allele counts</td>
</tr>
<tr>
<td>MLEAF</td>
<td>Maximum likelihood expectation (MLE) for the allele frequency</td>
</tr>
<tr>
<td>MQ</td>
<td>Root Mean-squared (RMS) Mapping Quality</td>
</tr>
<tr>
<td>MQRankSum</td>
<td>Z-score From Wilcoxon rank sum test of ALT vs. REF read mapping qualities</td>
</tr>
<tr>
<td>ReadPosRankSum</td>
<td>Z-score from Wilcoxon rank sum test of ALT vs. REF read position bias</td>
</tr>
<tr>
<td>SOR</td>
<td>Symmetric Odds Ratio of 2x2 contingency table to detect strand bias</td>
</tr>
<tr>
<td>REF.AD</td>
<td>Allele depth for REF</td>
</tr>
<tr>
<td>ALT.AD</td>
<td>Allelic depth for ALT</td>
</tr>
<tr>
<td>REF.AF</td>
<td>Calculated allelic frequency of REF (REF.AD/DP)</td>
</tr>
<tr>
<td>ALT.AF</td>
<td>Calculated allelic frequency of ALT (ALT.AD/DP)</td>
</tr>
<tr>
<td>mutect</td>
<td>Presence of call using MuTect with a panel of normals</td>
</tr>
<tr>
<td>gatk</td>
<td>Presence of call using GATK in tumour-only mode</td>
</tr>
<tr>
<td><strong>Population Features</strong></td>
<td></td>
</tr>
<tr>
<td>NHLBIEA.MAF</td>
<td>Minor allele frequency in percent for European population from NHLBI</td>
</tr>
<tr>
<td>NHLBLAA.MAF</td>
<td>Minor allele frequency in percent for African American population from NHLBI</td>
</tr>
<tr>
<td>NHLBI.ALL.MAF</td>
<td>Overall minor allele frequency from NHLBI</td>
</tr>
<tr>
<td>AF.1000Genomes</td>
<td>Global allele frequency based on AC/AN</td>
</tr>
<tr>
<td>AFR.AF</td>
<td>Allele frequency for the African population based on AC/AN from 1000 Genomes Project</td>
</tr>
<tr>
<td>AMR.AF</td>
<td>Allele frequency for the American population based on AC/AN from 1000 Genomes Project</td>
</tr>
<tr>
<td>ASN.AF</td>
<td>Allele frequency for the Asian population based on AC/AN from 1000 Genomes Project</td>
</tr>
<tr>
<td>EUR.AF</td>
<td>Allele frequency for the European population based on AC/AN from 1000 Genomes Project</td>
</tr>
<tr>
<td>LDAF</td>
<td>Allele frequency as inferred from haplotype estimation</td>
</tr>
<tr>
<td>TCGA.fraction.MAF</td>
<td>Proportion of recurrence in the TCGA dataset, position-based</td>
</tr>
<tr>
<td>TCGA.fraction.RecSNV</td>
<td>Proportion of recurrence in the TCGA dataset, gene-based</td>
</tr>
<tr>
<td><strong>Sequencing Features</strong></td>
<td></td>
</tr>
<tr>
<td>MappingQual</td>
<td>Mapping quality scores (in phred-scale) for probability of misplaced read</td>
</tr>
<tr>
<td>ReadPosition</td>
<td>Position within the read</td>
</tr>
<tr>
<td>TumourCoverage</td>
<td>Read depth from tumour bam file</td>
</tr>
<tr>
<td>DistanceSNP</td>
<td>Distance (in bases) to closest known germline SNP</td>
</tr>
<tr>
<td>DistanceIndel</td>
<td>Distance (in bases) to closest known germline Indel</td>
</tr>
<tr>
<td>StrandBias</td>
<td>Proportion of reads mapped to forward strand</td>
</tr>
<tr>
<td>RefForward</td>
<td>Quality of at least 13 REF bases on forward strand</td>
</tr>
<tr>
<td>RefReverse</td>
<td>Quality of at least 13 REF bases on reverse strand</td>
</tr>
<tr>
<td>NonRefForward</td>
<td>Quality of at least 13 non-REF bases on forward strand</td>
</tr>
<tr>
<td>NonRefReverse</td>
<td>Quality of at least 13 non-REF bases on reverse strand</td>
</tr>
<tr>
<td>SumRefBase</td>
<td>Sum of reference base qualities</td>
</tr>
<tr>
<td>SumSqRefBase</td>
<td>Sum of squares of reference base qualities</td>
</tr>
<tr>
<td>SumNonRefBase</td>
<td>Sum of non-ref base qualities</td>
</tr>
<tr>
<td>SumSqNonRefBase</td>
<td>Sum of squares of non-ref base qualities</td>
</tr>
<tr>
<td>SumRefMap</td>
<td>Sum of ref mapping qualities</td>
</tr>
<tr>
<td>SumSqRefMap</td>
<td>Sum of squares of ref mapping qualities</td>
</tr>
<tr>
<td>SumNonRefMap</td>
<td>Sum of non-ref mapping qualities</td>
</tr>
<tr>
<td>SumSqNonRefMap</td>
<td>Sum of squares of non-ref mapping qualities</td>
</tr>
<tr>
<td>SumTailRef</td>
<td>Sum of tail distance for ref bases</td>
</tr>
<tr>
<td>SumSqTailRef</td>
<td>Sum of squares of tail distance for ref bases</td>
</tr>
<tr>
<td>SumTailNonRef</td>
<td>Sum of tail distance for non-ref</td>
</tr>
<tr>
<td>SumSqTailNonRef</td>
<td>Sum of squares of tail distance for non-ref</td>
</tr>
<tr>
<td>RefAllele</td>
<td>Forward ref allele</td>
</tr>
<tr>
<td>NonRef Allele</td>
<td>Forward non-ref allele</td>
</tr>
<tr>
<td>BaseQual</td>
<td>Mapping quality scores (in phred-scale) for probability of misidentified base</td>
</tr>
<tr>
<td>DP.pileup</td>
<td>Number of reads covering or bridging position</td>
</tr>
<tr>
<td>VDB</td>
<td>Variant distance bias for indication of potential misalignment due to nearby SNP</td>
</tr>
<tr>
<td><strong>Genomic Context</strong></td>
<td></td>
</tr>
<tr>
<td>homopolymer.rate</td>
<td>Frequency of homopolymers within 100 bases up and downstream of position</td>
</tr>
<tr>
<td>GC</td>
<td>Proportion of bases that are GC within 100 bases up and downstream of position</td>
</tr>
<tr>
<td>trinucleotide</td>
<td>Trinucleotide context of position</td>
</tr>
</tbody>
</table>

Table 1. (Caption next page.)
Table 1. Classification Features — We extracted a total of 59 features for PRAD and 61 features for HNSC. Features are categorized based on origin and purpose. The PRAD model has two fewer features than the HNSC model (ClippingRankSum and SOR), as it was analyzed using an older version of GATK which lacked these two metrics.

point (i.e. the threshold used for binary classification, representing a specific precision-recall tradeoff) to be that with the maximum F\textsubscript{1} score. In HNSC, this resulted in a maximum F\textsubscript{1} score of 0.92 – presenting 99.96% specificity, 90.8% sensitivity, 93.49% precision and an overall accuracy of 99.91% (Supplementary Table 3). The PRAD dataset had significantly fewer true somatic mutations (µ\textsubscript{HNSC} = 184 vs. µ\textsubscript{PRAD} = 23, Students t-test, p < 2.2 × 10^{-16}), leading to a smaller training dataset and a lower F\textsubscript{1} score of 0.82. Nevertheless, this represented specificity of 99.98% and sensitivity of 86.6% (Supplementary Table 4).

To compare the performance of S22S with alternative approaches, we examined the precision of recall of five different tumor-only approaches. We first considered the use of a germline caller (GATK) on the tumor sample. Second, we looked at the use of MuTect on the tumor sample with a reference normal sample derived from the training dataset. We also considered three additional reference-free SNV prediction tools: Virtual Normal Correction (VNC), EBFilter and VarDict \cite{153,154,169}. None of these exceeded a precision of 0.10 for any operating point in either of the two tumors assessed, and S22S was Pareto superior to all of them (Figure 1B, Supplementary Table 3-4).

To demonstrate the impact of improved reference-free somatic classification on applications evaluating mutational landscapes, we looked at the recurrently-mutated genes reported by the original TCGA HNSC and PRAD studies respectively. For each recurrent gene, we calculated the proportion of correctly predicted mutations (true positive rate) and the proportion of correctly predicted non-mutated genes (true negative rate). S22S dramatically improved true negative rates for all recurrent genes (Figure 1C, Supplementary Figure 2).
While the use of TCGA data provided large cohorts for method benchmarking, smaller studies (e.g. rare tumor types) will have fewer training data available. To assess the sensitivity of S22S performance to training set size, we performed a subset analysis. S22S hyper-parameter optimization and modeling was repeated in a titration series with subsets of the training dataset and validated on the same independent held-out dataset used previously. We used samples of 5, 10, 25, 50 and 100 for PRAD and 5, 10, 25, 50, 100, 150 and 200 for HNSC. For both diseases, AUPRC increased in large steps up to 50 sample training cohorts, with smaller gains from additional samples beyond this threshold (Figure 1D, Supplementary Figure 3).

S22S has been shown to improve prediction accuracy in the presence of a training dataset related to the given experiment. However, there are many situations where disease-specific training data is unavailable. In this scenario, a surrogate dataset may prove sufficient. To assess the performance of S22S in such a case, we applied the HNSC model to PRAD test set and vice versa. As expected, performance was decreased by training on a different tumor type but the AUPRC remained higher than any other method, reaching an AUPRC of 0.81 for the PRAD-trained model evaluated on the HNSC testing dataset (Figure 1B). This suggests that a similar set of features is important for prediction performance across tumor types. To assess this, we evaluated the most discriminative features in each model by quantifying the mean decrease in the Gini coefficient. This is a metric that describes node purity and quantifies a variables usefulness in partitioning observations \[205\]. The same set of features was important in the two tumor types, although some variability in ranking that accounts for the models differential performance on other tumor types (Figure 1E).
Discussion

In summary, we have demonstrated accurate prediction of somatic variants without matched reference samples by integrating sequencing data with genomic and population prior knowledge. By using these effectively as a prior, S22S outperforms current standards of tumor-only SNV calling in all performance metrics, even when little or no training data are available. S22S is agnostic to sequencing platform, and can work with WGS, exome or targeted sequencing targeted panels, and its implementation is freely available. Many somatic SNV callers (both tumor/normal and reference-free) incorporate panels of reference samples, but the distribution of such panels can be restricted by their incorporation of germline SNP data. By contrast, S22S models can be readily shared, and indeed those generated here are available (Supplementary Data 1-2). Large consortia like ExAC, TCGA, ICGC and ICGCmed are ideally placed to generate, in a centralized way, robust models for somatic SNV detection that can be shared with the community as a whole.

Online Methods

Datasets

To build the training and test sets for our models, we leveraged publicly-available exome sequencing samples of primary tumors and their matched normal samples from two cancer types published by the The Cancer Genome Atlas (TCGA) network: prostate cancer (PRAD) and head and neck cancer (HNSC). BAM files were obtained from the Cancer Genomics Hub (CGHub, [https://cghub.ucsc.edu/](https://cghub.ucsc.edu/)) and realigned and recalibrated prior to downstream analyses [141, 155, 202]. Our final HNSC cohort is composed of 376 tumor/normal pairs while our PRAD cohort is composed of 188 tumor/normal pairs. The training and testing
matrices for model building for HNSC and PRAD are available in Supplementary Data 3-4 and Supplementary Data 5-6 respectively.

**Defining True/False Classes**

In order to train our model to accurately recognize true calls from false ones, we took a comprehensive approach when defining our truth and false sets. To create an adequate feature space, we performed SNV calling (focusing on substitutions rather than indels) in four manners. First, we called germline and somatic SNVs using the standard approach with tumor/normal pairs using GATK and SomaticSniper respectively [95, 97, 156, 157]. We also performed tumor-only SNV calling using GATK in tumor-only mode and MuTect with a panel of normal samples generated from the normal samples of the respective tumor type [99]. We defined the true class using variants identified from SomaticSniper on paired samples that were absent in the GATK germline calls. The false class encompassed all calls that were present in the germline calls, both tumor-only call sets and not found in the SomaticSniper set (Figure 1A). To eliminate ambiguity for our learner, we elected to omit any overlaps found between the germline and somatic calls from downstream analysis.

**Re-Alignment of Exome Sequencing Data**

For each BAM file obtained from CGHub, back-conversion to FASTQ files was done to allow realignment to the human reference genome for standardization, using the SamToFastq function from picard (v.1.92) [http://broadinstitute.github.io/picard](http://broadinstitute.github.io/picard). For PRAD, lane-level raw sequencing reads were realigned to human reference build hg19 using bwa aln (v0.5.7), while for HNSC, the realignment was performed with the human reference build GRCh37 with decoy (hs37d5) and bwa mem (v0.7.12) [93]. Merging of lane-level BAMs from the same library within each sample was facilitated via picard (v1.92), with
duplicates marked, and was followed by library-level merging of BAMs, without marking of duplicates. We used GATK (v2.4.9 for PRAD samples and v3.4.0 for HNSC samples) to perform local realignment and base quality recalibration on each realigned tumor/normal BAM pair [156]. Separate tumor and normal sample-level BAM files were extracted, followed by header correction using samtools (v0.1.19) and indexing using picard (v1.107) to generate the final realigned and recalibrated BAM file per sample.

**Germline SNP Calling**

We used GATK23 (v2.4.9 for PRAD and v3.4.0 for HNSC) to call germline single nucleotide polymorphisms (SNPs). For each PRAD sample, we used UnifiedGenotyper, followed by VariantRecalibrator and ApplyRecalibration on the realigned and recalibrated tumor/normal pair and removed all insertions/deletions (INDELs), somatic SNVs and ambiguous SNVs with more than one alternate base to obtain our final germline VCF callset. For each HNSC sample, we used HaplotypeCaller followed by VariantFiltration to hard-filter our callset using the following filter expressions: “QD > 10.0 || FS > 60.0 || MQ < 40.0 || DP < 50 || SOR > 4.0 || ReadPosRankSum < -8.0 || MQRankSum < -12.5” and “MQ0 >= 4 && ((MQ0 (1.0 * DP)) > 0.1)” to generate the final germline calls. We referred to the GATK Best Practices recommendations for the development of this pipeline [95, 157].

**Somatic SNV Calling**

We predicted somatic SNVs using SomaticSniper25 (v1.0.2 for PRAD and v1.0.4 for HNSC). First, candidate somatic calls were identified using bam-somaticsniper, with all default settings except for the -q option (mapping quality threshold) which was set to 1 instead of the default 0 at the recommendation of the developer. We then used a series of Perl scripts provided by the SomaticSniper package to filter out possible false positives
We filtered our candidate list using several methods. First, we applied standard and LOH filtering using pileup INDEL files generated for both the normal and tumor BAM files using SAMtools (v1.0.6). Then, we applied a bam-readcount filter (bam-readcount downloaded on Jan 10th, 2014) with a mapping quality filter -q 1 (using default settings otherwise). We also ran the false positive filter and a high confidence filter (with default parameters). The resulting VCF file with high confidence somatic SNV calls for each sample was subsequently filtered against several lists of known germlines to further reduce potential false positives.

In our filtering approach, we removed SNVs that were found in any of the following databases (also referred to as blacklists) using tabix to produce the final set of somatic calls: dbSNP142 (modified to remove somatic and clinical variants, with variants with the following flags excluded: SAO = 2/3, PM, CDA, TPA, MUT and OM), 1000 Genomes Project (v3), Complete Genomics 69 whole genomes, duplicate gene database (v68), ENCODE DAC and Duke Mapability Consensus Excludable databases (comprising poorly mapping reads, repeat regions, and mitochondrial and ribosomal DNA), invalidated somatic SNVs from 68 human colorectal cancer exomes (unpublished data) using the AccuSNP platform (Roche NimbleGen), germline SNPs from 477 samples used in our previous work in prostate cancer with an additional 10 WGS from prostate cancer patients with higher Gleason score (Fraser et al., Nature, in revision) and the Fuentes database of likely false positive variants; SNVs were whitelisted or kept (independent of presence in other databases) if they were found in the Catalogue of Somatic Mutations in Cancer (COSMIC) database (v71) [100, 158–164].

**Tumor-only SNV Calling**

To capture the error profiles of SNVs called in situations where a normal sample was not used, we also performed somatic SNV calling using only the tumor samples. We used two analytical
approaches that were most frequently employed in this situation, as seen and discussed across various sequencing forums. The first approach was to run variant calling with GATK in tumor-only mode and to retain the VCF file at the end of the “VariantFiltration” step following hard-filtering as the list of variants for the tumor sample. The second approach was to generate a panel of normal samples (PoN) by pooling together a cohort of normal samples using MuTect and to call mutations with the pooled list of germline variants serving as the normal surrogate \cite{99}. For this, we constructed two panels of normal samples, one for each of our tumor types, using the normal BAM files. The procedure for creating a PoN has been documented previously but in brief, each normal BAM file was passed to the tool separately under the input label of ‘tumor’ with the “artefact_detection_mode” set on \cite{99}. The output vcf per sample was merged together using CombineVariants from the GATK engine to generate the final normal panel. This process was done for both the PRAD normal samples and the HNSC normal samples to generate a PoN per tumor type.

Features

We hypothesize that true variants will exhibit different underlying sequencing properties, genomic background characteristics and population distributions when compared to germline variants or sequencing artifacts. Thus to characterize this, we extracted a set of features, based on variant positions found in the final callset, from the VCF file, following GATK processing with HaplotypeCaller, and the tumor BAM file, and annotated these with minor allele and population frequencies from the 1000 Genomes Project and NHLBI \cite{100, 165} in addition to TCGA recurrences calculated at both the position level and gene level using the latest release of TCGA MAF files. This resulted in a total of 59 features for the PRAD model and 61 features for the HNSC model. The discrepancy between the total number of features across the two models is a result of different versions of GATK used during processing (ClippingRankSum and SOR are features missing from the PRAD model).
full set of features used is described in Table 1.

**Model Training**

For our classifier, we chose to use a random forest (RF) an ensemble approach that is resistant to outliers and can effectively handle highly-correlated variables [166]. We trained our model using the randomForest package (v4.6-10) in the R statistical environment (v3.2.3). For each tumor type, we set aside 30% of our samples aside as a held-out test set and performed a grid search to parameterize our random forest using our 70% training set in a 10-fold cross-validation scheme (Figure 1A). To split samples into 70% training and 30% testing, we rounded 30% of the sample size per tumor type to the nearest integer and set this as the number of samples to allocate for each of our test sets. We then used the base function “sample” from the “stats” package in the R statistical environment (v3.2.3) with the seed ‘333’ to obtain indices to assign alphabetically-sorted sample names to the test set. The samples that remained were assigned to training.

**Random Forest Parameterization**

It has been noted that some parameters of random forest are more sensitive to tuning than others [167, 205]. In addition, due to the nature of our dataset, a large class imbalance exists between the negative and positive classes (the number of germline SNPs called greatly outweigh the number of somatic SNVs called). Thus to improve our predictive power, we elected to down-sample the major class as a function of the minor class and performed a grid search to tune parameters of random forest using our training set. We assessed the performance using a 10-fold cross-validation scheme to obtain the most optimal values to use for our full model (Figure 1A). We chose to tune three parameters for our HNSC dataset (mtry, nodesize and ntree) and four parameters for our PRAD dataset (mtry, nodesize, ntree and sampsize).
For the parameter mtry, we elected to test factor levels of the default value. Since the nature of our problem is classification and the default mtry for classification is the square root of the number of features, which in our case was 7, we thus used, in addition to the default value of 7, half of the default (4), twice the default (14) and three times the default (21). For ntree, we chose to test 1000, 5000 and 10,000 and values of 5, 25, 50 and 100 for the parameter nodesize. For our HNSC dataset, we down-sampled our major class at a ratio of 1:1 with our negative class but for our PRAD dataset, we also tested ratios of 5:1, 10:1, 20:1 and 50:1. Prostate cancer has been known to possess fewer somatic mutations than most other cancers [254]. Thus, down-sampling our data at a ratio of one-to-one would insufficiently cover the entirety of feature space for the negative class.

To split our 70% training set into 10 reasonably equal sets, we took an approach that was similar to the splitting of our training and test sets. We truncated what was 10% of the number of samples in our training set to the nearest integer and set this as the number of samples to allocate per fold. We then used the base function sample from the stats package in the R statistical environment (v3.2.3) with the seed ‘99’ to obtain array indices for our vector of sample names. Each time an index was chosen, that was taken out of the pool of indices we may sample from. We repeated this process nine times and used the indices to assign each sample to its respective set, with the remaining samples assigned to fold 10. The last fold also, on average, contained 1-2 more samples than the rest of the folds. For our grid search parameterization, we kept each fold consistent across all parameter combinations tested.

For each parameter combination, we calculated the precision, recall as well as the area under the precision-recall curve (AUPRC). We ranked each parameter combination based on the average area under the precision-recall curve AUPRC across ten folds (ties were broken using the standard deviations where a higher rank was attributed to the lower standard deviation).
and chose the set of parameters that generated the highest overall AUPRC. For the HNSC model, we proceeded to train the full model using mtry of 4, nodesize of 5 and 10,000 trees and for the PRAD model, we proceeded to train the full model using mtry of 4, nodesize of 5, 10,000 trees and a sampling ratio (sampsize) of 50:1 for false to true calls.

**Performance Assessment**

By varying the vote threshold from 0 to 1, we were able to calculate, across a continuous range, the number of false positive (FP), false negative (FN), true positive (TP) and true negative (TN) calls by using different cutoffs. This was facilitated by the pROC package43 (v1.8) in the R statistical environment (v3.2.3). We then used these to calculate metrics for assessing model performance such as sensitivity, specificity and precision. We also constructed a curve using our continuous range of precision and recall values and used the area under this precision-recall curve (AUPRC) as the main metric for comparing model performance. This was done using the trapezoid and calculated by:

\[
AUPRC = \frac{1}{2} \sum_{k=1}^{N} (x_{k+1} - x_k)(y_{k+1} - y_k)
\]  

(4.1)

where \(x\) is the the recall and \(y\) is the precision at cutoff \(k\). The set of parameters with the highest average rank of AUPRC across the ten folds was selected for the full model.

When selecting an operating point for our models based on the AUPRC, we elected to pick a threshold that maximized the harmonic mean of precision and recall, also known as the \(F_1\) score, which was calculated by:

\[
F_1 = \frac{2 \times precision \times recall}{precision + recall}
\]  

(4.2)
where precision is defined as:

\[
\text{precision} = \frac{TP}{TP + FP}
\]  

(4.3)

and recall is defined as:

\[
\text{recall} = \frac{TP}{TP + FN}
\]

(4.4)

For HNSC, an operating point of 0.93525 was selected while for PRAD, an operating point of 0.76175 was chosen.

**Cross-tumor Performance and Model Convergence**

We took two approaches to gauge the generalizability of each model and assessed the effects of both tumor type and cohort size on overall performance. First, we applied each model to the test set of the other tumor and evaluated its performance using the AUPRC metric. Due to discrepancies between features across models, when testing the HNSC model on the PRAD test set, we manually set the two missing features (ClippingRankSum and SOR) to 0 for all positions in the test set. Alternatively when testing the PRAD model on the HNSC test set, we omitted those two features from the test data frame when predicting using our RF classifier.

Second, to determine the effects of training size on model performance, we conducted a convergence experiment using reasonably-selected sample sizes in our training set and assessed its performance on the same test set by calculating the AUPRC. For the PRAD model, we tested sample sizes of 5, 10, 25, 50 and 100 and compared their AUPRCs to the full model (n = 132). Similarly for our HNSC model, we tested sample sizes of 5, 10, 25, 50, 100, 150, 200 and compared their AUPRCs to our full model (n = 263).
Data Visualization

Visualizations were generated in the R statistical environment (v3.2.3) using the lattice (v0.20-33), latticeExtra (v0.6-28), BPG (v5.6.8) and VennDiagram (1.6.17) packages \cite{148, 149}. Figures were compiled using LaTeX.

Benchmarking

To assess how our approach compares to methods developed under similar pretenses, we applied three published tools (EBFilter, VNC and VarDict) to our HNSC and PRAD test sets of 113 and 56 samples \cite{153, 154, 169}. We selected these tools based on their proposed functionality and the fact that they accepted similar inputs and generated comparable outputs to our method. We used the same AUPRC metric to assess performance across the different algorithms. Since each tool had its own specific method or score for annotating potential true variants or variants of interest, to generate a curve for each, we varied the threshold for each tool across its full range of values for the entire cohort. All tools were run with default options or with those suggested in the user guidelines for the tool unless otherwise specified.

VarDict was executable given a VCF file, BAM file and the human genome reference fasta as inputs, however, some of the other tools required additional input files. Both EBFilter and VNC took an approach that was similar to MuTect in that both required an additional list of normal samples to create a surrogate normal. For EBFilter, we compiled a list of all of the normal samples used towards the generation of our PRAD PoN and used this panel in all of our EBFilter runs. For VNC, we downloaded all available (n = 427) variant files of normal samples that were processed by Complete Genomics (CG) for the 1000Genomes Project from the FTP site \url{ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase3/data/}. We
used these samples instead of TCGA normal samples because the tool was designed to work with CG outputs and only accepted the varfile format as input to generate the virtual normal. To generate a range of precision and recall values for EBFilter, we varied the score that was outputted by the tool from 0 to the maximum score of the cohort. Since VNC implemented two fields for filtering calls – both relating to the number of samples out of the total pool of samples used in the construction of the virtual normal in which a variant was found – we varied both thresholds from 0 to 427 and calculated precision and recall at each combination of thresholds. In single-sample mode, VarDict implements a single allele frequency (AF) filter during variant identification, so to calculate precision and recall for the PR-curve, we varied the AF from 0 to 1. When selecting operating points for each tool, we used the same approach as with S22S and chose that which maximized the F₁ score.

Availability

Software and models are available at [http://labs.oicr.on.ca/Boutros-lab/software/s22s](http://labs.oicr.on.ca/Boutros-lab/software/s22s).
Appendix 2: Supplementary Materials

OncoScan Profiles

Figure S1: OncoScan profile for Affymetrix in-house reference cell line Ref103
Figure S2: OncoScan profile for cell line 93VU147T

Figure S3: OncoScan profile for cell line BICR56
Figure S4: OncoScan profile for cell line Cal27

Figure S5: OncoScan profile for cell line Cal33
Figure S6: OncoScan profile for cell line Detroit562

Figure S7: OncoScan profile for cell line FaDu
Figure S8: OncoScan profile for cell line HMS001

Figure S9: OncoScan profile for cell line HSC2
Figure S10: OncoScan profile for cell line JHU006

Figure S11: OncoScan profile for cell line JHU011
Figure S12: OncoScan profile for cell line JHU029

Figure S13: OncoScan profile for cell line PCI6A
Figure S14: OncoScan profile for cell line PCI6B

Figure S15: OncoScan profile for cell line PCI13
Figure S16: OncoScan profile for cell line PCI30

Figure S17: OncoScan profile for cell line PE/CA:PJ-49
Figure S18: OncoScan profile for cell line RF15A

Figure S19: OncoScan profile for cell line RF15B
Figure S20: OncoScan profile for cell line RF22A

Figure S21: OncoScan profile for cell line RF22B
Figure S22: OncoScan profile for cell line RF37A

Figure S23: OncoScan profile for cell line RF37B
Figure S24: OncoScan profile for cell line SCC4

Figure S25: OncoScan profile for cell line SCC9
Figure S26: OncoScan profile for cell line SCC15

Figure S27: OncoScan profile for cell line SCC25
Figure S28: OncoScan profile for cell line SCC47

Figure S29: OncoScan profile for cell line UPCI:SCC090
Figure S30: OncoScan profile for cell line UPCI:SCC152

Figure S31: OncoScan profile for cell line UPCI:SCC154
ASCAT Profiles

Figure S32: ASCAT profile for Affymetrix in-house reference cell line Ref103
Figure S33: ASCAT profile for cell line 93Vu147T

Figure S34: ASCAT profile for cell line BICR56
Figure S35: ASCAT profile for cell line Cal27

Figure S36: ASCAT profile for cell line Cal33
Figure S37: ASCAT profile for cell line Detroit562

Figure S38: ASCAT profile for cell line FaDu
Figure S39: ASCAT profile for cell line HMS001

Figure S40: ASCAT profile for cell line HSC2
Figure S41: ASCAT profile for cell line JHU006

Figure S42: ASCAT profile for cell line JHU011
Figure S43: ASCAT profile for cell line JHU029

Figure S44: ASCAT profile for cell line PCI6A
Figure S45: ASCAT profile for cell line PCI6B

Figure S46: ASCAT profile for cell line PCI13
Figure S47: ASCAT profile for cell line PCI30

Figure S48: ASCAT profile for cell line PE/CA:PJ-49
Figure S49: ASCAT profile for cell line RF15A

Figure S50: ASCAT profile for cell line RF15B
Figure S51: ASCAT profile for cell line RF22A

Figure S52: ASCAT profile for cell line RF22B
Figure S53: ASCAT profile for cell line RF37A

Figure S54: ASCAT profile for cell line RF37B
Figure S55: ASCAT profile for cell line SCC4

Figure S56: ASCAT profile for cell line SCC9
Figure S57: ASCAT profile for cell line SCC15

Figure S58: ASCAT profile for cell line SCC25
Figure S59: ASCAT profile for cell line SCC47

Figure S60: ASCAT profile for cell line UPCI:SCC090
Figure S61: ASCAT profile for cell line UPCI:SCC152

Figure S62: ASCAT profile for cell line UPCI:SCC154
Consensus Clustering

Figure S63: k=2.

Figure S64: k=3.

Figure S65: k=4.

Figure S66: k=5.
Figure S71: $k=11$.

Figure S72: $k=12$.

Figure S73: $k=13$.

Figure S74: $k=14$. 
Figure S75: $k=15$. 
Drug Profiling

Figure S76: The distribution of IC$_{50}$s calculated for each drug sorted by median.

Figure S77: The distribution of IC$_{50}$s calculated for each drug sorted by median, focused between responses of -2 and 3 in log$_{10}$ space.
Figure S78: The proportion of response classifications across samples for each drug, sorted by decreasing proportions of sensitive profiles.
Figure S79: The proportion of response classifications across samples with a goodness of fit of at least 0.9 and a calculable IC$_{50}$ for each drug. Sorting is done by decreasing proportions of sensitive profiles.
Drug Modelling

Figure S80: Two drugs had significant features following feature reduction. Following stepwise model fitting, a single feature was selected for each of the drugs and both features were found to be significant at an $\alpha$ of 0.05.
Figure S81: Nine drugs had significant features following feature reduction. Following stepwise model fitting, two features were found to be significant at an $\alpha$ of 0.05.
Power Analysis

Activity Areas

Figure S82: The power of each feature to detect a given effect size for the drug A-443654.
Figure S83: The power of each feature to detect a given effect size for the drug Atopaxar.
Figure S84: The power of each feature to detect a given effect size for the drug BEZ-235.
Figure S85: The power of each feature to detect a given effect size for the drug BI78D3.
Figure S86: The power of each feature to detect a given effect size for the drug BNTX maleate.
Figure S87: The power of each feature to detect a given effect size for the drug Bosutinib.
Figure S88: The power of each feature to detect a given effect size for the drug BYL719.
Figure S89: The power of each feature to detect a given effect size for the drug Cisplatin.
Figure S90: The power of each feature to detect a given effect size for the drug ER27319-maleate.
Figure S91: The power of each feature to detect a given effect size for the drug Everolimus.
Figure S92: The power of each feature to detect a given effect size for the drug Foretinib.
Figure S93: The power of each feature to detect a given effect size for the drug GDC-0941 bismesylate.
Figure S94: The power of each feature to detect a given effect size for the drug JTC801.
Figure S95: The power of each feature to detect a given effect size for the drug MK-1775.
Figure S96: The power of each feature to detect a given effect size for the drug NSC95397.
Figure S97: The power of each feature to detect a given effect size for the drug NSC146109-hydrochloride.
Figure S98: The power of each feature to detect a given effect size for the drug OSI-906.
Figure S99: The power of each feature to detect a given effect size for the drug Ouabain.
Figure S100: The power of each feature to detect a given effect size for the drug PRIMA-1.
Figure S101: The power of each feature to detect a given effect size for the drug Pyrrolidine.
Figure S102: The power of each feature to detect a given effect size for the drug Rapamycin.
Figure S103: The power of each feature to detect a given effect size for the drug RDEA119.
Figure S104: The power of each feature to detect a given effect size for the drug Ro31-8220-mesylate.
Figure S105: The power of each feature to detect a given effect size for the drug Ryuvidine.
Figure S106: The power of each feature to detect a given effect size for the drug SCH79797.
Figure S107: The power of each feature to detect a given effect size for the drug SP600125.
Figure S108: The power of each feature to detect a given effect size for the drug TCS2312-dihydrochloride.
Figure S109: The power of each feature to detect a given effect size for the drug A-443654.
Figure S110: The power of each feature to detect a given effect size for the drug Atopaxar.
Figure S111: The power of each feature to detect a given effect size for the drug BEZ-235.
Figure S112: The power of each feature to detect a given effect size for the drug BI78D3.
Figure S113: The power of each feature to detect a given effect size for the drug BNTX maleate.
Figure S114: The power of each feature to detect a given effect size for the drug Bosutinib.
Figure S115: The power of each feature to detect a given effect size for the drug BYL719.
Figure S116: The power of each feature to detect a given effect size for the drug Cisplatin.
Figure S117: The power of each feature to detect a given effect size for the drug ER27319-maleate.
Figure S118: The power of each feature to detect a given effect size for the drug Everolimus.
Figure S119: The power of each feature to detect a given effect size for the drug Foretinib.
Figure S120: The power of each feature to detect a given effect size for the drug GDC-0941 bismesylate.
Figure S121: The power of each feature to detect a given effect size for the drug JTC801.
Figure S122: The power of each feature to detect a given effect size for the drug MK-1775.
Figure S123: The power of each feature to detect a given effect size for the drug NSC95397.
Figure S124: The power of each feature to detect a given effect size for the drug NSC146109-hydrochloride.
Figure S125: The power of each feature to detect a given effect size for the drug OSI-906.
Figure S126: The power of each feature to detect a given effect size for the drug Ouabain.
Figure S127: The power of each feature to detect a given effect size for the drug PRIMA-1.
Figure S128: The power of each feature to detect a given effect size for the drug Pyrrolidine.
Figure S129: The power of each feature to detect a given effect size for the drug Rapamycin.
Figure S130: The power of each feature to detect a given effect size for the drug RDEA119.
Figure S131: The power of each feature to detect a given effect size for the drug Ro31-8220-mesylate.
Figure S132: The power of each feature to detect a given effect size for the drug Ryuvidine.
Figure S133: The power of each feature to detect a given effect size for the drug SCH79797.
Figure S134: The power of each feature to detect a given effect size for the drug SP600125.
Figure S135: The power of each feature to detect a given effect size for the drug TCS2312-dihydrochloride.
### Table S1: Clinical Information for HNC Cell Lines.

<table>
<thead>
<tr>
<th>Cell Line</th>
<th>HPV</th>
<th>Site</th>
<th>Age</th>
<th>Sex</th>
</tr>
</thead>
<tbody>
<tr>
<td>93-VU-147T</td>
<td>16</td>
<td>Floor of mouth</td>
<td>58</td>
<td>M</td>
</tr>
<tr>
<td>HMS-001</td>
<td>16</td>
<td>Tonsil</td>
<td>-</td>
<td>M</td>
</tr>
<tr>
<td>UM-SCC47</td>
<td>16</td>
<td>Tongue</td>
<td>45</td>
<td>M</td>
</tr>
<tr>
<td>UPCI:SCC090</td>
<td>16</td>
<td>Tongue</td>
<td>44</td>
<td>M</td>
</tr>
<tr>
<td>UPCI:SCC152</td>
<td>16</td>
<td>Tongue</td>
<td>44</td>
<td>M</td>
</tr>
<tr>
<td>UPCI:SCC154</td>
<td>16</td>
<td>Tongue</td>
<td>54</td>
<td>M</td>
</tr>
<tr>
<td>BICR56</td>
<td>-</td>
<td>Tongue</td>
<td>-</td>
<td>F</td>
</tr>
<tr>
<td>Cal27</td>
<td>-</td>
<td>Tongue</td>
<td>56</td>
<td>M</td>
</tr>
<tr>
<td>Cal33</td>
<td>-</td>
<td>Tongue</td>
<td>69</td>
<td>M</td>
</tr>
<tr>
<td>Detroit562</td>
<td>-</td>
<td>Pharynx</td>
<td>F</td>
<td></td>
</tr>
<tr>
<td>FaDu</td>
<td>-</td>
<td>Pharynx</td>
<td>56</td>
<td>M</td>
</tr>
<tr>
<td>HSC-2</td>
<td>-</td>
<td>Oral Cavity</td>
<td>69</td>
<td>M</td>
</tr>
<tr>
<td>JHU006</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>JHU011</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>JHU029</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>PCI13</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>PCI30</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>PCI6A</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>PCI6B</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>PE/CA:PJ49</td>
<td>-</td>
<td>Tongue</td>
<td>57</td>
<td>M</td>
</tr>
<tr>
<td>RF15A</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>RF15B</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>RF22A</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>RF22B</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>RF37A</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>RF37B</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>SCC15</td>
<td>-</td>
<td>Tongue</td>
<td>55</td>
<td>M</td>
</tr>
<tr>
<td>SCC25</td>
<td>-</td>
<td>Tongue</td>
<td>70</td>
<td>M</td>
</tr>
<tr>
<td>SCC4</td>
<td>-</td>
<td>Tongue</td>
<td>55</td>
<td>M</td>
</tr>
<tr>
<td>SCC9</td>
<td>-</td>
<td>Tongue</td>
<td>25</td>
<td>M</td>
</tr>
<tr>
<td>UWO1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>UWO2</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>UWO8</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
</tbody>
</table>
Table S2: Summary of current legal and clinical statuses as well as disease of interest for each drug in drug panel. See Abbreviations page for abbreviations listed.
Bibliography


[20] Davit Bzhalava, Peng Guan, Silvia Franceschi, Joakim Dillner, and Gary Clifford. A


[33] Tina Wiest, Elisabeth Schwarz, Christel Enders, Christa Flechtenmacher, and Franz X.


[38] Julia Anne Woolgar and Asterios Triantafyllou. Pitfalls and procedures in the


[50] Maria J. Worsham, Kang Mei Chen, Nivedita Tiwari, Gerard Pals, Jan P. Schouten, Seema Sethi, and Michael S. Benninger. Fine-mapping loss of gene architecture at the CDKN2B (p15INK4b), CDKN2A (p14ARF, p16INK4a), and MTAP genes in head


[69] Jacques Bernier, Christian Domenge, Mahmut Ozsahin, Katarzyna Matuszewska,


[73] Everett E. Vokes, Kerstin Stenson, Fred R. Rosen, Merrill S. Kies, Alfred W.


[112] Jordi Barretina, Giordano Caponigro, Nicolas Stransky, Kavitha Venkatesan, Adam A. Margolin, Sungjoon Kim, Christopher J. Wilson, Joseph Lehár, Gregory V. Kryukov, Dmitriy Sonkin, Anupama Reddy, Manway Liu, Lauren Murray, Michael F. Berger,


[128] Anthony T. C. Chan, Y. M. Dennis Lo, Benny Zee, Lisa Y. S. Chan, Brigette B. Y. Ma, Sing-Fai Leung, Frankie Mo, Maria Lai, Stephen Ho, Dolly P. Huang, and Philip J.


[133] Carole Fakhry, Qiang Zhang, Phuc Felix Nguyen-Tan, David Rosenthal, Adel El-Naggar, Adam S. Garden, Denis Soulieres, Andy Trotti, Vilija Avizonis, John Andrew


[144] Peter Van Loo, Silje H. Nordgard, Ole Christian Lingjærde, Hege G. Russnes, Inga H. Rye, Wei Sun, Victor J. Weigman, Peter Marynen, Anders Zetterberg, Bjørn Naume,


[155] Christopher Wilks, Melissa S. Cline, Erich Weiler, Mark Diekhans, Brian Craft, Christy Martin, Daniel Murphy, Howdy Pierce, John Black, Donavan Nelson, Brian Litzinger,


[164] Simon A. Forbes, David Beare, Prasad Gunasekaran, Kenric Leung, Nidhi Bindal,


[185] Yuker Wang, Martin Moorhead, George Karlin-Neumann, Nicholas J. Wang, James Ireland, Steven Lin, Chunmuan Chen, Laura M. Heiser, Koei Chin, Laura Esserman,


[191] António E. Pinto, Teresa Pereira, Márcia Santos, Mariana Branco, Angelo Dias, Giovanni L. Silva, Mónica C. Ferreira, and Saudade André. DNA ploidy is an independent predictor of survival in breast invasive ductal carcinoma: A long-term


[235] Claudia Di Serio, Silvia Pellerito, Maria Duarte, Daniela Massi, Antonella Naldini, Giuseppe Cirino, Igor Prudovsky, Marco Santucci, Pierangelo Geppetti, Niccolò...


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


