Making sense of cancer mutations

Looking into the wilderness beyond genes

Rashid, M.M.

DOI
10.4233/uuid:5b7f0c06-9664-4fa1-a1b8-0e5e64b500bb

Publication date
2020

Document Version
Final published version

Citation (APA)
Rashid, M. M. (2020). *Making sense of cancer mutations: Looking into the wilderness beyond genes*. [Dissertation (TU Delft), Delft University of Technology]. https://doi.org/10.4233/uuid:5b7f0c06-9664-4fa1-a1b8-0e5e64b500bb

Making sense of cancer mutations

Looking into the wilderness beyond genes

Mamunur Rashid

Making sense of cancer mutations
Looking into the wilderness beyond genes

Proefschrift

ter verkrijging van de graad van doctor
aan de Technische Universiteit Delft,
op gezag van de Rector Magnificus Prof.dr.ir. T. H. J. J. van der Hagen
voorzitter van het College voor Promoties,
in het openbaar te verdedigen op
woensdag 9 September 2020 om 12:30 uur

door

Mamunur RASHID
M.Sc. in geavanceerde methoden in de informatica, Queen Mary Universiteit van Londen, UK
geboren te Noakhali, Bangladesh

Dit proefschrift is goedgekeurd door de promotor:

Samenstelling promotiecommissie bestaat uit:

Rector Magnificus	voorzitter
Prof. dr. M. J. T. Reinders	Technische Universiteit Delft, promotor
Dr. ir. J. de. Ridder	University Medical Center Utrecht, copromotor

Onafhankelijke leden:

Dr. M. P. J. K. Lolkema	Erasmus Medical Centre
Prof. dr. M. A. Swertz	University Medical Center Groningen
Prof. dr. T. Lenaerts	Université Libre de Bruxelles (ULB)
Dr. D. J. Adams	Wellcome Sanger Institute, UK
Prof. dr. L. F. A. Wessels	Technische Universiteit Delft
Prof. dr. R. C. H. J. van Ham	Technische Universiteit Delft, reserve member

Dr. D. Tax heeft als begeleider in belangrijke mate aan de totstandkoming van het proefschrift bijgedragen.

Printed by: Ridderprint | www.ridderprint.nl

Front & Back: Cover art : Search for cancer driving mutations in genes and beyond by Mamunur Rashid.

ISBN 978-94-6416-052-9

An electronic version of this dissertation is available at
http://repository.tudelft.nl/.

The infinite personality of human comprehend the universe.
There is nothing that cannot be subsumed by the human personality, and
this proves that the truth of the universe is human truth.

Rabindranath Tagore in a conversation with Albert Einstein

Contents

1

Introduction

1

1.1. Mutations in the Human Genome

Human DNA is a three billion base pair long sequence composed of only four nucleic acid alphabets Adenine (A), Cytosine (C), Guanine (G) and Thymine (T). A mutation is an alteration in and individuals persons DNA sequence, such that the sequence differs from what is observed in the population as a whole. Mutations are broadly classified into two major categories: hereditary and acquired (Figure 1.1) (Loewe, 2008; Milholland et al. 2017). Hereditary mutations, commonly known as germline mutations, are inherited from either of the parents through germ cells and are present throughout every cell of an individual. Acquired or somatic mutations, on the other hand, occur at various times points during a person's life and are their presence is mostly limited to certain types of cells. Many intrinsic factors can cause these changes such as erroneous DNA repair and external mutagenic insults such as ultraviolet radiation from the sun, smoking (Stratton et al. 2009; Konnick and Pritchard, 2016).

Mutations play a vital role in human evolution by enabling genetic diversity and protecting the population by enhancing disease resistance and survival (Lacy, 1997). Mutational changes that occur more frequently, for example, in more than 1% of the population are called polymorphisms or population variation and are responsible for many of the normal differences between people such as blood type, eye colour and hair colour(Karki et al. 2015; 1000 Genomes Project Consortium 2010). Mutations observed in less than 1% of the population are referred to as rare variants. Based on their impact on human health they are broadly categorized as advantageous to human health ('good'), harmful to health ('bad') and have little or no impact on health ('neutral') (Loewe and Hill, 2010; Landrum et al. 2014).

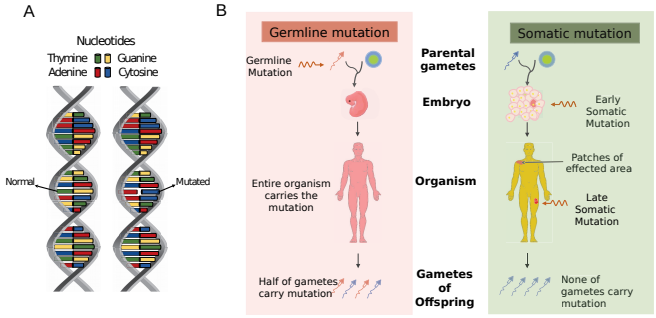


Figure 1.1: (A) Mutations are alterations of nucleic acids in human DNA. In the normal DNA the nucleotide base was thymine in the mutated DNA it changed to adenine (B) Germline mutations are inherited from either of the parents via germ cells while somatic mutations are acquired during a person's lifetime.

1.2. Somatic mutation in cancer

Cancer is commonly used to refer to more than one hundred distinct diseases, all displaying at least one of the phenotypic hallmarks suggested by Hanahan and Weinberg (Hanahan and Weinberg, 2011). Each cancer type has its own unique risk factors and epidemiology. One common factor that binds them all together is that they all arise from a changes in the DNA (Siegel et al. 2013). There are several models proposed as how a tumour arises and evolves from genetic changes. One reported by Stratton et al. suggests a single cell acquires the hallmarks of cancer through somatic alterations and clonally expand to form a tumour (Stratton et al. 2009; Hanahan and Weinberg, 2011). Somatic mutations can occur either in the protein-coding segment of the genome or in the part that does not actively transcribe and translate into functional protein, commonly known as the 'noncoding genome' (Nature Education, 2018). Mutations in these two different regions of the genome confer their influences in tumorigenesis via two distinct routes. Field cancerization is another popular model where a field of pre-malignant heterogeneous cell populations with their distinct mutational and expression profile can arise due to some epithelial histopathological alterations or mutagenic event (Dakubo et al., 2007, Parikh, K. et al. 2019). In chapter 5, we briefly discussed the effect of field cancerization in the context of adnexal tumours.

1.2.1. Coding and noncoding mutations in cancer

The protein coding genome is divided into functional sub-units called 'genes', which themselves are composed of one or more exons. Proteins are produced through transcription of exons into RNA and it's subsequent translation into amino acids are responsible for most of the work in a eukaryotic cell (Figure 2) (Cohen, 2004). As a result, mutations in the coding elements such as exons have an immediately quantifiable impact on protein production and human health. Mutations that contribute to tumour development or progression by increasing protein production in a gene are called activating mutations and the gene is referred to as an oncogene. Mutations that facilitate tumour development by repressing tumour suppressor proteins are called inactivating mutations (Vogelstein et al. 2013).

Historically, noncoding elements and their mutations were considered to have little or no influence in human health. In recent years, however, large-scale genome and epigenome profiling studies such as the Encyclopedia of DNA elements (ENCODE) and the Roadmap Epigenomics project have revolutionized our perspective of the noncoding genome. Conservative estimates suggest that as much as 40% of the human genome are directly or indirectly involved in some form of functional regulation (Encode Project, 2012; Roadmap Epigenomics Consortium et al. 2015). Noncoding cis-regulatory elements such as promoters and transcription factor binding sites (TFBS) regulate the transcription of nearby gene and recent studies by (Horn et al. 2013; Vinagre et.al. 2013; Larrayoz et al. 2016) have clearly demonstrated capabilities of mutations in these regions to drive tumorigenesis.

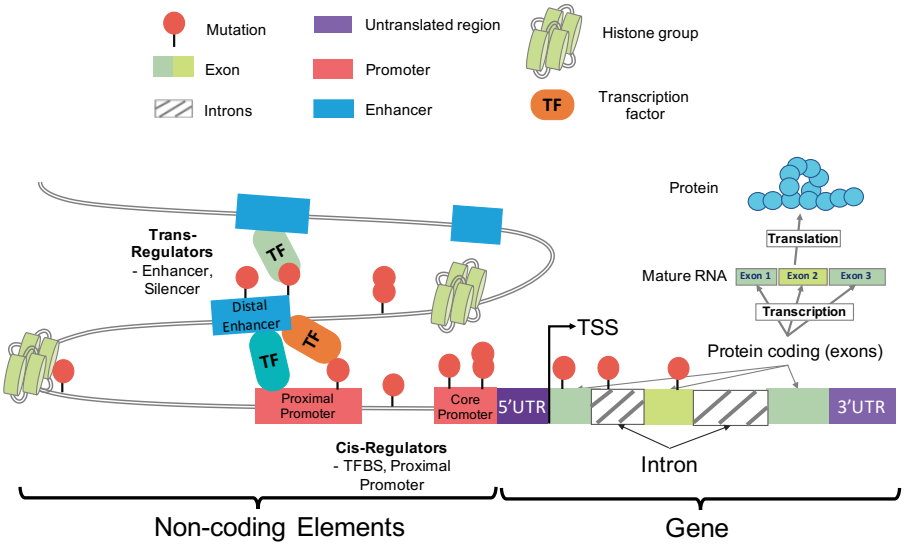


Figure 1.2: Protein coding and non-coding elements of the human genome. Exons are transcribed to messenger Ribonucleic acids (mRNAs) which are then translated to proteins. Noncoding elements such as introns, promoters and enhancer do not translate into proteins. These elements directly or indirectly regulate the transcription and translation process.

1.3. Next generation sequencing in cancer mutation detection

The first human genome sequencing took about 15 years and estimated cost was nearly three hundred million dollars (The Cost of Sequencing a Human Genome, NIH). It was performed using what was state-of-the-art technology at the time, the sequential Sanger genome sequencing technique. In contrast, modern massively parallel sequencing technologies can sequence millions of DNA fragments in parallel and produce around 45 human genomes in a single day for less than 1000 dollars each (Figure 1.3 a & b) (Illumina, 2015). These technologies are collectively known as Next Generation Sequencing (NGS). HiSeq Novasq from Illumina, SMRT sequencing from Pacific Bioscience (PacBio) are examples of a few popular available platforms. Figure 1.3 shows sequenced reads from a tumour and a normal DNA of an individual aligned against the human reference genome. Advancements of these technologies and rapid reduction in cost allows us to sequence at higher depth of coverage i.e. more reads per nucleotide and even detect mutations observed only in a small fraction of cells.

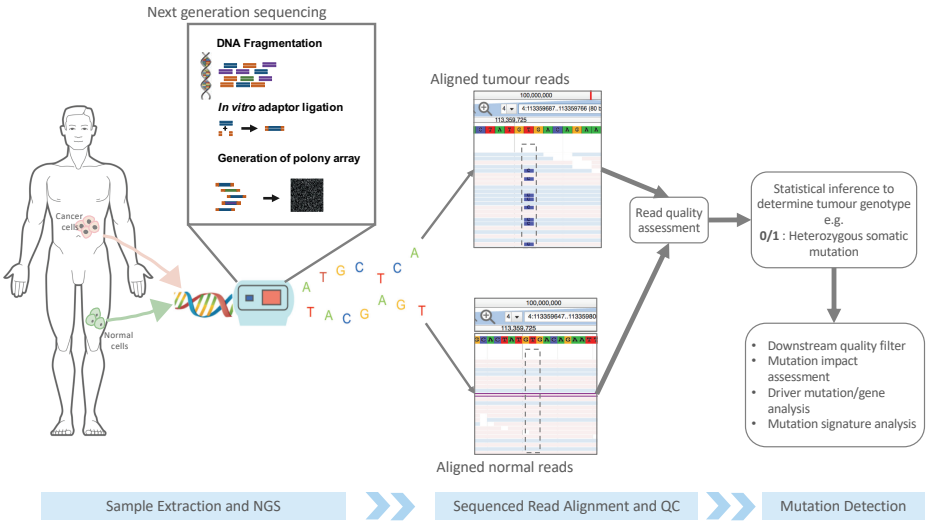


Figure 1.3: Somatic mutation detection using next generation sequencing technology : DNA samples are extracted from tumour and normal cells and fragmented for sequencing library preparation. Prepared library is sequenced in parallel using NGS instruments. Sequenced reads are then aligned to a reference genome. Reads marked as blue inside the dotted rectangle in tumour reads indicates the presence of a heterozygous mutation while the the reads from the normal DNA shows none. Detected mutations go through a number of post-calling quality control steps

1.4. Somatic mutation detection

Somatic mutations can provide selective survival advantage to cancerous cells allowing them to metastasize. Detection of these mutations will allow us to better characterize tumours, understand the mutational processes operative in them, study the perturbed biological pathways and to develop novel course of treatment (Stratton et al. 2009; Vogelstein et al. 2013). A rapid decline in sequencing cost has opened a new era of patient genomic data-driven personalized cancer treatment where cancer treatment is tailored to an individual patient based on the mutational landscape of patient tumours. The accurate detection of somatic mutation is therefore a key element of this process (Jackson and Chester, 2015).

1.4.1. Challenges of somatic mutation detection

In theory, somatic mutations can be distinguished by simply comparing the mutant read proportion between a tumour and a normal sample obtained from a cancer patient. For example, in Figure 1.3, a **T** to **C** nucleotide variation is observed in more than 50% of the tumour reads but the reads from the normal DNA remain invariant indicating the presence of a tumour specific somatic heterozygous mutation. In reality, however, this process becomes less trivial due to a number of

challenges mostly originating from two sources: (i) sample extraction/preparation and (ii) technical artefacts in DNA sequence (Alioto et al. 2015).

A DNA sample's journey from the tissue of origin to the sequencing machine involves several biochemical preservation and preparation steps. Sample cross-contamination can occur at any of these steps. Cross-individual and within-individual contamination are the most common types and one of the largest sources of artefacts in somatic mutation detection (Cibulskis et al. 2013). Cross-individual contamination is when DNA molecules from a different individual get into the admixture. Even a small level of contamination can introduce a large number of low allelic fraction false positives. Within-individual contamination, on the other hand, occurs when tumour DNA contaminates the normal or vice versa. DNA material from normal tissue adjacent to the tumour have been routinely used as a source of germline DNA in many retrospective cancer profiling studies (Emami et al. 2017; McLendon et al. 2008). These tissues are often infiltrated by tumour cells and can lead to a severe loss in sensitivity during somatic mutation detection.

Formalin induced artefacts are another major source of false positive in many cancer studies. Formalin fixed paraffin-embedding (FFPE) is a century old technique for tissue preservation and is one of the primary sources of cancer samples in many retrospective cancer profiling studies. Hydrolytic deamination, the transformation of a cytosine base to uracil/thymine (C>T), is a frequently occurring DNA damage observed in FFPE tissues. Following polymerase chain reaction (PCR) to amplify DNA material, these errors appear as low allelic fraction mutations in NGS data (Oh et al. 2015; Do and Dobrovic, 2012). Cancer tissues often contain multiple sub-clones and these mutations also appear at a very low frequency in sequencing data (Yates et al. 2015). Distinguishing the true low-frequency somatic mutations from FFPE induced sequencing artefacts remains a big obstacle when studying FFPE tissues.

Despite considerable improvements in DNA sequencing technologies over the last decade sequencing error is still one of the biggest rate-limiting factors in distinguishing true somatic mutations. Comprehensive analysis of tumour-normal pairs from chronic lymphocytic leukaemia and medulloblastoma by Alioto et al. (2015) demonstrated that issues such as low sequencing depth, imbalance in depth of coverage between tumour and normal sample, poor read quality, low read mapping quality complicates things further. In addition, our analysis of several cancer data sets (e.g. Rashid et al. 2013; Rashid et al. 2016; Rabbie et al. 2017) presented through chapters 2-5 revealed that misalignment of sequencing reads by alignment tools around repetitive regions of the genome and structural variants can also give rise to a considerable amount of artefacts. In the next couple of sections we will discuss some popular somatic mutation detection tools and possible avenues to improve accuracy in somatic mutation detection.

1.4.2. Somatic mutation detection tools

Several tools have been developed in recent year to detect somatic mutations from paired tumour-normal sequencing data and they broadly belong to two classes. The first group perform an independent analysis of tumour and normal sequencing reads followed by a statistical test to confirm if the tumour has a different genotype than the normal (Pleasance et al. 2009; Koboldt et al. 2012). The second group of methods such as (Larson et al. 2012; Goya et al. 2010; Cibulskis et al. 2013) take the somatic mutation rate into account and use joint probability-based statistical approaches to simultaneously analyse matched tumour and normal data.

The agreement between these tools is often considerably low, mostly due to the differences in their core algorithms (Kim and Speed, 2013; O'Rawe et al. 2013). Each tool has a slightly different error model and prior assumptions of the underlying somatic mutation rate to tune sensitivity and specificity (Xu et al. 2014). For example, SomaticSniper (Larson et al. 2012) calculates tumour and normal genotype likelihood (Li et al. 2008) for each site using a uniform prior for somatic mutation rate (default : 0.01) and reports the phred-scaled probability of them being different as somatic score. MuTect (Cibulskis et al. 2013), on the other hand, uses different prior probabilities at sites of common germline variation versus the rest of the genome. Finally, VarScan2 (Koboldt et al. 2012), which belongs to the first group, performs a Fisher's exact test to assess if the tumour and normal genotypes are significantly different. So, not surprisingly applying different variant-calling algorithms to the same data often result in a partially overlapping set of somatic mutations (Chapter 2 and Rashid et al. 2013). Table 1.1 below gives a quick overview of some of the most popular somatic mutation detection tools available.

Somatic caller	Method used	SNV	Indel	Comment
Mutect2	Bayesian classifier	✓	✓	Postprocess filter included; Low allelic fraction mutations
Varscan2	Allele frequency based heuristics, Fisher's exact test	✓	✓	Option to perform copy number analysis
JointSNVMix2	Probabilistic graphical models	✓	✗	No postprocess filter included
CaVEMan	Expectation maximization	✓	✗	Postprocess filter available
SomaticSniper	Genotype likelihood model	✓	✗	Standard VCF output format. No post process routine included
Bambino	Allele frequency based heuristics	✓	✗	Simplistic allele frequency based interpretation
Strelka	Bayesian statistics	✓	✓	Postprocess filter included

Table 1.1: Overview of popular somatic mutation detection tools: the table highlights the methods used by these tools, their output e.g. single nucleotide variants or larger variants and whether they include any inbuilt post mutation detection quality control filters.

1.4.3. Combination of multiple tools and filtering strategy

Due to the heterogeneity in their outputs, selecting the ideal somatic mutation detection tool appropriate for the task in hand can be challenging (Xu et al. 2014). For example, in clinical cancer diagnostics settings, false discoveries can lead to misleading prognosis and prescribing an incorrect course of treatment. Biomarker research groups, however, can settle for low specificity in order to identify novel target genes. In chapter 2, we propose a combinatorial approach to harness the strengths of multiple somatic mutation detection tools to mitigate the variability issue. This allows end users to adjust sensitivity and specificity based on the research question in hand (Rashid et al. 2013). In addition, we developed a set of post-mutation detection quality control measurements to address many issues that give rise to sequencing artefacts. Our analysis on a set of published human breast cancer samples (Nik-Zainal et al. 2012) and hepatocellular carcinomas (Guichard et al. 2012) presented in chapter 2 indicates that this framework considerably improves the sensitivity and specificity of the somatic mutation detection process. This analysis framework was also applied to several large-scale cancer genome sequencing studies. These include analysis of 55 colorectal adenoma tumours (presented in chapter 3 and (Rashid et al. 2016), the tumour genome of a melanoma patient (presented in chapter 4 and (Rabbie et al. 2017), tumour exomes of 24 mice representing pre- and post-haematopoietic malignancy (presented in Horton et al. 2017).

1.4.4. Orthogonal mutation validation

Even in the most stringent settings, somatic mutation detection frameworks can produce false calls (Alioto et al. 2015). As mentioned in the previous section, this can have a significant impact in clinical diagnostics setups. To mitigate these uncontrollable factors, it is essential to validate detected somatic mutations orthogonally. Orthogonal validation refers to verification of mutations using a different technology (e.g. Sanger sequencing) other than the platform on which the mutations were originally detected. An orthogonal validation of a handful randomly selected set of detected somatic mutations using a new aliquot of DNA can consolidate the findings and provide useful insight about the false discovery rate of the system (Beck et al. 2016). We used a number of orthogonal validation techniques throughout this thesis (chapter 2-5 and Rashid et al. 2013; Rashid et al. 2016; Rabbie et al. 2017) to validate reported somatic mutations. A further discussion on a selection of these technologies, their strengths and limitations can be found in chapter 7.

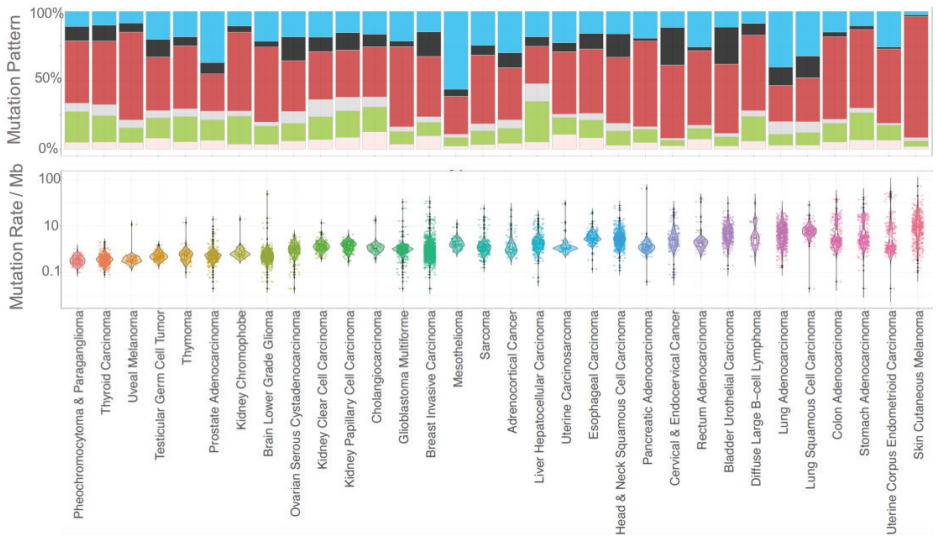


Figure 1.4: Landscape of somatic mutation across various cancer types (a) shows the mutation pattern (nucleotide base change) across different cancer types while (b) shows the distribution of somatic mutation burden

1.5. Somatic mutation burden and signatures

The declining cost of sequencing has enabled large-scale genome profiling studies such as the International Cancer Genome Consortium (ICGC) (Zhang et al. 2011) and the Cancer Genome Atlas (TCGA) (Weinstein et al. 2013) to sequence unprecedented numbers of cancer genomes across many different cancer types. These colossal data sets have revealed a number of remarkable properties of various cancer sub-types including their mutation burden, mutation patterns and potential cancer-driving genes. Figure 1.4 shows the somatic mutation load and nucleotide change spectra for 32 distinct cancer types. Skin cutaneous melanomas exhibit the highest mutation load (median: 16.60 mutations/Mb) occurring most likely due to ultra-violet ray damage to melanocytes. Alternatively, pheochromocytomas, a benign tumour of adrenal glands, have the lowest mutation burden (median: 0.35 mutations/Mb). The cytosine to thymine (C>T) transition is the most common type of single nucleotide change across most of the cancer types. Results from recent clinical trials in melanoma by (Lauss et al. 2017), in multiple myeloma by (Miller et al. 2017) and several other cancer types have indicated that mutational load has a strong correlation with the expression of neoantigens that allows immune checkpoint inhibitors to better identify cancer cells and improves disease free survival. In an effort to discover any such clinical phenotypes associated with mutation burden for early stage human malignancies such as colorectal adenomas (discussed in chapter 3 & Rashid et al. 2016), adnexal tumours (discussed in chapter 5) we

compared their mutational load with that of several cancers published by the TCGA consortium. In section 1.6, we also discussed the significance of mutational burden in the context of paediatric melanoma.

Mutations in tumour cells are the consequence of aberrant endogenous processes such as defective DNA repair or due to exogenous factors such as exposure to carcinogens. The imprint of a mutational process on tumour DNA sequence is commonly referred to as a mutational signature (Alexandrov et al. 2013). For example, excessive exposure to ultra violet light dramatically increases the number of cytosine to thymine (C>T) mutations, a common signature observed in many melanoma patients. Analysis of mutational signatures allows us to better understand underlying biological processes associated with a number of cancers and has also allowed patient stratification for therapy (Nik-Zainal et al. 2012). Mutational signature detection approaches available in current literature broadly falls under two categories: *de-novo* signature detection vs reconstruction of samples based on published signatures. De-novo signature analysis tools such as EMu (Fischer et al. 2013), SomaticSignatures (Gehring et al. 2015) delineate the operative mutational processes without any prior knowledge of cancer type or known mutational signatures. Considerably large sample cohorts are required for reliable estimation of de-novo signatures. On the other hand, the second class of methods e.g. deconstructSigs by Rosenthal et al. (2016) estimates the contribution of known mutational signatures in each individual tumour. Using a *de-novo* signature analysis approach, we identified two distinct mutations processes operative in early-stage colorectal adenoma tumours (chapter 3 and Rashid et al. 2016). In chapter 5, we followed the second approach to identify the contribution of signatures published by Alexandrov et al. 2013 in different adnexal tumour subgroups. Finally, as reported in Horton et al. 2017, using a custom analytical approach, we compared published human cancer signatures (Alexandrov et al. 2013) with de-novo signatures identified in 24 mice tumours that developed haematopoietic malignancy to assess the efficacy of this mouse model to study human disease.

1.6. Significance of mutation burden in paediatric melanomas

Childhood cancers are rare and mostly comprise haematopoietic tumours (about 40%), various solid tumours (about 35%) and central nervous system (CNS) tumours (about 25%). Compared to the commonly occurring adult tumours, paediatric tumours differ in their underlying pathology and behaviour and are hence treated differently (Murphy et al. 2013). For example, immunotherapies have shown great potential in treating adult melanoma patients with higher expression of neoantigens a feature directly correlated with a higher mutational burden. In the clinic, however, paediatric patients are not routinely considered for these therapies (Rabbie et al. 2017). In chapter 4 we described the clinical course of a 15 year old primary melanoma patient treated with conventional treatment. We presented the complete genomic profile of her tumour and compared this to a further series of

13 adolescent melanomas published by (Lu et al. 2015) and 275 adult cutaneous melanomas from the TCGA consortium (Zhang et al. 2011). Based on our findings in chapter 4, we suggested that paediatric melanomas can have a mutational load as high as adult cutaneous melanomas and the genomic profile of paediatric melanoma patients should be taken into account when determining the course of treatment (Rabbie et al. 2017).

1.7. Tumour heterogeneity and field cancerization

Cancer is an evolving disease that originates from a single mutated cell and during its course of progression, tumours generally become more heterogeneous. This leads to the presence of a diverse collection of cell populations also known as subclones within the bulk tumour, harbouring distinct mutational patterns and often different levels of sensitivity to treatment (Dagogo-Jack and Shaw, 2018, Yates et al., 2015). Understanding these diverse cell compositions will give us a better insight into tumour evolution and potential therapeutic intervention. While malignant tumours remain the focus of the majority of cancer studies some scientists such as (Marino-Enriquez and Fletcher, 2014) argue that more emphasis should be given on benign tumours. Many benign tumours transform into malignant tumours (e.g. colon polyp to adenocarcinoma, skin mole to cutaneous melanoma) and a comprehensive characterization of these transformations in an early stage will lead to early cancer detection and improved prognosis (Atkin and Saunders 2002; Tsao et al. 2003). Sequencing multiple tumour regions, longitudinal analysis, liquid biopsy samples and single-cell sequencing are a few emerging techniques to better understand a tumour's journey from benign stage to complex heterogeneous malignancy (Dagogo-Jack and Shaw, 2018). Recent cancer single cell sequencing efforts such as colon epithelial cell sequencing by Parikh et al. (2019) and Topographic Single Cell Sequencing (TSCS) based breast tumors profiling by Casasent et al. have also significantly improved our understanding of these early tumour transformation (Lawson et al., 2019). In chapter 3, we study this heterogeneity in early stage colon adenomas by examining multiple polyps from individual patients and detected significant differences in the somatic mutation rate as well as driver genes between the tumours from the same individual (Rashid et al. 2016). To explore the journey of a benign tumour to malignancy we also analyzed distinct components of several skin adnexal tumours (presented in chapter 5). Based on our findings, we argued that malignant skin adnexal tumours do not necessarily arise from their benign counterparts and can originate from independent lineage.

'Field cancerization' is an alternative cancer development model first proposed by Slaghter et.al. (1953) after observing multi-centric tumour origin in oral carcinoma patients. According to this process, instead of arising from one single cell and evolving to multiple subclones, there exists a field of pre-malignant cancer cells due to some epithelial histopathological alterations or mutagenic event from which multiple independent lesions occur, leading to the development of multi-focal tumours (Dakubo et.al., 2007). With the advancements in molecular profiling of tumour genome, other works have documented its presence in different cancer types such

as Brodsky Jones (2004) in haematopoietic malignancies, Heaphy et al. (2006) in breast carcinoma and Shen et al. (2005) in colorectal cancer. Field cancerization has significant clinical implications in cancer treatment. Cancer fields often remain after surgical resection of the primary tumour leading to new cancer development. Validated biomarkers from cancer fields can also be useful in risk assessment, early detection and chemo-prevention (Dakubo et al. 2007). In our study of human adnexal tumours presented in chapter 5, we identified driver mutations in both the tumour and normal tissue collected from the vicinity of several tumours indicating the possible presence of cancer fields in these tumour types.

1.8. Driver and passenger mutations

Only a handful of somatic mutations among the thousands observed in a tumour genome confer a selective survival advantage to the tumour cells. These mutations are commonly referred to as driver mutations and often occur at a very early stage of tumour development, triggering the tumorigenesis (Gonzalez-Perez et al. 2013; Vogelstein et al. 2013). Knudson (1971) proposed a 'two hit' hypothesis of cancer development in the 1970s after studying retinoblastoma tumours. According to this hypothesis, in dominant inherited form, one mutation is inherited from germ cells (e.g. BRCA1/2 in familial breast and ovarian cancer) and the second mutation is acquired by somatic cells (Miki et al. 1994). In the nonhereditary form, however, both mutations occur in the somatic cell. Because of their role in tumour initiation and providing selective growth advantage, driver mutations are seen as the 'Achilles' heel' of tumours. They are the primary objective of many cancer research programs because of the potential to tailor therapeutic interventions based on the patient's own tumour DNA sequence.

Unlike driver mutations, a large fraction of mutations in tumour genomes do not confer any selective growth advantage and are categorized as 'passenger mutations' (Vogelstein et al. 2013). As a result, these mutations have never been the topic of active research in cancer genomics. As discussed in section 1.6, some recent clinical data indicated an association between passenger mutation burden and response to checkpoint inhibitors, mostly due to an increase in neo-antigen load (Lauss et al. 2017; McFarland et al. 2017). Our own analysis on a set of UV treated mouse melanoma cell lines has also shown that an increase in ultra violet exposure associated mutations, which in turn manifest in higher neoantigens load, enhance response to checkpoint blockade treatment (Lo AJ and Rashid M et.al. : Submitted to Science transnational medicine). However, any causal link between any passenger mutation and cancer has yet to be established.

1.8.1. Driver mutations in the coding genome

Driver mutations in the coding region are broadly classified into two categories, oncogenic and tumour suppressor. Oncogenic driver mutations mostly occur in specific codons – missense or focal amplification - causing increased protein production. *BRAF*, *KRAS*, *APC1* are examples of oncogenes operative across number

of cancer types. Tumour suppressor genes such as *TP53* and *RB1*, on the other hand, manifest through the loss of function or deleterious mutations (Vogelstein et al. 2013).

Because of their directly measurable impact in cancer progression and potential therapeutic opportunity, identifying driver mutations in the coding genome remains one of the fundamental focuses of many cancer genomics studies. A plethora of tools have been developed over last decade to deconvolute the complex genomic signal and identify a handful of driver mutations from a large pool of passenger ones (Gonzalez-Perez et al. 2013; Dees et al. 2012). By and large, these algorithms search for the enrichment of protein-altering mutations within a gene body given the background mutation rate of that particular gene (Porta-Pardo et al. 2017). Gene-specific characteristics, such as length, replication timing are also taken into consideration when assessing the propensity for acquiring mutations (Lawrence et al.2013). IntOGen, developed by Gonzalez-Perez et al. (2013), on the other hand, combines several deleterious metrics (e.g. SIFT (Adzhubei et al. 2013) and PolyPhen2 (Ng and Henikoff, 2003)) to calculate functional impact bias (FM bias) of mutations in genes against a background null distribution. Another method *dNdScv*, published by Martincorena et al. (2017) computes the ratio of non-synonymous to synonymous mutation ratio (dN/dS) per gene, to infer positive selections. An application of both these methods can be found in chapter 5. In chapter 3, we applied an in-house driver gene detection method similar to *dN/dS* on a set of colorectal tumours to establish the significance of accumulation of loss of functions mutation in the *WTX* gene (Rashid et al. 2016). In chapter 5 we extended the analysis process further by applying an application of both dNdScv and IntOGen

The term 'driver mutation' is often associated with somatic mutations but germline mutations also play a critical role in driving cancer development. Unlike somatic mutations that trigger the tumorigenesis, these mutations predispose individuals to cancer risk. Germline mutations in *BRAC1* and *BRAC2* have been associated to a number of cancers including breast (Peto J. et al. 1999) and ovarian cancer (Kanchi et al. 2014). Individuals carrying a germline mutant allele of *POT1* gene have a higher chance of developing of cutaneous melanoma (Robles-Espinoza et al. 2014). Identifying these germline risk alleles can lead to early prevention and better patient management. Distinguishing these mutations poses a fundamentally different challenge than that of somatic driver somatic mutations discussed above. In chapter 5, we reported a custom workflow to assess germline risk alleles of cutaneous adnexal tumour patients and reaffirmed the role of *CYLD* as a germline driver in these tumours.

1.8.2. Driver mutations in the noncoding genome

Until very recently driver mutations have been exclusively associated with coding genes because of their ability to alter protein production. Recent large-scale cancer-genome sequencing efforts, such as TCGA and ICGC, have revealed that the vast majority of somatic variation occurs in the 98% of the genome that is considered

to be noncoding, i.e. outside of gene bodies (Zhang et al. 2011; Chang et al. 2013). Mutations in the promoter regions can however lead to the creation of new transcription recruitment sites (Horn et al. 2013) or the reduction (Cooper et al. 2002) of Transcription Factor (TF) binding affinity (Katainen et al. 2015). Work by Lopes-Ramos et al. (2017) & Bhattacharya and Cui (2016) have shown evidence of aberrant gene expression as a consequence of mutations in microRNA binding sites. Pan-cancer analysis by Weinhold et al. (2014) has shown a strong enrichment of mutations in the regulatory regions of several cancer-driving genes. These findings unambiguously highlighted the importance of noncoding mutations as potential cancer drivers. A well-characterized set of noncoding drivers can open new diagnostic and therapeutic avenues for many cancers. In the next section, we discuss a few existing tools to prioritize noncoding drivers, explore some of the challenges and also lay a foundation for our own efforts to prioritize them (chapter 6).

1.9. Prioritization of noncoding mutations

Noncoding driver mutations are thought to exert their influence on tumour growth via regulatory elements and as a result gene-centric enrichment tests to identify protein altering hotspots are no longer effective in the noncoding genome. The research community is still in the early days of cancer whole genome sequencing and the lack of sufficient validated noncoding drivers makes the task of establishing any common pattern very challenging. A wide range of computational approaches have been developed to distinguish noncoding driver mutations from benign passenger ones. These tools leverage the wealth of large-scale cancer genome profiling studies such as ICGC (Hudson et al. 2010) and comprehensive epigenome profiling studies such as ENCODE (Encode Project, 2012) to provide a rich characterization of the mutations in the noncoding genome. We will discuss some of these tools in the subsequent section

Machine learning based approaches have already been successfully adopted to solve a wide range of biological data analysis problems from protein structure prediction Rost and Sander (1994) and classification Weston et al. (2005) to biomarker discovery in cancer Perou et al. (1999). Unlike rule-based techniques, which inherently rely on user-defined feature weights (e.g. Fu et al. 2014), machine learning based techniques learn the underlying distribution of the data in an unbiased manner. In the context of noncoding mutation prioritization, assigning absolute weight on molecular features such as transcription factor binding activity or DNA accessibility is quite impractical due to the absence of a precise characterization regarding how they operate in cancer cells. This makes machine learning based systems a more favourable choice in mutation prioritization tasks.

Machine learning algorithms can be divided into two main types: unsupervised or supervised learning. Unsupervised methods partitions the data into meaningful clusters without any explicit data label (e.g. breast gene expression pattern by Perou et al. 1999). Clustering (e.g. k-means, hierarchical), dimensionality reduc-

tion methods such as Principle Component Analysis (PCA) are examples of unsupervised learning. Supervised learning approaches, on the other hand, learn from a set of labelled observations (e.g. protein classification from amino acid sequence Weston et al. 2005). Classifiers such as support vector machine and decision trees are examples of supervised learning. The performance of supervised learning systems rely heavily on good training examples and the absence of sufficient validated noncoding drivers makes unsupervised learning an appealing alternative for noncoding driver prediction. However, because of several factors such as feature scaling and ambiguity around the interpretation of identified clusters of mutations, supervised learning methods have been dominating the prioritization landscape (Kircher et al. 2014; Ritchie et al. 2014). We will briefly discuss some of these methods in the following section.

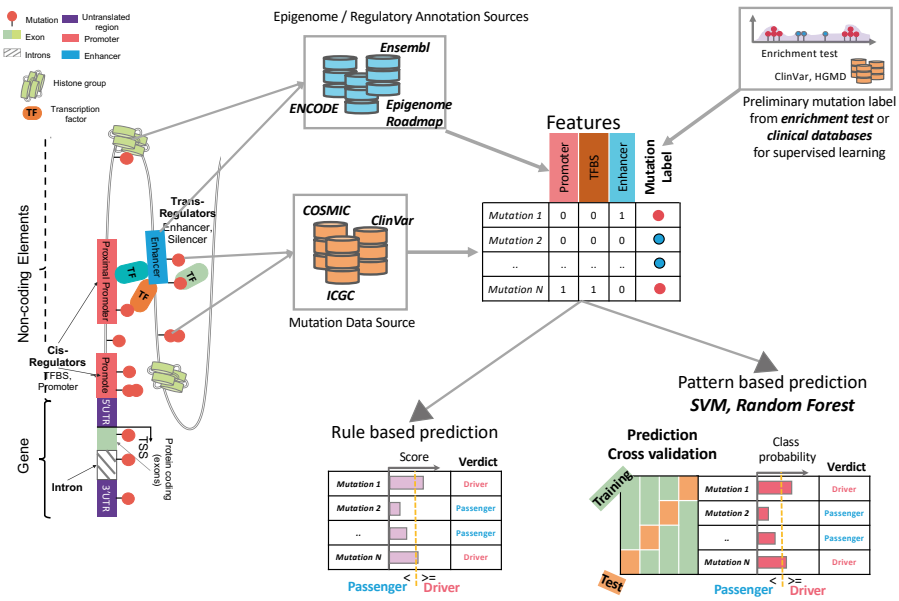


Figure 1.5: Overview of noncoding driver prioritization: Mutations are annotated with genomic, epigenomic and regulatory features. Mutations are then scored for pathogenicity either using a predetermined feature weights or machine learning techniques via cross validation

1.9.1. Noncoding driver prediction tools

Numerous tools have been developed in recent years to predict noncoding driving mutations. They can be broadly classified in to two groups: machine learning based approaces such as Combined Annotation Dependent Depletion (CADD) by Kircher et

al. (2014), GWAVA (Ritchie et al.), FATHMM (Shihab et al. 2015), DeepSea by Zhou et al. (2015) and rule-based such as Funseq2 (Fu et al. 2014) and SuRFing Ryan et al. (2014). These workflows have exploited a compendium of genomic, epigenomic and regulatory information to annotate noncoding mutations collected from various data sources (e.g. the Human Gene Mutation Database(HGMD), ClinVar, ICGC). Mutations are annotated for a range of features such as their overlap with a known regulatory region (e.g. promoter, enhancer), conservation of the nucleotide base across various species or it affects on TF binding affinity. Rule-based approaches such as FunSeq2 score every mutation based on pre-determined feature weights. Supervised approaches, however, require labelled data. Mutations are labelled either pathogenic or passenger based on experimentally validated clinical associations (e.g. ClinVar or HGMD) or some heuristics. For example, GWAVA and FATHMM use a set of curated heritable germ-line mutations from the HGMD database as positive instances and benign polymorphic variations (SNPs) as negative. CADD, on the other hand, trained it's SVM model on a set of 29.4 million simulated mutations and observed SNPs in the human genome. Finally, the pathogenicty of every single mutation is assessed via a cross-validation, dividing the data into multiple training and test folds. Figure 1.5 illustrates a generic noncoding mutation prioritization workflow adopted by almost every single methods described above. A brief description of these methods, the underlying algorithm, and the data used for training and testing can be seen in table 1.2.

Tool	Approach	Data Sets used	Trained on	Validation
GAWAVA	**Random Forest**	1000 Genome COSMIC; HGMD	HGMD (P) 1000 Genome SNPs(N)	HGMD (P) COSMIC Recurrent (P)
CADD	**SVM**	Whole Genome ClinVar; TRP3	SNPs with high derived allele frequency and simulated mutations	ClinVar; TP53
DANN	**Deep Learning**	Whole Genome ClinVar;	Data from CADD	
FATHMM	**Multiple Kernel Learning**	HGMD	HGMD (P) 1000 Genome SNPs(N)	HGMD
Kyoon Lab	**Random Forest**	ICGC Breast and Lung Cancer Mutations	ICGC Breast and Lung Cancer Mutations	
SuRFing	Combined Weighted Rank	HGMD(P) ENCODE(Background) RAVEN; ClinVar	HGMD(P) ENCODE(Background) RAVEN; ClinVAR	SORT1 EGR2 TCFL2
FunSeq2	Rule Based	Cosmic Recurrent Mutations ICGC (570 Tumours)	NA	HGMD TERT

Table 1.2: Overview of noncoding driver prioritization tools : table lists machine approaches (e.g. rule based or pattern based) used by these tools and data set used for training, test and validation

The studies discussed above have laid the initial foundation for noncoding mutation prioritization and demonstrated that properties of driver mutations can indeed be learned. In their effort to better understand the properties of noncoding driver mutations they collectively gathered a large compendium of curated annotation sources allowing subsequent research projects to investigate the properties of noncoding mutations in a data-driven approach. Yet there remains scope for

considerable improvements in several areas.

As mentioned previously, a well defined positive (i.e. driver) and negative (i.e. passenger) set of mutations is essential for a supervised learning system to predict reliably. A number of computational approaches (e.g. FATHMM, GWAVA, SuRFing) described above have used generic pathogenic variants reported in databases such as HGMD or ClinVar instead of cancer specific mutations. Cancer causing mutations are fundamentally different from other disease associated mutations because of their ability to introduce one of the cancer hallmarks in affected cells (Hanahan and Weinberg, 2011). Moreover, many noncoding cancer mutations reported in these databases are of germline origin. In section 1.8.1 we have already discussed how somatic and germline mutation confers their influence in cancer development through two unique routes and we argue therefore that their detection also requires distinct approaches. To train a model for somatic noncoding driver mutation prediction, only somatic mutations should be taken into account. Weinhold et al. (2014) proposed a window based pan-cancer somatic mutation burden analysis to identify mutational hotspots across the tumour genome. These mutational hotspots are indicative of genomic regions under positive selection in cancer genomes. In the absence of a large set of experimentally validated noncoding driver mutations, these approaches provide a good approximation of a true driver set. In chapter 6, we adopted a similar approach to label mutations for the downstream classification task. Detection of known noncoding driver mutations such as *TERT* promoter mutations consolidated our argument to use this approach to generate data label for supervised learning.

Several tools discussed previously (e.g. Kircher et al. 2014; Quang et al. 2015) exploited a mixture of coding and noncoding features and the features set are dominated by protein coding features (e.g. consequence, PolyPhen) many of which are not relevant to noncoding mutations. As a result, they often make excellent predictions for mutations in the protein coding regions but perform poorly in prioritizing noncoding mutations.

Due to the scarcity of driver mutations and an abundance of passenger mutations, any prioritization tool aiming to distinguish between these two classes faces a serious class imbalance challenge (Longadge et al. 2013). Oversampling of the minority class (e.g. SMOTE by Chawla et al. 2002) and undersampling of the majority class (e.g. Tomek link by Tomek, 1976) and have previously been shown to offer some improvements in type 2 diabetes prediction (Ramezankhani et al. 2016). In chapter 6, we explored several avenues to mitigate the class imbalance problem in the context of noncoding driver mutation prediction and proposed a class dependent loss function to address this issue.

1

1.10. Contribution of this thesis

Cancer is a multifaceted disease and understanding the complex interplay between protein coding genome and noncoding genetic elements is the key to this battle. In this thesis we explored various computational methods for somatic mutation detection and distinguishing drivers from passenger ones within and beyond the coding genome. The schematic diagram below presents a simple layout of various inter-related topics discussed in this thesis.

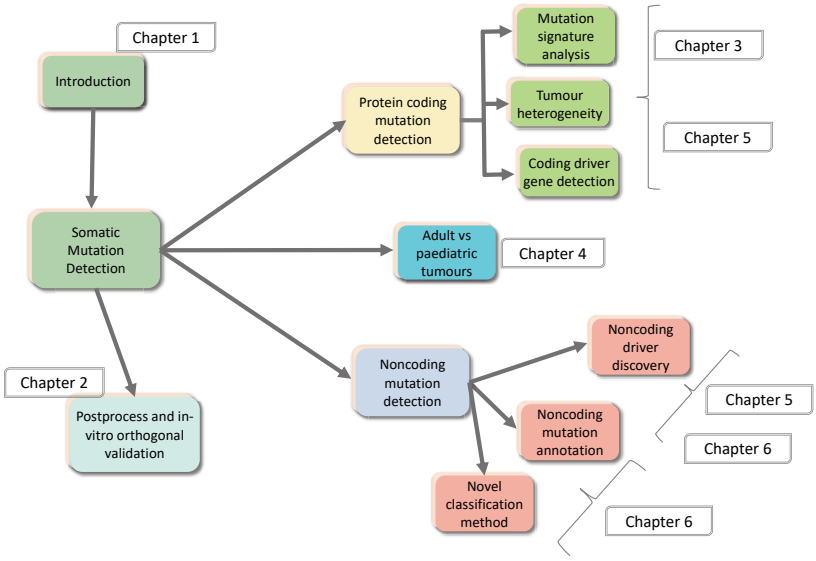


Table 1.3: A schematic diagram of different topics discussed in this thesis

We proposed a novel computational framework for the accurate detection of somatic mutations and demonstrated its application in a number of cancer studies. We also investigated several well-known sources of artefacts that frequently contaminates the mutation detection process and outlined approaches to tackle them combining in-silico and experimental procedures. We followed very early stage tumours to understand their journey from benign skin abnormality to malignancy and heterogeneity in their genetic composition. Using the help of unsupervised machine learning approaches we examined the mutational signatures that distinguish early stage colon adenomas. Combining published and in-house driver gene discovery methods we reported novel cancer-driving genes in human colorectal and adnexal tumours. In conjunction with in-silico analysis, we deployed a number of in-vitro experiments to confirm these findings. Our comparative analysis of paediatric and adult melanoma patients indicated the necessity of incorporating genomic data in paediatric melanoma patient management in the clinic. Finally, we explore beyond the traditional boundaries of coding genome and proposed a novel work-flow to

annotate and prioritize noncoding cancer-driving mutations. We aimed to address several computational challenges associated with this task such as the lack of training data and the class imbalance problem. We strongly believe, taken together these findings will provide useful guidelines for future tumour genome analysis and therapeutic target discovery studies at a genomic scale.

2

Cake

We have developed Cake, a bioinformatics software pipeline that integrates four publicly available somatic variant-calling algorithms to identify single nucleotide variants with higher sensitivity and accuracy than any one algorithm alone. Cake can be configured to run on a high-performance computer cluster or used as a standalone application.

2.1. Introduction

The development of Next Generation Sequencing (NGS) technologies has made it possible to generate more comprehensive catalogs of somatic alterations in cancer genomes than ever before. Software tools to find these variants deploy different mathematical approaches to interrogate the genome sequences of tumour / germline pairs. For example, the variant detectors Bambino (Edmonson, et al., 2011) and VarScan 2 (Koboldt, et al., 2012) both identify somatic variants by comparing alternative allele frequencies between tumour and normal sequences. VarScan 2 uses a Fisher's exact test and Bambino a Bayesian scoring model to identify somatic variants in paired samples. Other algorithms include CaVEMan (Nik-Zainal, et al., 2012; Stephens, et al., 2012) and SAMtools mpileup (Li, et al., 2009), which compute the genotype likelihood of nucleotide positions in tumour and normal genome sequences by use of an expectation-maximization method.

Putative, raw variant calls made by these algorithms typically undergo further filtering. For example, known single nucleotide polymorphisms (SNPs) present in dbSNP (Sherry, et al., 2001) or the 1000 Genomes dataset (The 1000 Genomes Project Consortium, et al., 2012), or sites with low mapping qualities are usually filtered from the final somatic call set. Validation rates ultimately depend on the stringency of this filtering of putative sites.

Intriguingly, applying different somatic calling algorithms to the same data often results in a set of only partially overlapping single nucleotide variant (SNV) sites. To illustrate this phenomenon, we deployed four publicly available somatic variant calling algorithms (Bambino, CaVEMan, SAMtools mpileup and VarScan 2) on a dataset composed of 24 human hepatocellular carcinoma exomes (Guichard, et al., 2012). Since this study reported 994 validated somatic variants identified using the independent CASAVA pipeline, we used these data to gauge the performance of each algorithm. This analysis revealed at best a 5.82% overlap between SNV calls made by any two of these widely used callers, and at worst a 0.11% overlap. Notable, however, was the fact that the majority of validated calls were identified by two or more algorithms, suggesting that a merging approach may improve both the sensitivity and accuracy of somatic variant calling. See the Supplementary Information for more details.

In an effort to take advantage of existing software tools and to improve variant detection we developed Cake (Supplementary Figure 1). Cake is a fully configurable bioinformatics pipeline that integrates four single nucleotide somatic variant calling algorithms (Bambino, CaVEMan, SAMtools mpileup, and VarScan 2), and deploys an extensive collection of fully customizable post-processing filtering steps. We show that the performance of Cake exceeds any one algorithm for somatic variant detection making it an optimal tool for cancer genome analysis.

2.2. Implementation

Cake is implemented in Perl, enabling the configuration, execution and monitoring of the four callers in a high-performance computing environment using a job scheduler. Alternatively Cake can be configured to run in standalone mode on a single computer (See the User Manual on SoureForge for more details). The existing choice of algorithms can be easily modified using a template we provide. A package containing the callers and the post-processing modules and install script is available for download.

		Hepatocellular carcinoma (24 samples / 842 validated sites)			Breast cancer (2 samples / 264 validated sites)			
		Validated mutations identified (total)*	% Sensitivity	Average number of variant calls per sample	Validated mutations identified (total)**	% Sensitivity	Average number of variant calls per sample	Validation success rate (Sequenom)
Single algorithms (after filtering)	Bambino	742	88.1%	2503 ± 1070	248	93.9%	3456 ± 324	
	CaVEMan	801	95.1%	1072 ± 1055	(263)	(99.6%)	(961 ± 90)	
	mpileup	727	86.3%	429 ± 226	181	68.6%	329 ± 32	
	VarScan 2	805	95.6%	926 ± 527	205	77.7%	929 ± 91	
Cake	≥ any 2 callers	812	96.4%	634 ± 299	254	96.2%	613 ± 42	51.5%
	≥ any 3 callers	794	94.3%	270 ± 132	214	81.1%	326 ± 50	81.7%
	4 callers	652	77.4%	168 ± 98	166	62.8%	178 ± 42	88.3%

Table 2.1: Summary of the results of different somatic variant calling algorithms and Cake on two human exome data sets.

2.3. Result

To evaluate the performance of Cake we used the above-mentioned human hepatocellular carcinoma dataset composed of 24 exome tumour / germline pairs and two human breast cancer exomes for which we had genomic DNA for follow-up validation (Nik-Zainal, et al., 2012). The performance of each variant calling algorithm was evaluated by running each one individually using their default settings and filtering the results using the post-processing filters implemented in Cake. The results are summarized in Table 1.

Human hepatocellular carcinoma dataset: In their study, Guichard, et al. (2012) experimentally validated 994 variants. Pre-processing of the original NGS files however left a reference set of 842 experimentally validated SNV positions. Using Cake with an intersection of two or more algorithms, 812 validated variants were retained (Supplementary Figure 2), representing an overall sensitivity of 96.4%. An average of 634 variants were predicted per exome (Table 1). Cake outperformed the

best single algorithm in terms of specificity and the number of variants reported per sample.

Human breast cancer exome dataset: Since the above analysis will favour callers that perform like CASAVA, and because we did not have DNA from the hepatocellular carcinomas for follow-up analysis to ascertain the false positive and negative rates, we next used exome data for two breast tumours for which whole genome amplified material was in hand. Using Cake and an intersection of two or more callers, we made a total of 1,225 calls (per sample 613±42), of which 254 were from a reference call set representing a subset of positions [264] covered by the capture baits where a somatic mutation had resulted in a non-synonymous change; a sensitivity of 96.2% [Table 1, Supplementary Figure 3]. Excluding CaVEMan, which was used in the original study, Cake again outperformed all other algorithms (Table 1).

To assess the specificity of the somatic variant calling by Cake used the Sequenom MassARRAY SNP genotyping platform on tumor and germline DNA samples. A total of 400 variants were randomly selected from the 1,225 calls made by at least two callers in the Cake pipeline; 200 from each sample. Two hundred and seventy variants were validated including 95 somatic mutations confirmed in the original study, 111 somatic mutation that were not described previously, and 64 germline variants. Importantly we called variants in a greater target region than the original study by analyzing positions in 5′ and 3′ UTRs, and introns (Supplementary Information). Nonetheless 22 novel non-synonymous somatic variants were discovered and confirmed (Supplementary Figure 4). These were all positions called by CaVEMan in the original study that had been filtered during post processing. Of the 22 novel non-synonymous calls, we find variants in HUWE1, MAP3K5 and RRM2, all of which have been implicated as cancer driver genes.

A further 400 variants (Supplementary Information) were included as a true negative set resulting in a worst-case accuracy for Cake of 75.8%. Although we used our default of at least two callers as part of the above-mentioned analysis, we note that 87.6% of validated calls were reported by all four callers (Supplementary Figure 3, Table 1). This indicates that merging predictions increases the probability of identifying true mutations.

2.4. Summary

Here we describe Cake, a software tool integrating four somatic variant detection algorithms to call variants with higher accuracy and specificity than any one algorithm alone. Cake performs well on whole genomes, exomes and targeted NGS data, as well as on both human and mouse samples. Cake is freely available to the research community.

2.5. Acknowledgements

We thank Patrick Tarpey and David Jones from the Can-cer Genome Project at the Sanger Institute for their assis-tance. Funding: Supported by CR-UK and the

Wellcome Trust. CDRE was supported by Consejo Nacional de Ciencia y Tecnología (CONACYT) and the Wellcome Trust.

2.6. Supplementary Materials:

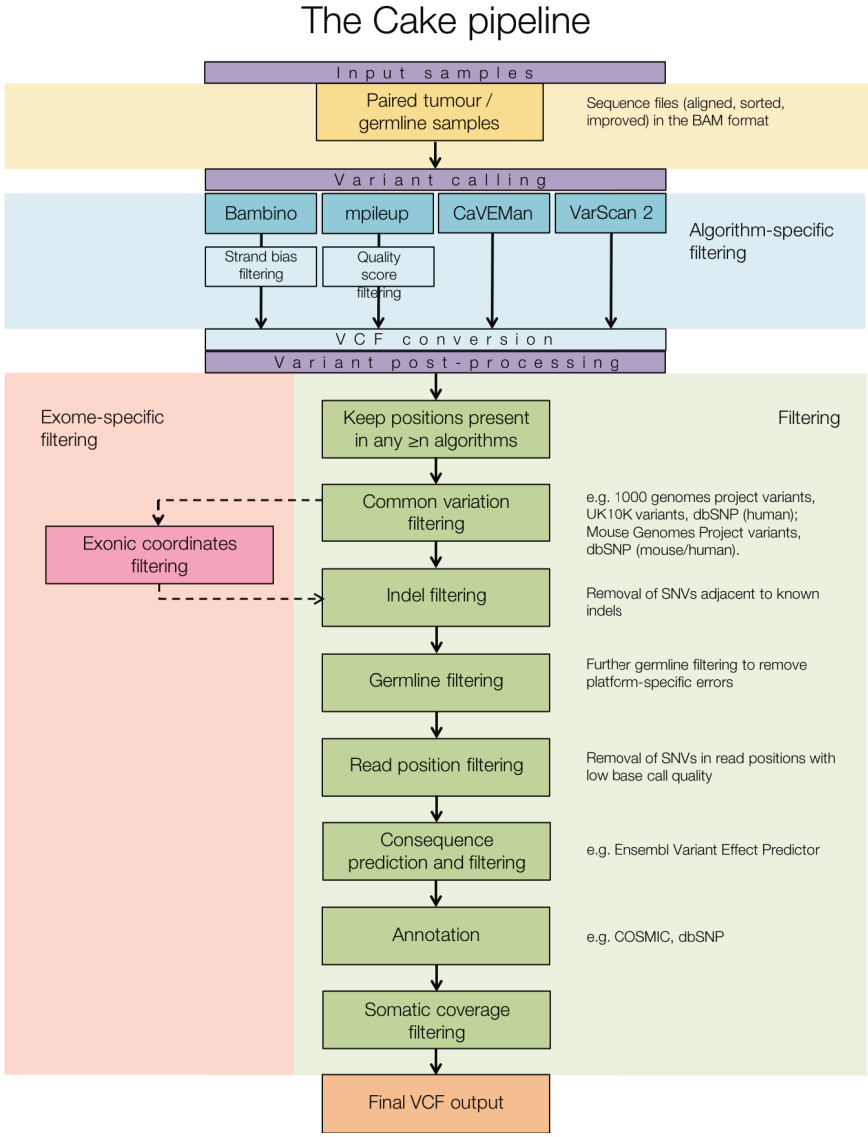


Figure 2.1: Cake somatic mutation detection pipeline.

2.6.1. Variant intersection strategy

Somatic variant callers produce outputs in different formats, e.g. genotype (VCF) or read counts. For uniformity and better compatibility, all outputs are converted to VCF format. By default, variants identified by at least any 2 out of 4 callers and reporting the same genotypes are processed through variant filtering. In the example below, all algorithms have called the same genotype in Position1 in both the tumour and the normal samples, and thus it will be considered for variant filtering in any of the 4 intersection approaches (right side of the Figure 2a). Conversely, Position4 is identified by 3 callers, but only two of them have called the same genotype. Then, it will be passed to the variant filtering stage only in the 'any out of 4' and 'at least 2 out of 4' callers strategies (Green and orange dotted rectangle).

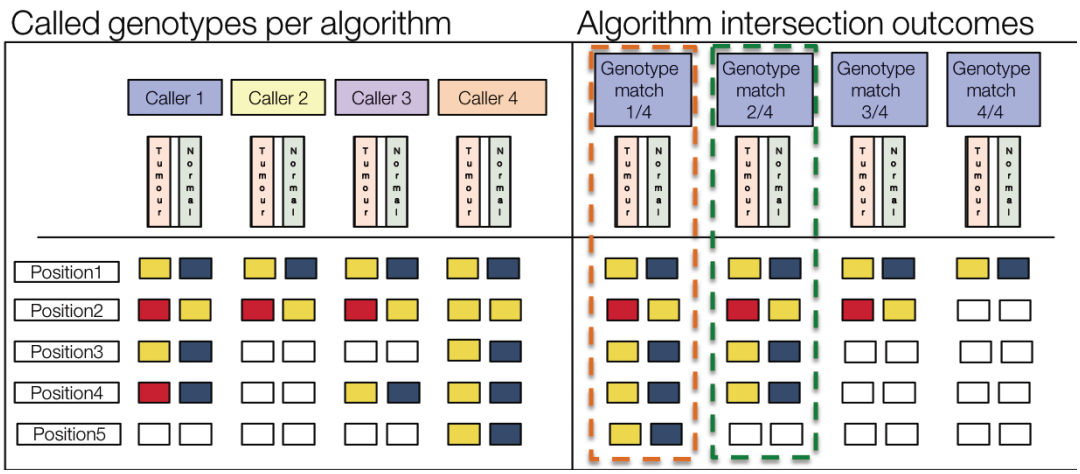


Figure 2.2: Variant intersection among multiple callers

Through this flexible intersection approach, Cake seeks to improve the sensitivity as well as the specificity. For a variant to pass through the intersection stage, it has to be identified by at least any n (n = number of callers specified by the user in the configuration file) out of all (4 by default) somatic callers. Variants missed by one caller may be detected by others, contributing to achieve a higher sensitivity. Moreover, overlapping across multiple callers controls the false positive rate.

2.6.2. Best intersection strategy

Choosing the best overlapping strategy is a non-trivial problem considering the complex landscape of cancers. For example, variable mutation rates across different cancer types combined with differences in sequencing technologies make it difficult to generate generic simulation data replicating the underlying complexity of different cancer types.

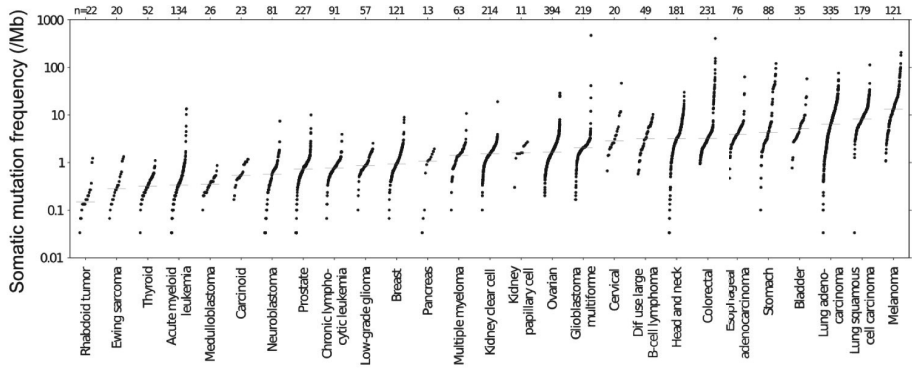


Figure 2.3: Spectra of somatic mutations across cancer types (Taken from Lawrence et al. (2013) Nature, in press).

Validation capacity (the availability of large-scale validation technologies and resources) also restricts the number of mutations/genes to follow-up. In Supplementary Table 3 we have tried to provide a rough guidance to users for selecting the best strategy for their data based on our experience. Users should take these as general guidelines rather than hard and fast rules, and assess each study individually based on the research question.

2.6.3. Additional data set analysis

In their study, Guichard et al. (2012) validated 850 single-nucleotide somatic variants from 24 human hepatocellular carcinoma tumour / normal exome pairs. At eight of these sites, we were unable to find read coverage following re-alignment of the data. These positions were excluded from downstream analysis. This left a control set of 842 somatic SNV positions. Using the Cake merge approach (≥ any 2 out of 4 callers), we identified 812 of these positions. The Venn diagram below shows the breakdown of these calls by caller and the overlapping calls.

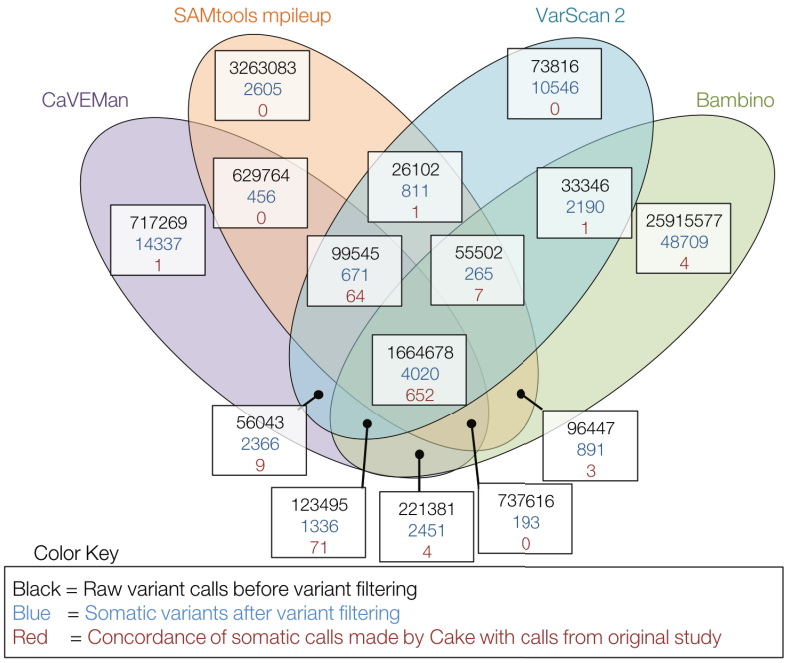


Figure 2.4: Human hepatocellular carcinoma data. Overlap between various somatic callers

Stephens et al. (2012) validated 264 somatic mutations from two breast cancer exome / germline pairs. Here we show somatic variant calls made by the algorithms in the Cake pipeline. Using an intersection of calls made by ≥ any 2 out of 4 callers, followed by variant filtering, we identified 254 of the 264 validated positions.

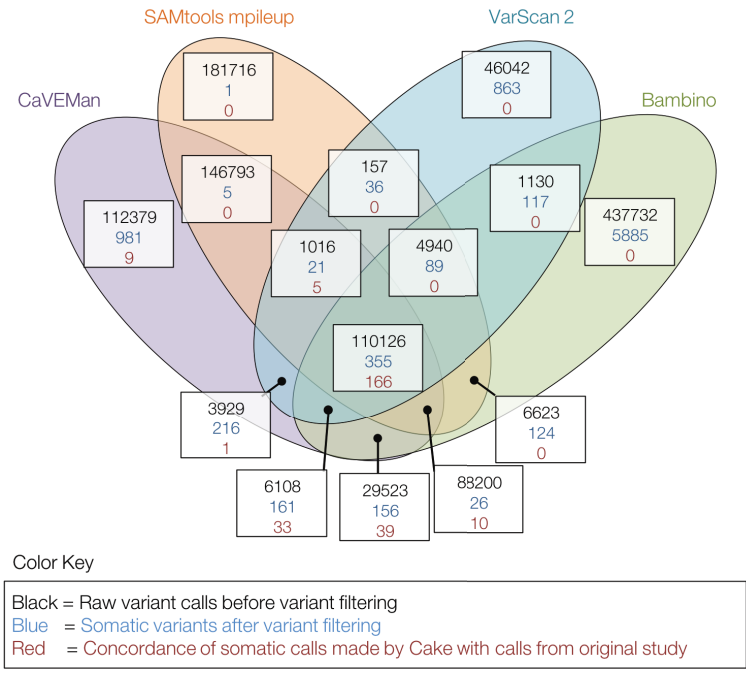


Figure 2.5: Human breast cancer data. Overlap between various somatic callers

To assess the sensitivity and specificity of the Cake merge approach (≥ any 2 out of 4 callers), we sent 400 predicted somatic variants for independent validation in the Sequenom MassArray platform. The Venn diagram below shows the distribution and breakdown of these variants (novel as well as those validated in original study). Variants that failed at any stage during the validation and for which we could not determine a result are not depicted.

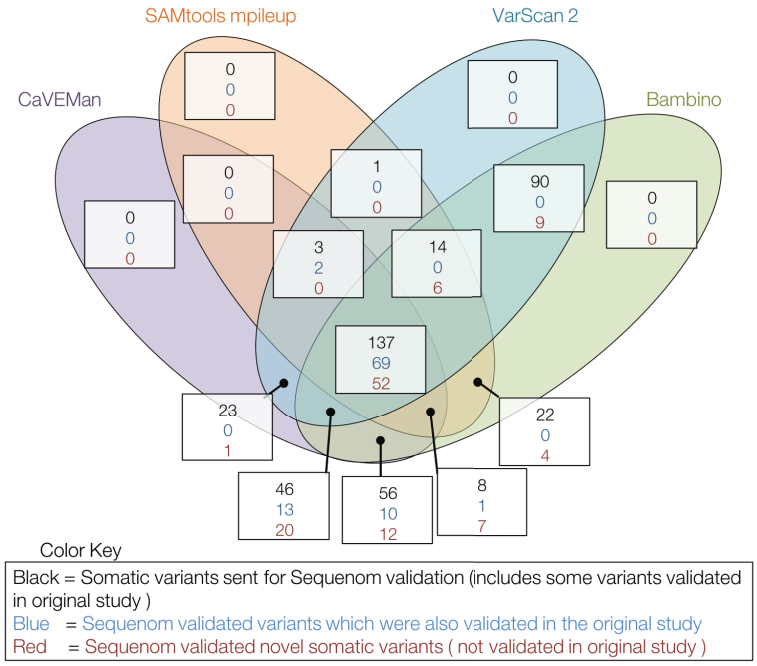


Figure 2.6: Validation of human breast cancer mutations using Sequenom

Algorithms	Number of overlapping variants	Total number of variants	Percentage overlap	Percentage of identified Validated Variants
CaVEMan & VarScan 2	1,943,761	4,438,557	43.80%	94.5%
CaVEMan & mpileup	3,131,603	7,690,925	40.72%	85%
mpileup & VarScan 2	1,845,827	6,859,437	26.90%	86%
Bambino & CaVEMan	2,747,170	30,350,663	9.05%	86.3%
Bambino & mpileup	2,554,243	32,866,536	7.77%	78.6%
Bambino & VarScan 2	1,877,021	29,103,548	6.45%	86.8%
Cake - ≥ any 2 out of 4	3,743,919	33,713,664	11.11%	96.4%
Cake - ≥ any 3 out of 4	2,680,836	33,713,664	7.95%	94.3%
Cake – 4 out of 4	1,664,678	33,713,664	4.94%	77.4%

Table 2.2: This table shows an overlap of raw variants called from the 24 human hepatocellular carcinoma / germline pairs using the algorithms in the Cake pipeline. No filtering was performed on these variants. A subset of these data are displayed graphically in Supplementary Figure 3.

	≥ any 2 out of 4 algorithms intersection	≥ any 3 out of 4 algorithms intersection	4 out of 4 algorithms intersection
Targeted gene follow-up (with prior biological hypothesis)	✓		
Large mutation sets for landscape discovery		✓	✓
Number of genes	Dozens	Hundreds/thousands	Whole genome/exome
Smaller sample cohort	✓		
Large sample cohort		✓	✓
Follow-up Validation (Capillary sequencing / Sequenom MassArray)		✓	✓
Follow-up Validation (454 pyrosequencing)	✓		

Table 2.3: General user guidelines for algorithm intersection strategy.

Somatic Variant	Gene symbol	Consequence	Status in the Stephens, *et al* (2012) study	Disease
1:8418341	*RERE*	Missense	Not seen	leukemia, squamous cell carcinoma
21:47810651	*PCNT*	Missense	Not seen	breast carcinoma
12:122517135	*MLXIP*	Missense	Seen but filtered	
17:55183040	*AKAP1*	Missense	Seen but filtered	prostate cancer
19:51413945	*KLK4*	Missense	Seen but filtered	breast carcinoma
X:153276548	*IRAK1*	Missense	Seen but filtered	non-small cell lung carcinoma

Table 2.4: Breakdown of novel missense variants called by Cake, human breast cancer data. Additionally, we validated 15 more somatic variants, found in non-coding transcripts, that were seen by CaVEMan but filtered out.

3

Adenoma development in familial adenomatous polyposis and MUTYH-associated polyposis: Somatic landscape and driver genes

Familial adenomatous polyposis (FAP) and MUTYH-associated polyposis (MAP) are inherited disorders associated with multiple colorectal adenomas that lead to a very high risk of colorectal cancer. The somatic mutations that drive adenoma development in these conditions have not been investigated comprehensively. Here we perform analysis of paired colorectal adenoma and normal tissue DNA from individuals with FAP or MAP, sequencing 14 adenoma whole exomes (eight MAP; six FAP), 55 adenoma targeted exomes (33 MAP, 22 FAP) and germline DNA from each patient, and a further 63 adenomas by capillary sequencing (41 FAP, 22 MAP). With these data we examine the profile of mutated genes, the mutational signatures, and the somatic mutation rates observing significant diversity in the constellations of mutated driver genes in different adenomas, and find loss-of-function mutations in WTX (9%; P < 9.99e-06); a gene implicated in regulation of the WNT pathway and p53 acetylation. These data extend understanding of early events in colorectal tumorigenesis in the polyposis syndromes.

3.1. Introduction

Over the last three decades there has been a dramatic improvement in our understanding of the genetic basis of germline susceptibility to colorectal cancer (CRC) (Ewing L et al. 2014) . This began with the identification of the adenomatous polyposis coli gene (APC) (Kinzler K et al. 1991) in which germline mutations cause familial adenomatous polyposis (FAP), followed by the discovery of other genes such as MSH2 in Lynch syndrome (Lynch HT et al. 1994), and MUTYH (mutY Homolog) in MUTYH-associated polyposis (MAP) (Sampson JR 2003) (also called MYH-associated polyposis). Germline variants in the genes LKB1, SMAD4, GREM1, PTEN, BMPR1A and AXIN2 have also been implicated in predisposition to colorectal cancer, and highlight a role for many pathways in tumorigenesis of the colon (Patel SG 2012). More recently, germline variants in the exonuclease domains of the polymerase epsilon catalytic subunit gene (POLE) and in the DNA polymerase delta catalytic subunit (POLD1) gene have been linked to colorectal adenoma and carcinoma development (Palles C 2013). Variants in POLE and POLD1 dramatically increase the somatic mutation rate resulting in C:G > T:A somatic base changes (Heitzer E 2014). While the majority of colorectal cancers are sporadic, variants in the abovementioned genes, and in several other high penetrance susceptibility genes including MLH1, MSH6 and PMS2, collectively account for around 5% of all cases (Patel SG 2012). While colorectal cancer is a common endpoint of germline variants in these genes they initiate tumorigenesis in different ways meaning that the landscape of somatic mutations, the genes that are mutated, and the paths to malignancy are likely to differ. As a corollary the multiplicity of tumours in patients with germline mutations in these genes differ, and clinical outcomes and responses to therapy can vary suggesting a complex interplay between the germline genetics of each patient and the somatic landscape of their tumours (Grover S 2012). Despite major initiatives to analyse sporadic colorectal cancer [TCGA 2012] little is know about the somatic landscape of mutations in tumours from patients with hereditary forms of the disease. Here we set out to profile somatic mutations in pre-malignant adenomas in two hereditary colon cancer syndromes; FAP and MAP.

In FAP, adenomas may develop following somatic inactivation of the wildtype allele of APC, an event that is thought to be among the earliest somatic changes occurring during tumorigenesis in these patients (Levy DB 1994). APC normally binds to GSK3β as a part of a complex called the destruction complex which regulatess β -catenin stability, and hence the output of the WNT pathway [11]. Loss or attenuation of the activities of the destruction complex results in elevated levels of β-catenin, and of downsteam effectors such as Cyclin-D, AXIN2 and BIRC5 (Clevers H 2014). These proteins participate in the cell cycle, growth, and the regulation of cell death, respectively. The location of the germline mutation in APC in a FAP patient and the mode by which the wildtype allele of the gene is inactivated during adenomagenesis influences the degree to which the WNT pathway is activated (Cheadle JP 2002). The level of WNT pathway activation influences the multiplicity of intestinal polyposis and the growth of the adenomas that form, and is described by the "just-right hypothesis" whereby optimal levels of WNT activation

drive cell growth without tipping cells into apoptosis or evoking cell death pathways (Lowe SW 2004). Fine-tuning of the WNT pathway is thus central to colorectal tumorigenesis (Segditas S 2006). MUTYH (MutY homolog) is a DNA glycosylase that removes adenines misincorporated opposite 8-oxo-7,8-dihydro-2′-deoxyguanosine (8-oxodG) (Sulpska MM 1996). Patients with MAP carry biallelic loss-of-function mutations in the MUTYH gene, which in targeted sequencing studies of adenomas was found to manifest as an increase in somatic G:C > T:A mutations at the APC locus (Al-Tassan N 2002). Loss of MUTYH by itself is not oncogenic akin to some other colorectal cancer syndromes such as Lynch syndrome. With the exception of the aforementioned mutations in APC, transversions in RAS resulting in the generation of a G12C amino acid change are the only other established somatic event in MUTYH-driven tumorigenesis. The landscape of somatic changes, the rate of somatic mutation, and the genes that are mutated in this disease are unknown.

Several studies have used next-generation sequencing of human colorectal cancers to survey their somatic mutational landscape. The Cancer Genome Atlas (TCGA) characterized the genomes of 276 sporadic colorectal cancers focusing almost exclusively on invasive cancers and metastatic tumours (TCGA 2012). Other studies have analysed the exomes of microsatellite instable (MSI) primary cancers (Timmerman B 2010), while Nikolaev et al., (Nikoleiv SI 2012) performed a detailed and comprehensive analysis of 24 sporadic adenomas revealing a signature of deamination (C > T at CpG sites), suggesting a role for replication stress in mutational acquisition. Here we focus on the early evolution of adenomas from MAP and FAP patients, and investigate the somatic mutation rate and the pattern of mutation. We also identify mutated driver genes, including truncating mutations in WTX (also known as FAM123B and AMER1).

3.2. Materials and Methods

3.2.1. Tumor collection :

Ethical approval and written informed consent from each participant was obtained under UK NHS Research Ethics Committee approvals 02/09/22 and 12/WA/0079. Adenomas were harvested at colectomy or polypectomy from the colorectum of patients with confirmed germline mutations in APC or MUTYH. Larger adenomas were halved longitudinally with one part being snap frozen in liquid nitrogen and the other formalin-fixed for histopathology. Smaller lesions were snap frozen and histopathology performed using a small sample cut from the frozen material. A description of the germline mutations carried by all patients and a summary of their clinical histories is available in Supplementary Table 1. Three sets of adenomas were analysed using either whole exome sequencing, targeted exome sequencing, or capillary sequencing of WTX, as described below. Histopathological analysis of adenomas was performed by a clinical gastrointestinal histopathologist (GTW). A summary of the pathology reports is provided in Supplementary Table 2. DNA was extracted using the Qiagen DNeasy Kit. Lesions selected for analysis were of similar

(sub-cm) size with most showing low-grade dysplasia.

PatientID	Cohort	Whole Exome Sequencing (WES)	Targeted Gene Sequencing (TGS)	Capillary Sequencing
1	MAP	5 adenomas	4 adenomas	4 adenomas
2	MAP	3 adenomas	5 adenomas	
3	MAP		3 adenomas	
4	MAP		21 adenomas	14 adenomas
5	FAP	2 adenomas		5 adenomas
6	FAP		10 adenomas	
7	FAP		11 adenomas	
8	FAP	4 adenomas	1 adenoma	
9	FAP			5 adenomas
10	MAP			4 adenomas
11	FAP			2 adenomas
12	FAP			5 adenomas
13	FAP			17 adenomas
14	FAP			7 adenomas
Total		14 adenomas	55 adenomas	63 adenomas

Table 3.1: Seven samples were sequenced by both whole-exome and targeted-exome sequencing but are not shown here; details are available in Table S2 (see supplementary material).

3.2.2. Whole exome and targeted exome sequencing:

Whole exome sequencing was performed using the Agilent whole exome capture kit (SureSelectXT Human All Exon 50Mb) as described previously (Coffey AJ, 2011). Captured material was indexed and sequenced on the Illumina platform at the Wellcome Trust Sanger Institute. Targeted capture sequencing was performed using baits designed against the genes APC, WTX/FAM123B, ATRNL1, BCL9L, BRCA1, BRCA2, CXCR5, DMD, FBXW7, GPR112, HUWE1, KMT2C/MLL3, NF1, PTEN, SLFN5, SMAD4, SORCS1, TP53, UBR2 and ZNF37A, genes selected for sequencing on the basis of being recurrently mutated, or truncated, in the unfiltered whole exome sequencing data. These genes are also enriched for non-silent mutations in the genome sequencing of sporadic CRC (TCGA 2012). A breakdown of the sequencing metrics for each sample is provided in Supplementary Figure 1. We collected and whole exome sequenced 14 adenomas (eight MAP and six FAP) from two MAP patients and two FAP patients, and corresponding germline control DNA for each patient. (Figure 1, Supplementary Figure 1 & Supplementary Table 2). In a similar way, 22 FAP and 33 MAP adenomas and corresponding blood leukocyte DNA from three FAP and four MAP patients were sequenced using the targeted bait set (Supplementary Table 2). DNA from all of the adenomas and corresponding leukocyte controls was available for follow-up genotyping/validation. Seven cases that were whole exome sequenced were also targeted exome sequenced (four FAP and three MAP) to compare these platforms (Supplementary Table 2). Table 1 provides a summary of the sequenced samples.

3.2.3. Somatic single nucleotide variant calling:

DNA sequence data from paired adenoma/normal constitutional DNA samples was presented to the Cake pipeline, which uses the somatic variant callers Bambino, CaVEMan, SAMtools mpileup and VarScan2 (Rashid M 2012). As described previously (Rashid M 2012) we used a somatic caller merging approach to identify somatic variants selecting only those detected using three or more of these algorithms for further analysis. We have previously shown that this approach increases the sensitivity and specificity of variant detection (Rashid 2012). These calls were further filtered using modules such as the 1000 Genomes Project phase 1 single nucleotide polymorphisms (SNPs), excluding variants with minor allele frequencies greater than 0.01, and by standard variant filtering.

3.2.4. Variant validation by Sequenom:

We attempted to validate all non-silent somatic variant calls from both the targeted and whole exome sequencing experiments using the Sequenom platform. Both normal tissue and adenoma DNA samples were analysed as described previously (Thomas RK 2007).

3.2.5. Capillary sequencing of WTX and KRAS:

WTX was one of several genes found to carry truncating mutations and was capillary sequenced in a larger panel of adenomas (41 FAP and 22 MAP; Supplementary Table 2) to extend the data collected from the whole exome and targeted exome sequencing experiments. In brief, primers were designed against each exon of the gene, amplicons were bi-directionally sequenced, and variants called using the Mutation Surveyor Software followed by manual inspection. KRAS sequencing was performed as described by Jones et al (Jones S 2004).

3.2.6. Mutation signature analysis:

We interpreted the mutational catalogue of MAP and FAP using validated variants called from the whole exome sequence data, and also the raw variant calls made by the Cake pipeline. For this task we used EMu, a probabilistic algorithm that infers the number of mutational processes operative and their individual signatures (Fischer 2013). Mutations were mapped to the 96 possible trinucleotide combinations taking into account the possibility for each mutation to occur in the context of each trincleotide type within the human genome. As the model underlying EMu assumes that the input samples are independent, we further collapsed the mutation data by patient and performed a patient centric signature analysis.

3.2.7. Statistical analysis of *WTX* mutations:

To determine whether *WTX* was significantly enriched for nonsense mutations we used Monte-Carlo simulations. Over 100,000 iterations were generated where six nucleotide changes were randomly introduced into the *WTX* sequence (six being

the number of changes found in *WTX* in the targeted sequencing experiment; five nonsense and one synonymous) using the underlying base change probability from TCGA data across all tumour types. We then computed all possible outcomes for these mutations from each iteration and compared these frequencies to the frequency of truncating mutations found in the targeted exome analysis.

3.3. Results

3.3.1. Calling and validation of somatic variants:

We attempted to validate all non-silent positions at which a candidate somatic variant call had been made from the whole (573) or targeted (45) exome data using the Sequenom platform. To do this we genotyped DNA from each adenoma, and a matched normal tissue or leukocyte control DNA sample. We successfully designed assays against 434/573 positions from the whole exome sequencing experiment, and 42/45 positions from the targeted exome sequencing experiments. The overall validation rate for the 434 calls from the whole exome sequencing of the MAP/FAP polyps was 80.87% (351 successfully genotyped somatic SNVs). The validation rate for the targeted exome experiment was 95.23% (40/42). Supplementary Table 3 & 4.

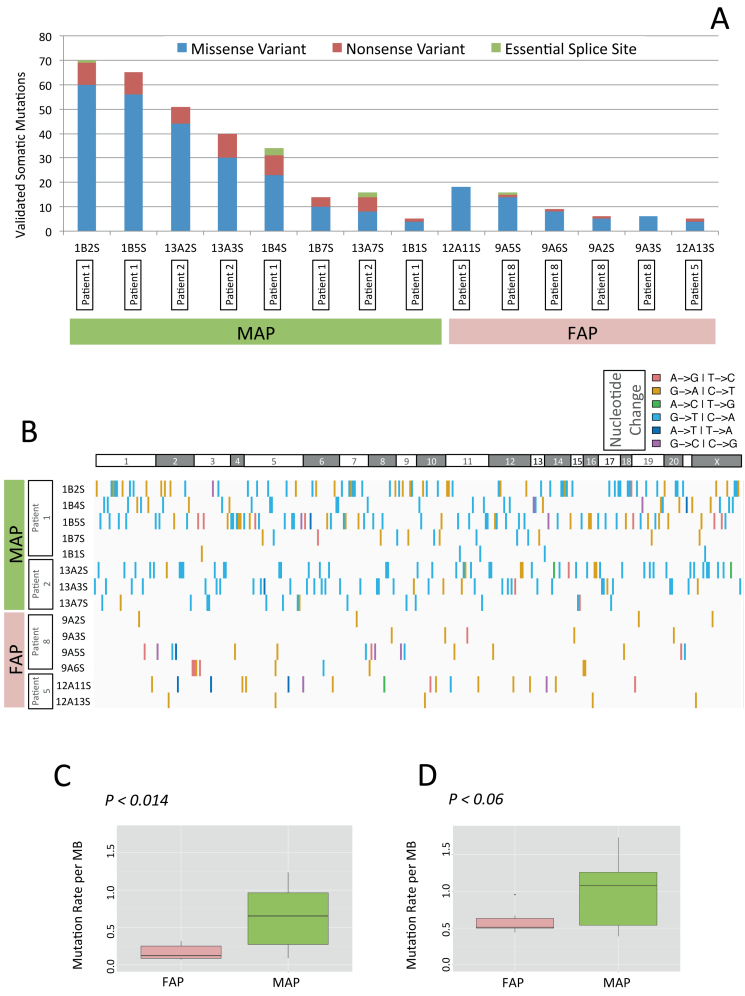


Figure 3.1: Somatic mutation calls from adenomas from patients with MAP or FAP. (A) Validated protein-changing somatic mutation calls fromMAP and FAP adenomas: missense, nonsense and splice site variants are shown. (B) Profile of the validated somatic calls for the variants shown in (A): the type of nucleotide change is indicated by the colour of each bar for each patient; adenoma IDs for each patient are shown on the y axis (for clinical details, see AQ2 supplementary material, Tables S1, S2); vertically printed column at left denotes chromosomes. (C) Box plots showing, per megabase (Mb), median and 25th and 75th percentiles of validated somatic variant calls calculated from the data shown in (A). (D) Box plots showing, per Mb, mean and SD of all somatic variant calls made using the Cake pipeline (Rashid M 2012); p values represent the results of Student's two-tailed t-test

3.3.2. The frequency and distribution of somatic mutations in FAP and MAP adenoma exomes

Figure 1a shows the breakdown of Sequenom-validated protein-changing or disruptive somatic variants by adenoma and disease and their mutational class; missense, nonsense or essential splice site. Figure 1b shows the breakdown of validated non-silent somatic variants by mutational class, disease, tumour, and patient. All variant calls including synonymous variants are shown in Supplementary Figure 2a, and their mutational profile is shown in Supplementary Figure 2b. Variant validation metrics are provided in Supplementary Figure 3. Analysis of the somatic mutational landscape of protein-changing variants in this way revealed a mean somatic mutational burden of 0.65 mutations per Mb in MAP adenomas compared to 0.16 per Mb in FAP adenomas ($P < 0.014$) (Figure 1c). When using the raw output of the Cake pipeline there were 0.98 and 0.59 variants per Mb for MAP and FAP, respectively ($P < 0.06$) (Figure 1d). These findings suggest an increased mutational burden resulting from loss of *MUTYH* activity in the range of 1.6-3.9 fold, comparable to findings reported in a recent study of lymphoblastoid cell lines established from MAP patients (Grasso F 2014). Importantly, we observed a significant increase ($P < 0.012$; Student's two-tailed t-test) in the proportion of truncating mutations found in adenomas from MAP patients compared to those from FAP patients, Figure 1a. This observation may reflect the fact that there are different profiles of chromosomal imbalances in FAP and MAP adenomas (Cardoso 2006), such that tumour suppressor genes undergo allelic loss in FAP rather than being disrupted by point mutations.

3.3.3. The mutational signatures of FAP and MAP:

We determined the pattern and distribution of somatic nucleotide changes found in the eight MAP and six FAP adenomas that were whole exome sequenced. Using both validated variant calls (Figure 1) and also the output of the Cake pipeline (Supplementary Figure 2 & Supplementary Table 3) we used EMu software to discern mutational signatures as a way of identifying mutational processes that may be operative. This analysis revealed strong statistical support (Delta Bayesian information criterion [Delta-BIC] score > 171) for two distinct mutational processes; Signatures A and B when all 573 variant positions were used (Supplementary Figure 4). Analysis of validated calls assuming that two signatures are present led to similar results (Figure 2a). Both signatures include C > T mutations at XpCpG sites, compatible with spontaneous deamination, but signature A also included C > A mutations, especially at TpCpX sites. The latter may result from sequence motifs that have enhanced mutability. We also used EMu to estimate the mutational composition of the 14 individual adenomas with respect to the two mutational processes (Figure 2b). Signature A, which is composed primarily of C:G > A:T transversions that are typically associated with the failure to remove misincorporated adenines opposite 8-oxo-7,8-dihydro-2'-deoxyguanosine (8-oxodG), was the dominant signature in MAP adenomas (Figure 2b). In contrast the dominant mutational signature in FAP adenomas was Signature B (Figure 2b).

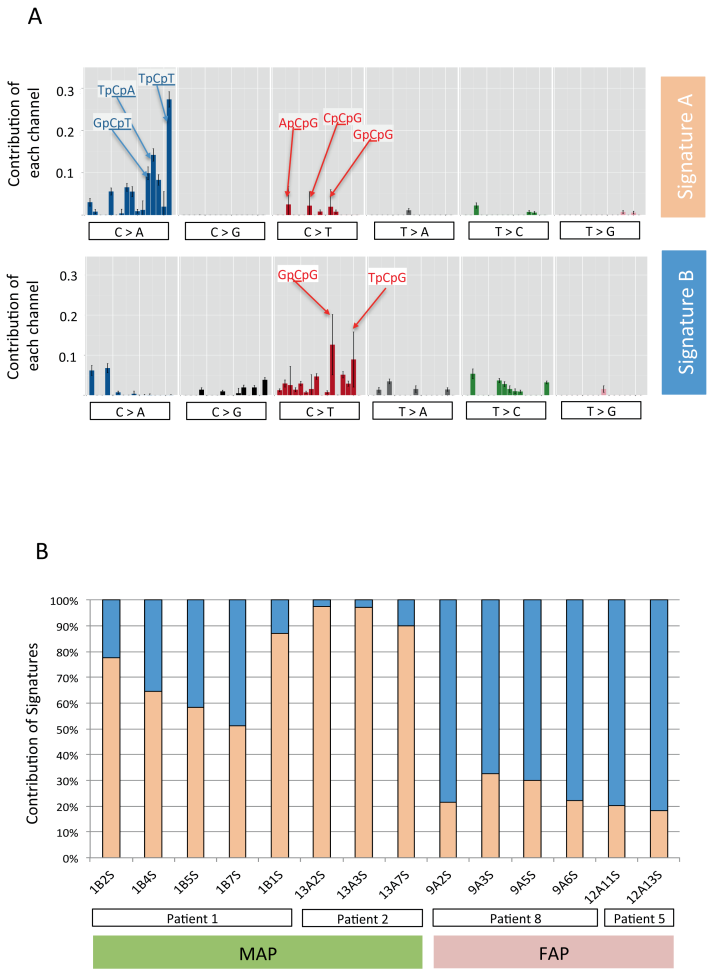
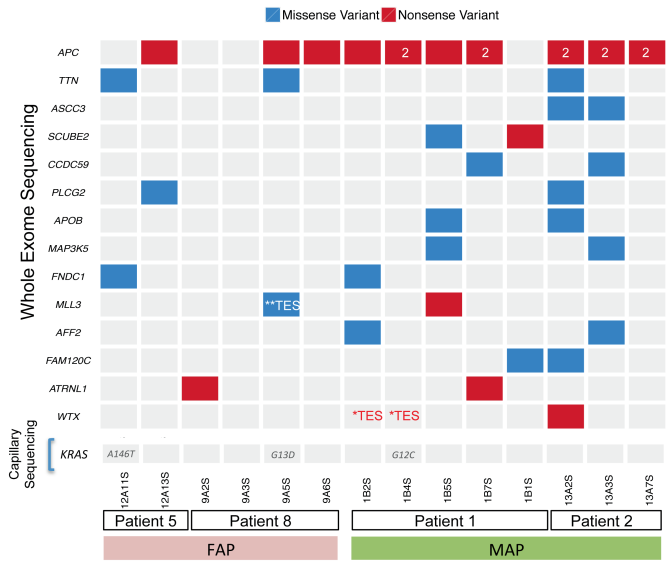


Figure 3.2: Mutational signatures in MAP and FAP. (A) The mutation spectra across 96 mutational channels (each representing a trinucleotide context, as described previously Alexandrov et al. 2013). (B) Mutational signature activity plot, indicating the proportion of somatic mutations found in adenomas from MAP and FAP patients that can be attributed to either signature A or signature B; the validated positions in Table S3 (see supplementary material) were used for this analysis.

3.3.4. Driver mutations in MAP and FAP adenomas:

We used the 351 validated somatic variants from exome sequencing (Supplementary Table 3) to ask which genes are likely to be driver genes in colorectal adenoma formation. Of the six FAP adenomas, three had somatic nonsense variants in APC, all falling into the β-catenin binding domains of APC and distal to the germline APC

mutation found in these patients (Figure 3). Of the eight MAP adenomas five had biallelic nonsense APC mutations. Truncating mutations were also validated in the genes *SCUBE2, RELN, FBXW7, MLL3, WTX/FAM123B, OTUD7B* and *KPRP* across the MAP and FAP adenomas (Supplementary Table 3). Two adenomas (Figure 3) were found to carry truncating mutations in the attractin-like 1 (ATRNL1) gene. Known driver genes from the cancer gene census in which we identified missense mutations included MAP3K5 and NRAS (p.Q61K) (Figure 3 & Supplementary Table 3). Two adenomas carried missense mutations in the phospholipase C, Gamma2 gene (*PLCG2*), which is related to *PLCG1* recently described as a driver gene in angiosarcoma (Behjati S et al. 2014). We also found three adenomas carrying protein-changing KRAS mutations; MAP polyp 1B4S carried a p.G12C, while FAP polyps 12A11S and 9A5S carried p.A146T and p.G13D changes, respectively. All these variants were validated by capillary sequencing. We next designed a custom capture bait set against genes identified as carrying truncating mutations or as being recurrently mutated in the unfiltered whole exome sequencing data (see Methods). Analysis of 55 adenomas (33 MAP and 22 FAP) and corresponding control DNA using this bait set yielded the Sequenom-validated somatic mutations shown in Figure 4a.



Overviews of the somatic mutations called by the Cake pipeline, the results of their validation by Sequenom genotyping, and their mutational profile are shown in Supplementary Figure 5 & Supplementary Table 4. Of particular note was the identification of truncating mutations in *WTX* and mutations in genes such as *TP53, FBXW7* and *PTEN*. Four MAP polyps were found to carry canonical KRAS p.G12C mutations resulting from a somatic G:C > T:A change; GGT > TGT, a figure in accord with a previous report (Jones S et al. 2004). Several adenomas that were

Figure 3.3: Candidate driver genes in adenomas fromMAP and FAP patients. The most frequently mutated genes inMAP and FAP adenomas are shown: red boxes, nonsense mutations; blue boxes, missense mutations; a number in a variant box indicates where multiple mutations of the same class are found. KRAS mutational status, determined by capillary sequencing, is also shown; some of the tumours were also sequenced by targeted-exome sequencing (TES); **TES, missense mutations; * TES, nonsense mutations (see supplementary material, Table S2). All positions were validated by Sequenom genotyping of tumour and control DNA; grey indicates no mutation found

whole exome sequenced were also targeted exome sequenced (Table 1, Supplementary Table 2 & Supplementary Table 4). Owing to the extremely high depth of sequence coverage obtained in the targeted exome studies we identified and validated additional driver mutations not seen by whole exome sequencing (Figure 2 & Supplementary Table 4).

3.3.5. *WTX* mutations in FAP and MAP:

During the targeted sequencing of 33 MAP and 22 FAP adenomas we identified and validated five truncating *WTX* mutations, all in MAP lesions, representing a statistically significant enrichment of truncating mutations in this gene (P < 9.99e-06). *WTX* mutations have been reported in advanced colorectal cancer, but their role in early stages of colorectal tumorigenesis is unknown. In order to determine whether there were differences in the frequency or profile of *WTX* mutations between early MAP- and FAP-associated adenomas we employed capillary sequencing to screen the exons and exon-intron boundaries of WTX in a further 22 MAP and 41 FAP adenomas. We identified nine further truncating mutations (including one frameshift mutation), six in FAP adenomas and three in MAP adenomas, and one missense mutation in a MAP adenoma (Figure 4b & Supplementary Table 5). The 17 truncating mutations of *WTX* identified in the different phases of our study are all likely to impact the function of its β-catenin binding region Figure 4b. Although *WTX* is on the X chromosome we identified mutations in adenomas from both male and female patients. We did not observe somatic biallelic mutations in adenomas from females, suggesting that lyonization may be responsible for loss of *WTX* function. *WTX* was originally identified as a gene involved in the development of Wilms' tumour of the kidney (Rivera MN et al. 2007) and has reported roles in the regulation of the WNT pathway, TP53 and cell fate, and in the localisation of the tumour suppressor protein WT1 (Kim WJ et al. 2012, Rivera MN et al. 2009, Moisan et al. 2011). Germline truncating mutations in *WTX* have also been linked to a sclerosing skeletal dysplasia (OSCS; MIM300373) and are not considered to be associated with an increased risk of tumours, although relatively early onset colorectal cancer occurred in one of 25 adult patients in the original report (Jenkins ZA et al. 2009). Mass spectrometry studies have revealed that *WTX* forms a complex with β-catenin, *AXIN1*, β-TrCP2 (β-transducin repeat–containing protein 2), and APC (adenomatous polyposis coli) to promote the ubiquitination and degradation of β-catenin (Major MB et al. 2007). Knockdown experiments have shown *WTX* to be a negative regulator of the WNT pathway [33]; thus mutation of *WTX* may result in activation of this pathway and the promotion of tumorigenesis. Figure 5 shows the location of the somatic *APC* muta-

tions identified in each adenoma and their *WTX* mutational status. The majority of the *APC* mutations found in adenomas with WTX mutations fell into the β-catenin or mutator cluster region (*MCR*) potentially resulting in impaired formation of the destruction complex rather than its complete loss (Chandra SHV et al. 2012), a scenario that may allow further modulation of β-catenin signalling via changes in WTX expression.

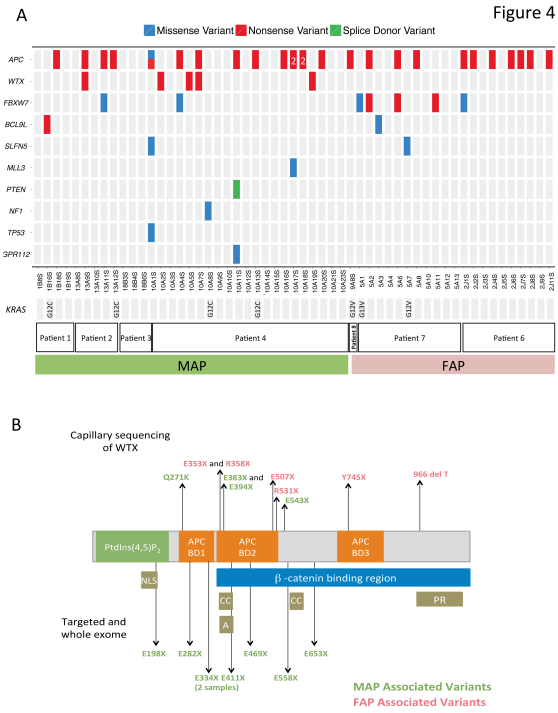


Figure 3.4: WTX is a cancer driver in adenomas inMAP and FAP. (A) The validated somatic mutations identified by targeted resequencing of 33 MAP and 22 FAP adenomas are shown (see supplementary material, Table S4): red, nonsense somatic mutations; blue, missense variants; green, essential splice site variants; KRAS mutations, which were assessed by capillary sequencing, are also shown; grey indicates no mutation found. (B) Somatic variants identified in WTX by whole- and targeted-exome sequencing (top) and capillary sequencing (bottom): the protein domains and positions are derived from Ensembl (ENST00000330258); NLS, nuclear localization signal; AA, acidic region; CC, coiled-coil domain; PR, proline-rich region; BD, binding domain

3.4. Discussion

In this study we explore the somatic mutational landscape of early stage pre-malignant adenomas from patients with germline mutations in APC and MUTYH

as a first step towards defining the catalogue of mutated genes. This task is essential for identifying which sets of mutations are most likely to lead adenomas to progress to colorectal cancer. We reveal that MAP adenomas have approximately two-to-four times the number of coding region somatic mutations when compared to FAP adenomas and that these mutations are overwhelmingly G:C > T:A mutations, in keeping with the expected signature associated with MUTYH loss. This observation confirms, for the first time, the expectation that deficiency of MUTYH leads to a mutator phenotype in colorectal tumours and is consistent with the observation of substantial colorectal cancer risk in MAP, even in the absence of dense polyposis (Nieuwenhuis MH et al. 2012). We find significant complexity in the patterns of mutated genes such that, with the exception of APC, KRAS and WTX mutations, few adenomas have the same set of mutated driver genes, a novel observation that may have implications for the definition of high risk adenomas in the era of molecular pathology.

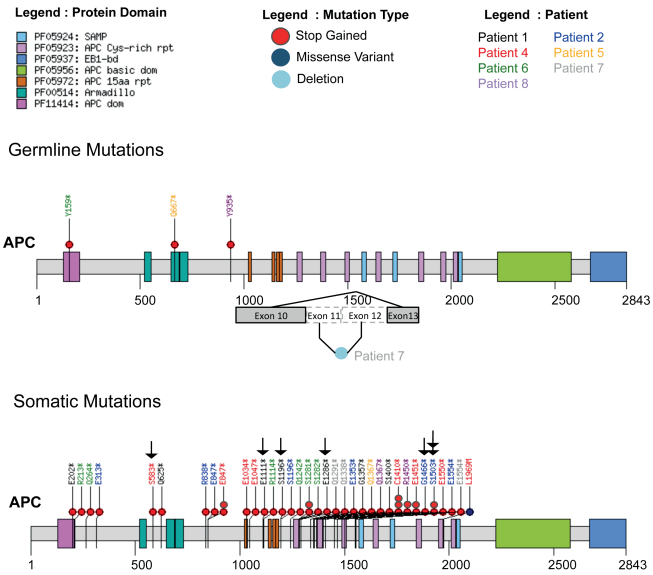


Figure 3.5: Germline and somatic APC mutations and somatic WTX mutational status. Predicted consequences of germline and somatic APC mutations found in adenomas from FAP and MAP patients are shown. Colour-coded circles indicate mutation type. Ensembl APC protein isoform ENST00000257430 and colour-coded domains are shown: FAP, patients 5–8; MAP, patients 1, 2 and 4; mutations are colour-coded by patient; arrows, lesions also carrying truncating WTX mutations identified by next-generation sequencing (see Figures 2, 4). Where an adenoma was found to carry bi-allelic APC mutations, we indicated the presence of a truncating WTX mutation against both mutations. We sequenced and validated somatic APC mutations in two adenomas from patient 2 that resulted in an amino acid change, p.S1503*. Both of these adenomas also carried truncating somatic WTX mutations (at positions p.E411X and p.E558X)

In FAP and MAP we found 50% of tumours carried somatic APC mutations (in the whole and targeted exome experiments combined). Notably the frequency of biallelic mutations was found to differ between our whole exome and targeted exome study despite an identical ascertainment of samples. In those lesions in which we did not find loss-of-function mutations in APC it is possible that the gene is disrupted by an imbalance of chromosome 5 (Cardoso J 2006) or by copy number neutral changes at the APC locus (Segditsas S et al. 2012). The number of APC mutations we observe here is slightly lower than estimates from the TCGA who report frequencies of between 60-80% for hypermutated and non-hypermutated cancers, respectively, and probably reflects the difficulties of identifying somatic mutations in lesions with low tumour cellularity. Intriguingly, across all protein-coding genes we observed a significant enrichment for truncating mutations in MAP adenomas compared to FAP ($P < 0.012$). We also observed a significant ($P < 9.99e-06$) incidence of truncating WTX mutations, with only APC being more frequently mutated. Our study demonstrates that in patients with MAP or FAP diverse molecular mechanisms are operational, even at early stages of colorectal tumorigenesis. This suggests that medical therapies for these disorders may be most effective if they target the initiating events of tumorigenesis prior to the development of mutationally diverse adenomas.

3.5. Author Contributions

MR, AF, AGR, VM performed computational analysis. CHW, JT, PS, SI, JM performed lab experiments and contributed to the analysis. GTW performed histopathological analysis. MR, CHW, JSR, DJA designed the experiments and wrote the paper.

3.6. Acknowledgements

Members of the DNA sequencing group at the Sanger Institute and the patients and their families. Dr. Victor Quesada for Monte Carlo code. Drs. Chi Wong and Daniela Robles Espinoza for helpful comments.

3.7. Grant Support

This work was supported by Cancer Research UK, the Wellcome Trust, the ERC Synergy Programme, and the Wales Gene Park.

3.8. Supplementary materials:

List of online supplementary materials:

- Figure S1. Sequencing metrics for the sequencing of adenomas and matched normal tissue samples from patients with MAP or FAP
- Figure S2. Somatic variant calls made by the Cake pipeline from MAP and FAP cases

- Figure S3. Results of Sequenom validation experiments of somatic variant calls made using Cake against adenoma/matched normal tissue pairs from MAP and FAP cases
- Figure S4. Mutational signatures in MAP and FAP using all 573 calls made by the Cake pipeline
- Figure S5. Results of targeted sequencing of adenoma/matched normal tissue pairs from MAP and FAP patients
- Table S1. A summary of the clinical details of each patient
- Table S2. A summary of the samples sequenced as part of this study
- Table S3. Whole-exome variant calls
- Table S4. Targeted-exome variant calls
- Table S5. Capillary sequencing of WTX and variant calls

All supplementary materials available at :
https://onlinelibrary.wiley.com/doi/abs/10.1002/path.4643

4

Genomic analysis and clinical management of adolescent cutaneous melanoma

Melanoma in young children is rare, however its incidence in adolescents and young adults is rising. We describe the clinical course of a 15-year-old female diagnosed with AJCC stage IB non-ulcerated primary melanoma, who died from metastatic disease four years after diagnosis despite three lines of modern systemic therapy. We also present the complete genomic profile of her tumour and compare this to a further series of 13 adolescent melanomas, and 275 adult cutaneous melanomas. A somatic BRAFV600E mutation and a high mutational load equivalent to that found in adult melanoma, and composed primarily of C>T mutations was observed. A germline genomic analysis alongside a series of 23 children and adolescents with melanoma revealed no mutations in known germline melanoma-predisposition genes. Adolescent melanomas appear to have genomes that are as complex as those arising in adulthood, and their clinical course can, as with adults, be unpredictable. The survival from advanced melanoma in adults has been revolutionised by the introduction of immune checkpoint inhibitors and molecular targeted therapies, however children and adolescents younger than 18 years have had limited access to the registration clinical trials. We present a detailed genomic analysis of a series of adolescent melanomas and the clinical course of one such patient who died from metastatic disease. A high mutational load was observed, and suggests that immune-based therapies may be relevant, but response cannot be guaranteed. Germline mutations in established adult melanoma-predisposing genes were not evident in an extended childhood and adolescent series. Given the complexities around diagnosis and the paucity of prospective clinical studies for younger individ-

uals, melanoma in this age group represents a particular clinical challenge requiring specialist management by a dedicated multidisciplinary team.

4.1. Introduction

Melanoma in children is rare, accounting for only 2% of all malignancies in patients younger than 20 years (Howlader, 2016). Melanoma in infancy and early childhood (1-10 years) comprise around 8% of newly diagnosed cases in young people, whereas adolescents (11-20 years) account for the majority (92%) of melanoma cases (Lorimer et al., 2016). Importantly, melanoma incidence in the adolescent population is rising at a rate of 2% per year (Austin et al., 2013). Melanocytic lesions in children comprise a heterogeneous group of neoplasms that can be broadly divided based on histology and age onset, and three major subtypes have been described (Barnhill and Kerl, 2006). Firstly, melanoma can arise in association with a pre-existing, usually large, Congenital Melanocytic Neavus (CMN) (Guegan et al., 2016; Trozak et al., 1975). The lifetime risk of malignant transformation from a CMN is 5-10%, and 50% of these transformations are said to occur in the first decade of life (Bett, 2005; Krengel et al., 2006). The second type are termed Spitzoid melanocytic tumours, which comprise a wider spectrum of histological variants including Spitzoid melanoma and atypical Spitz tumours. The vast majority of Spitz naevi occur in individuals younger than 20 years and often arise on the extremities and face (Reed et al., 2013). The third subtype, generally occurring in adolescents, has been termed 'conventional' or adult-type melanoma, owing to its shared clinical and histological features typical of adult melanoma. In contrast to infantile and childhood cases, post-pubertal melanoma is most often sporadic, occurring as a de novo lesion in patients with fair-coloured skin and substantial sun exposure (Wood, 2016).

Cutaneous melanoma in adults is characterised by a high prevalence of somatic mutations and the mutational pattern depicts a characteristic ultraviolet-light (UV)-induced signature associated with frequent transitions at dipyrimidine sites (Cancer Genome Atlas Network, 2015). A recent comprehensive genomic analysis found that tumours from adolescents bear a remarkably similar mutational rate and spectrum to tumours from adults, suggesting that sun protection practices are important in early life (Anderson et al., 2009; Lu et al., 2015). In addition to its rarity and the low clinical suspicion for malignancy, there is recognition that melanomas in young people are commonly amelanotic and the clinico-pathologic features may overlap with proliferative nodules and other benign skin lesions that are generally more common in children than adults (Cordoro et al., 2013; Moscarella et al., 2012). This can lead to delays both in diagnosis and treatment (Neier et al., 2012)

Several high-risk mutations have been identified in melanoma-dense families, including mutations in the cyclin-dependent kinase inhibitor 2A (CDKN2A) gene (Cannon-Albright et al., 1992), the cyclin-dependent kinase 4 (CDK4) gene (Zuo et al., 1996), and more recently in the Breast cancer 1 (BRCA1) associated protein 1 (BAP1) (Aoude et al., 2013; Wiesner et al., 2011) and protection of telomeres 1 (POT1) genes (Robles-Espinoza et al., 2014; Shi et al., 2014). However, the prevalence of these predisposing mutations amongst younger patients is largely unknown. A deeper understanding of the molecular drivers in childhood and ado-

lescent melanoma would advance our understanding of its pathogenesis, particularly the role of gene-environment interactions in susceptible cases, and could help define particular high-risk subgroups that might benefit from specialist screening and surveillance.

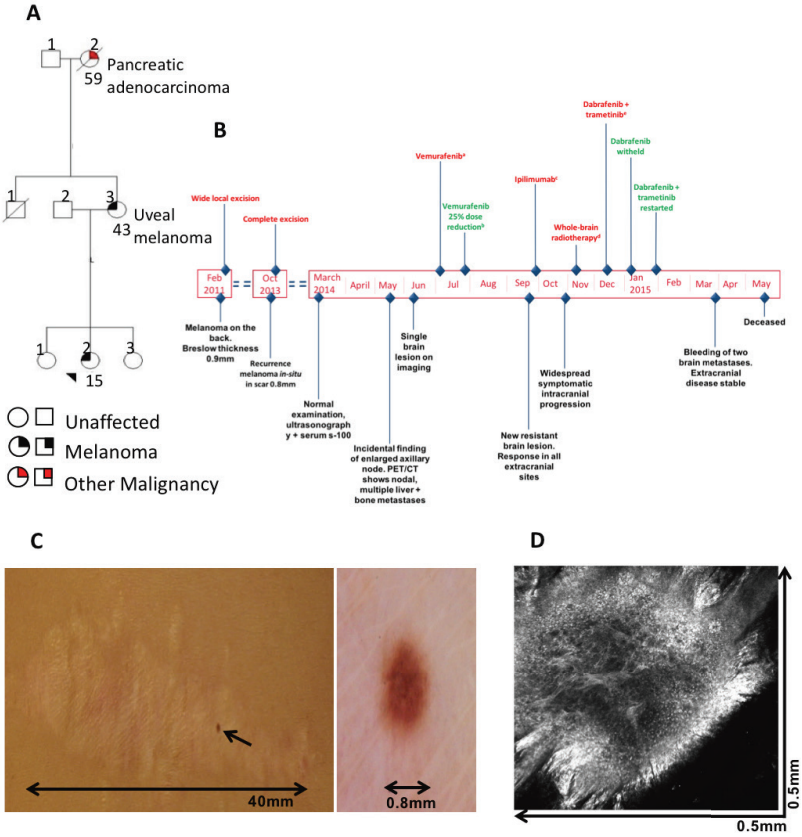


Figure 4.1: Clinical timeline of the 15 year old index patient (M_4180). (A) Family pedigree. The proband is indicated with an arrow, ages at diagnosis are shown. (B)Timeline of diagnosis and treatment. (C) New pigmented melanoma in situ appearing in the centre of previous melanoma wide local excision scar. Accompanied dermoscopic image of the insitu melanoma prior to further wide local excision (beside). (D) Reflectance confocal microscopy at the dermoepidermal level, showing proliferation of dendritic atypical melanocytes. [a]Vemurafenib starting dose 960 mg twice a day. [b]Dose reduction vemurafenib to 720 mg twice a day. [c]Ipilimumab 3 mg/kg every 3 weeks. [d]Wholebrain radiotherapy 10 Gy in 10 fractions. [e] Dabrafenib 150 mg twice a day, trametinib 2 mg once a day.

4.2. Result

In this study we present a detailed clinical history of one patient and an extensive genomic analysis of this patient's germline and tumour, and that of a wider series

of adolescent melanomas.

Patient Presentation

The 15-year-old female described had blonde hair, blue eyes, skin phototype II on the Fitzpatrick Classification Scale (Fitzpatrick, 1988), and a history of multiple (>50) benign skin naevi. Her mother had a history of uveal melanoma and her maternal grandmother had pancreatic adenocarcinoma (Figure 1a). She presented in February 2011 with an enlarging symmetric raised light brown papule on the right lower posterior chest wall at the level of the costal margin, which measured less than 1cm in diameter. The lesion was removed by shave-excision at her local hospital and was found to be a non-ulcerated cutaneous malignant melanoma, Clark's level IV, Breslow thickness 0.9mm and 6 mitoses/mm2 (Figure 1b). As the lesion extended to the excision margins, wide local excision was undertaken with subsequent clear margins. Ultrasonography revealed no pathological regional lymph nodes, and she underwent active multimodality six-monthly surveillance.

Two and a half years later in October 2013, a 0.8mm pigmented lesion appeared in the centre of the existing wide local excision scar (Figure 1c). Dermoscopic examination revealed a homogeneous pattern and reflectance confocal microscopy showed atypical dendritic cells in the dermo-epidermal junction (Figure 1d). This lesion was diagnosed as melanoma in situ, which was completely excised. In March 2014, no abnormalities were detected on surveillance clinical examination or ultrasonography and serum s-100 levels were recorded as normal at 0.13µg/L (<0.15µg/L). However, two months later, during a separate clinic consultation for acne treatment, an enlarged lymph node was detected in the right axilla and serum s-100 levels were now elevated at 0.7µg/L. A PET/CT scan revealed avid FDG uptake in multiple liver and bone metastases as well as right axillary lymph nodes (Figure 2a). A single asymptomatic brain metastasis was also identified on imaging (Figure 2b). Three cutaneous metastases were evident, one of which was excised. Molecular analysis of the excised metastasis using PCR revealed a BRAFV600E mutation.

In July 2014, she was commenced on systemic therapy with the BRAF kinase inhibitor, vemurafenib. Ten days into therapy she experienced arthralgia, blepharitis, meibomian gland inflammation (presenting with suppuration from the sebaceous gland at the rim of the eyelids and treated with topical and oral antibiotics), as well as a widespread cutaneous rash necessitating interruption of treatment (Erfan et al., 2016) (Supplementary Figure 1). Treatment was re-introduced two weeks later at a 25% dose reduction. Repeat cross-sectional imaging two months later showed a response in all the nodal and liver lesions (Figure 2c). There was also response in the brain lesion, but a new brain metastasis within the amygdala was now evident (Figure 2b). Vemurafenib was therefore stopped and, following a three-week washout, immune checkpoint inhibitor therapy with ipilimumab was commenced. Following the second cycle, she was admitted to hospital with migraine and unsteadiness of gait and neuroimaging revealed widespread multiple brain metastases (Figure 2b). Her symptoms improved with corticosteroids and whole-brain radiotherapy. In December 2014, combination MAP kinase inhibitor therapy with dabrafenib and

trametinib was commenced. Treatment was associated with pyrexia necessitating brief interruption of dabrafenib, but subsequent resumption of the combination regimen. At the end of March 2015 she was re-admitted with a sudden-onset severe headache. Imaging revealed bleeding and perilesional oedema into two existing brain metastasis and the appearance of a further new brain metastasis. She died from progressive metastatic melanoma two months later.

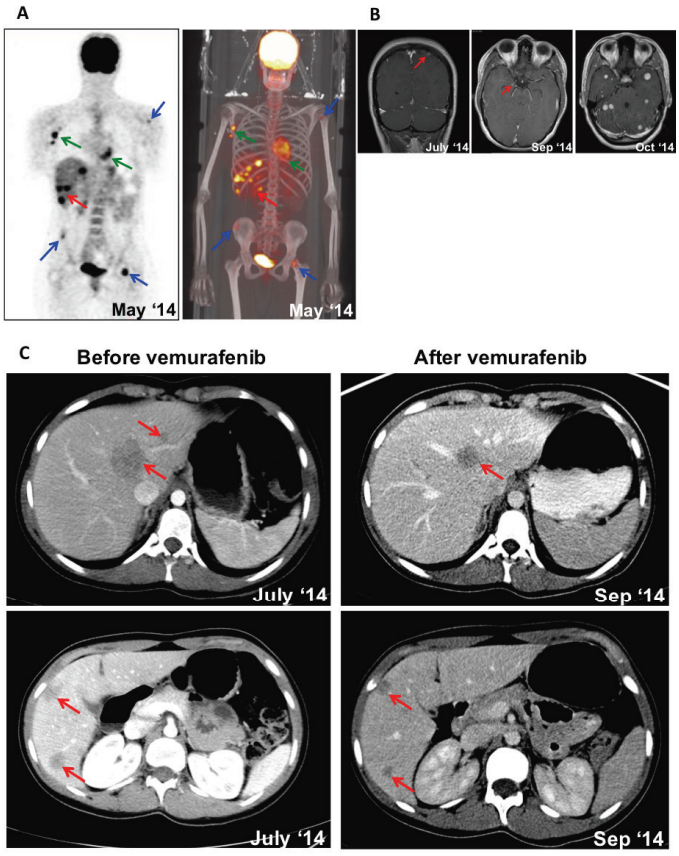


Figure 4.2: Radiological evaluation through treatment. (A) 18FFDG PETCT alongside 3D colour reconstruction. Arrows indicate avid FDG tracer uptake in the right axilla, left humeral head, left femoral neck and right iliac crest (blue) as well as widespread liver uptake (red). (B) Postcontrast T1weighted MR images showing tiny enhancing lesions in the left parietal lobe (July 2014) and right amygdala (September 2014). Axial post□contrast MR images prior to whole□brain radiotherapy showing multiple and supra□ and infratentorial lesions with no significant mass effect (October 2014). (C) Cross□sectional CT images of the liver postIV□contrast in the portal phase. Baseline images show hypodense focal lesions corresponding to segment 1 in the left hepatic lobe (left upper) and the caudal segments of the right hepatic lobe (left lower). On the right, post□treatment images indicate a partial response in all liver lesions (arrows).

Tumour Genomic Analyses

Whole genome sequencing of a cutaneous metastasis and matched germline DNA from the index patient described above revealed somatic mutations in melanoma driver genes including a BRAFV600E mutation, and a truncating CDKN2A mutation (Figure 3a). In total, we identified 133 mutations in the protein-coding region of the genome, of which 89 were protein-altering and 44 were silent (non-synonymous to silent mutation ratio = 2.022) (Supplementary Table S4 and S5). 15,853 somatic mutations were identified genome-wide with a mutation frequency of 5.12 mutations per megabase (Figure 3a and 3c). The tumour displayed a disproportionately high level of cytidine to thymidine (C>T) transitions accounting for >80% of all nucleotide changes and bore closest resemblance to the UV-exposure signature (signature 7) described by Alexandrov et al (Alexandrov et al., 2013) (cosine similarity test 0.63) (Supplementary Figure 2). We validated 42 randomly selected loci via Sanger sequencing of tumour and germline DNA and found 36 (86%) to be true somatic variants (Supplementary Table S7). A further 13 'conventional' melanomas (so-called due to their shared clinical and histological features typical of adult cutaneous melanoma) were identified from Lu et al (Lu et al., 2015). These patients had a median age of 16 years (13-20) and stage IB-IV disease. The tumours were generally from sun exposed sites (6 head and neck, 3 trunk, 3 extremities and 1 unknown) and were mainly of common histological subtypes (6 nodular, 5 superficial spreading, 1 acral and 1 unknown) (Supplementary Table S1). Pooling variants from our patient with somatic variant calls from these 13 conventional melanomas revealed a median of 10.23 mutations per megabase (3.21-52.65) (Figure 3a) (Supplementary Table S6). We obtained further mutation data of 275 adult cutaneous melanomas from The Cancer Genome Atlas (mean age 56.62 years). A Wilcoxon test comparing these to the adolescent melanoma series did not reveal any significant difference between the mutation rates of adolescent vs adult cutaneous melanoma (P value = 0.2721).

Germline Genomic Analyses

We investigated this 15-year-old patient's germline genome for known melanoma-predisposition genes including CDKN2A, CDK4, and BAP1 but failed to find any rearrangements, copy number neutral changes, point mutations or other alternations that may explain her presentation. A wider analysis of 23 additional children and adolescents, including 5 with resected primary melanoma that we whole genome sequenced for this study, and 18 children described in Lu et al (Lu et al., 2015), also failed to identify variants in established melanoma-predisposition genes. These 5 new cases had a median age of 10 years (6-16 years), were all of the superficial spreading histological subtype and had AJCC stage I disease, while the remaining 18 cases identified from Lu et al had a median age of 15 years (9 months–20 years) and included a wider spectrum of both stages and histologic subtypes (Supplementary Table S1). We noted that our patient carried R142H and V60L alleles in the melanocortin 1 receptor (MC1R) gene, contributing to her pale complexion (Garcia-Borron et al., 2014) (Supplementary Figure 1). Other MC1R variants were also discovered in the children and adolescents analysed in our study (Supplementary

Table S3).

Figure 3

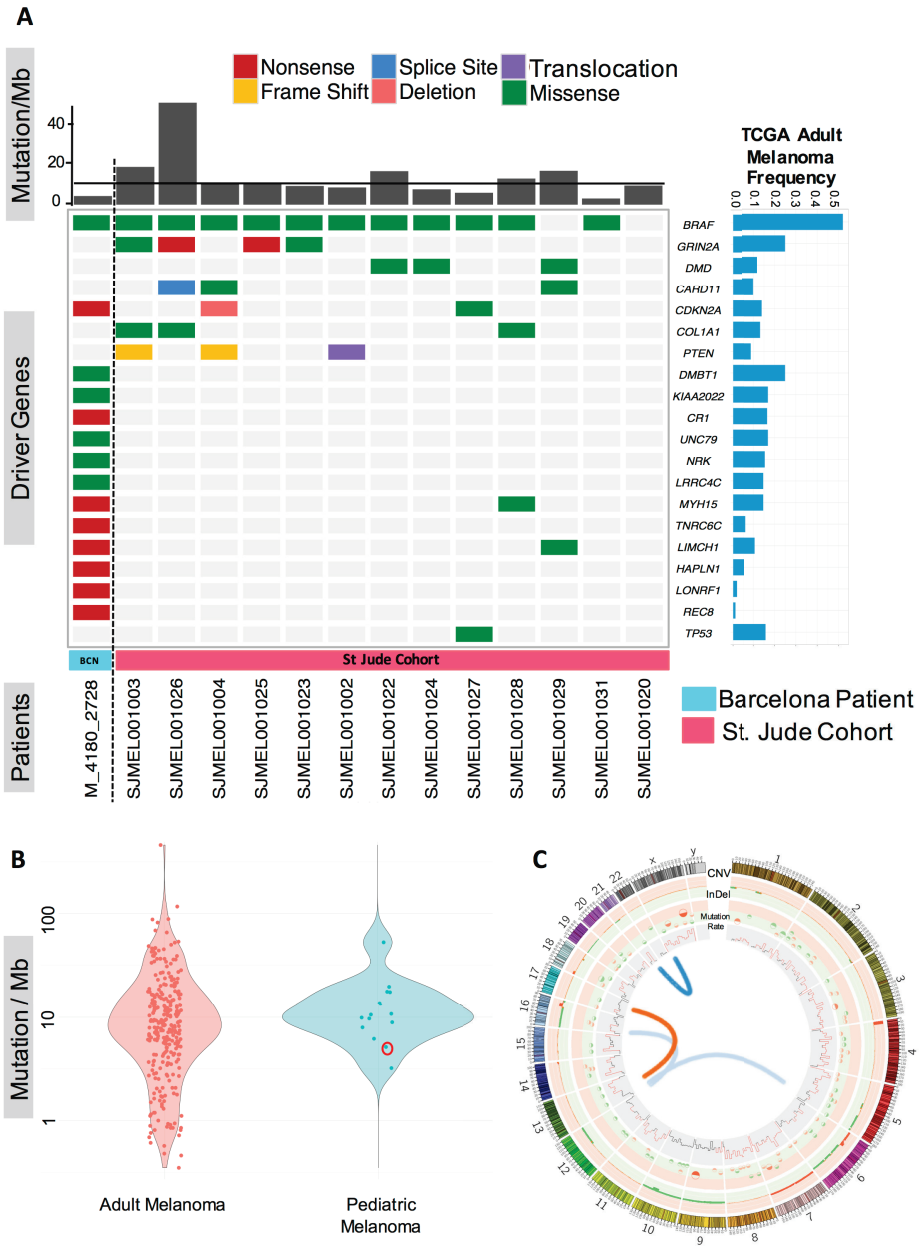


Figure 4.3: Somatic genomic analyses of adolescent melanoma. (A) Mutational landscape of adolescent melanoma. Driver mutations from the patient described are shown in the first column on the lefthand side. Remaining cases are from Lu et al. (2015) and indicate the 13 conventional adolescent melanoma patients described within this cohort and for whom genome sequencing data was available. Bar chart across the top panel shows the mutation rate per megabase (Mb) while the right panel shows the mutational frequency in adult cutaneous melanoma found in The Cancer Genome Altas (TCGA; Cancer Genome Atlas Network, 2015), straight line indicates the median number of mutations across all patients. Genes were selected based on those most frequently mutated in The Cancer Genome Atlas (adult) and in Lu et al. (childhood and adolescent; Lu et al., 2015), as well as the lossoffunction mutations detected in this 15yearold patient. A number of commonly mutated genes identified in the TCGA melanoma cohort are omitted owing to the absence of mutations of these genes in our adolescent data set (including NRAS,NF1,MAP2K1 and RB1). (B) Cluster plot of mutational frequency of adolescent versus adult cutaneous melanoma. The index patient described is circled in red. (C) Circos plot of somatic changes in the 15yearold patient described. The outermost track shows large copy number gains (red) and losses (green)(Table S8). Middle track shows small insertions and deletions (Table S9). The inner most track shows mutations per Mb (regions marked in red have mutation rates higher than 15 mutations/Mb). Interchromosomal translocations are shown in the centre and were seen in t(12 6)(q21 q2),t(12 15)(q14 q1), t(16 12)(q23 q2) and t(20 22)(q13 q32) (Table S10).

In view of the emerging evidence implicating telomere dysregulation in familial melanoma (RoblesEspinoza et al., 2014; Shi et al., 2014) we further searched for alterations in genes encoding the shelterin protein complex that protect the ends of telomeres. In particular, the protection of telomeres 1 (POT1) gene, adrenocortical dysplasia homolog (ACD) gene and telomeric repeat binding factor 2 interacting protein (TERF2IP) genes have been shown to be important in some melanoma families (Aoude et al., 2015). We found 1/24 patients carried a missense mutation in TERF2IP (allele frequency 0.00378 in The Exome Aggregation Consortium (ExAC) (Supplementary Table S3), although the pathogenicity of this mutation is unknown.

4.3. Discussion

Metastatic spread of melanoma is relatively rare amongst children, however there is data that suggests that when this occurs the prognosis is particularly poor (Strouse et al., 2005). The adolescent described in this study presented with a AJCC stage IB primary melanoma, which is associated with a 95% 5year survival (Balch et al., 2009). Despite this she developed extensive metastases three years after diagnosis and died of metastatic disease within 12 months despite three lines of modern systemic therapies known to offer potential for survival gain.

Notably, and as reported previously (Lu et al., 2015), we identified a preponderance of UV-induced mutations across 'conventional' adolescent melanomas, which was unexpected given the relatively limited exposure to UV light compared to an adult population. This 15-year-old patient had intermittent sun exposure amounting to approximately 120 hours/year, yet was always appropriately sun protected. We were unable to find germline predisposing alleles in an extended series of children and adolescents, suggesting that established high-penetrance predisposition genes do not explain most cases. However, many of these patients carried R variants of MC1R associated with red hair, freckles and pale skin (Valverde et al., 1995).

Given the variability in clinical behaviour, wide histological variation and the rarity of melanoma in infancy and early childhood, studies in this population are scarce. Consequently, our understanding of the pathogenesis in this younger cohort is more limited. Analysis of a recent large national dataset of over 350,000 melanoma patients showed that children (1-10 years) and adolescents (11-20 years) had differing survivals, suggesting inherent differences in the biology of the disease (Lorimer et al., 2016). The distinct clinical and histopathologic features of melanomas arising in a CMN and Spitzoid tumours suggest that their molecular features are likely to be very different from the 'conventional' adolescent tumours described herein (Kinsler et al., 2013; Lu et al., 2015; Shakhova et al., 2012). Additional studies on the genomic evolution of these rarer subtypes could help facilitate improved diagnostics and tailored therapies.

The reason for the observed rise in incidence of melanoma during adolescence remains unclear, however the finding of a high mutational load driven by UV exposure supports the need for education and behavioural modification as an important preventative strategy starting in early life (Green et al., 2011). The strong therapeutic effect of immune checkpoint blockade in some patients has been linked to the expression of neoantigens, mutant peptides presented by MHC Class 1. A higher overall mutational burden would be expected to lead to the expression of more neoantigens, with mutation number being associated with improved efficacy of immunotherapy (Snyder et al., 2014; Van Allen et al., 2015). This adolescent developed metastatic disease at 18 years and accessed a range of modern treatments through clinical trials. It is imperative that adolescents are given the opportunity to participate in relevant clinical trials that include novel therapies (Pappo, 2014).

4.4. Materials and Methods

Patient enrolment

We whole genome sequenced six patients as part of our study. Our first patient, whose treatment we detail, was a 15-year-old female who attended the Department of Dermatology at the University Hospital Clínic of Barcelona, Spain. Five additional children with resected primary melanoma were also identified from the University Hospital Clínic of Barcelona and from Leiden University Medical Center, Netherlands. The remaining cases were selected from a cohort of paediatric melanomas identified and sequenced at St Jude Children's Hospital, Memphis, TN (Lu et al., 2015), as part of the Paediatric Cancer Genome Project (Downing et al., 2012) (study accession through the European Genome-phenome Archive; EGAS00001000901) (Supplementary Table S1). Written informed consent was obtained from the patients' parents.

Dermoscopy, Histopathology and Imaging

Total body photography and digital dermoscopy were performed by SP using MoleMaxTM HD and DermLite® FOTO. Histopathological analyses were performed by

an expert dermatopathologist.

Sample Processing
Tumour DNA extraction from the index 15-year-old patient M_4180 was performed using the Qiagen DNA Micro Kit. Germline DNA was extracted from peripheral blood mononuclear cells using the salting out method.

Tumour Genomic Analyses
DNA from a metastatic cutaneous deposit and whole blood DNA from the index 15-year-old patient were genome sequenced on the Illumina X10 platform (Supplementary Table S2). Whole genome sequenced reads were aligned against the human reference genome (GRCh37) using the Burrows-Wheeler Aligner (Li and Durbin, 2009) (Supplementary Table S2). We used a somatic caller merging approach to identify somatic variants selecting only those detected using four or more algorithms for further analysis (Rashid et al., 2013). These calls were further filtered for germline polymorphic variants using the 1000 Genomes Project (Auton et al., 2015), other standard quality filters were also applied (e.g. depth of coverage ≥10, read mapping quality ≥15). Small insertions and deletions were identified using Scalpel (Narzisi et al., 2014). Randomly selected candidate variants were validated by capillary sequencing. Large somatic copy number aberrations were detected using the Batternberg algorithm. Somatic variants from a series of 13 'conventional' paediatric melanomas (so-called due to their shared clinical and histological features typical of adult cutaneous melanoma) described by Lu et al (Lu et al., 2015) and for whom genome sequencing data was available, were used for a comparative analysis. Exome sequencing data from a further 275 adult cutaneous melanomas was downloaded from The Cancer Genome Atlas and used for comparison with adult-onset disease.

Germline Genomic Analyses
Germline DNA from the peripheral blood of five children with resected primary melanoma were whole genome sequenced on the Illumina HiSeq2500 platform (Supplementary Table S1 and S2). These sequences, and that of the index case, were combined with whole genome and whole exome sequences from a collection of 18 children sequenced at St Jude Children's Hospital (Lu et al., 2015) (comprising 13 children from the 'conventional' melanoma cohort described above and 5 from the other histological subgroups described therein) (Supplementary Table S3). Germline variants were called using samtools mpileup (Li et al., 2009) and bcftools (Narasimhan et al., 2016). These variants were annotated for consequence using Ensembl Variant Effect Predictor (McLaren et al., 2016), and filtered for non-synonymous variants and then further restricted to those variants known to be rare (allele frequency < 10-3) by comparison to the Exome Aggregation Consortium (ExAc) dataset (Lek et al., 2016), or that were private to a single child.

M_4180_2728	EGAN00001232866	Tumour of patient M_4180
M_4180	EGAN00001195811	Germline of patient M_4180
M_509	EGAN00001197185	Germline of patient M_509
M_1064	EGAN00001197186	Germline of patient M_1064
M_3629	EGAN00001197186	Germline of patient M_3629
M_4117	EGAN00001197188	Germline of patient M_4117
$D1\_10\_02707$	EGAN00001197189	Germline of patient $D1\_10\_02707$

For details see Supplementary Table S1

4.5. Acknowledgments:

First we would like to sincerely thank this patient and her family for giving us the opportunity to undertake this research and for allowing us to share this data. We extend this immense personal appreciation to all the patients and families involved. We would also like to thank Armita Bahrami and Alberto Pappo from St Jude Children's Hospital for their helpful discussions and for sharing their experience and data.

4.6. Conflict of interest statement:

No conflicts of interest to declare. Role of the funder/sponsor: The funders did not play a role in the design of this study or in the interpretation of the results.

4.7. Supplementary materials:

List of supplementary figures

Supplementary fig 1	Cutaneous toxicities associated with vemurafenib in this patient.
Supplementary fig 2	Mutational spectra

List of supplementary figures

Table S1	Clinical summary
Table S2	QC metric of NGS data
Table S3	Germline Profile
Table S4	All Somatic Mutations [M_4180]
Table S5	Exonic Somatic Mutations [M_4180]
Table S6	Mutation Rate per Mb
Table S7	All SNV Validations
Table S8	All CNVs calls from Battenberg
Table S9	All InDel calls from Scalpel
Table S10	Translocation calls from Lumpy

Supplementary materials can be found at :
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5435926/

5

The genomic landscape of skin adnexal tumors: spiradenoma, cylindroma and their malignant counterpart spiradenocarcinoma

Spiradenoma and cylindroma are distinctive skin adnexal tumors with sweat gland differentiation and potential for malignant transformation and aggressive behaviour. Here we present the genomic analysis of 75 samples from 58 representative patients including 15 cylindromas, 17 spiradenomas, 2 cylindroma-spiradenoma hybrid tumours, 24 low- and high-grade spiradenocarcinoma cases together with morphologically benign precursor regions of these cancers. Somatic or germline alterations of the CYLD gene were found in 15/15 cylindromas, 5/16 spiradenomas but only 2/24 spiradenocarcinoma cases. Notably, we observed a recurrent missense mutation (22 tumors, 18 patients) in the kinase domain of the ALPK1 gene in spiradenoma and spiradenocarcinoma, that was mutually exclusive from mutation of CYLD and was shown to activate NFkB activity in reporter assays. In addition, high-grade spiradenocarcinomas were found to carry loss-of-function TP53 mutations, while 3/15 cylindromas had disruptive mutations in DNMT3A. Collectively, these results provide important insights into the genetic events that drive skin adnexal tumor development and reveal potentially actionable mutations.

5.1. Introduction

Spiradenoma and cylindroma are closely related benign skin adnexal tumors with sweat gland differentiation. They show histological similarities and may represent part of a morphological spectrum, further evidenced by rare spiradenomacylindroma hybrid tumors. The majority of tumors are sporadic and present as solitary nodules. Spiradenomas show a predilection for the extremities, while cylindromas commonly occur on the head and neck(Singh et al. 2013). Occasionally, they may be multiple in the setting of the BrookeSpiegler syndrome (BSS), a rare autosomaldominant inherited disorder characterized by cylindromas, spiradenomas and/or trichoepitheliomas in individuals with germline mutations of the CYLD gene (Young et al. 2006). Malignant transformation in spiradenoma (spiradenocarcinoma) and, less frequently, cylindroma (cylindrocarcinoma) is a rare event. Histologically these tumors are composed of a benign precursor and a morphologically distinct malignant component, which can be further subdivided into lowgrade and highgrade(van der Horst et al. 2015). The morphology of these tumors appears to be a good predictor of outcome. Morphologically lowgrade tumors have potential for local recurrence, while disseminated disease and diseaserelated mortality is largely limited to highgrade carcinomas (Dai et al. 2014, Granter et al. 2000, van der Horst et al. 2015, Kazakov et al. 2009). Little is known about the underlying genetic events that drive these tumors. Cylindromas are characterized by mutations in the CYLD gene and approximately two thirds of cylindromas have also been reported to carry the MYBNFIB fusion gene, which leads to overexpression of MYB, analogous to adenoid cystic carcinoma (Bignell et al. , Fehr et al. , Persson et al.). No genetic data are available for spiradenomas and the events leading to malignant transformation and to the more aggressive behavior of the highgrade tumors are largely unknown. As yet, only mutations in the TP53 gene have been reported in the malignant tumors (Kazakov et al. 2010).

To improve understanding of these rare diseases we performed a comprehensive genomic characterisation of samples from a large collection of representative patients.

5.2. Result

5.2.1. Sample ascertainment and whole exome sequencing

Samples were obtained through the University of Edinburgh Tissue Bank with ethical approval obtained under REC 15/ES/0094. Cases were independently reviewed by two dematopathologists to confirm diagnoses. In total 75 samples underwent nextgeneration sequencing, 52 with paired adjacent normal/germline DNA (from 43 patients), while the remaining 23 samples (15 patients) without matched normal/germline DNA were used as a validation cohort (Supplementary Table 1ab). Capillary sequencing was also performed on 10 additional cases to validate a hotspot mutation as described below. Thus, in total we had 68 patients, samples from 58 of whom underwent nextgeneration sequencing while 10 samples were capillary sequenced. A full breakdown of the samples used in the various stages of analysis and available clinical characteristics of each patient is provided in Supplementary

Table 1ab. Briefly, high and lowgrade spiradenocarcinoma, benign spiradenoma and dermal cylindroma patients had a median age of 68.5, 61.5, 58, 60 years at diagnosis, respectively. Notably, five patients (one cylindroma, one spiradenoma, one cylindromaspiradenoma hybrid and two highgrade spiradenocarcinoma) patients were previously diagnosed with BrookeSpiegler syndrome. Half of tumors (37/68 patients; 54%) were located on the head and neck area while the remaining cases were from the trunk (19/68 patients; 28%) or extremities (7/68 patients; 10%). The tissue sites for the remaining 8% of tumors (5/68) were unknown. FFPE cores were collected from each tumor and DNA extracted, while uninvolved adjacent skin was used to obtain normal/germline DNA where available (referred to here as adjacent normal/germline). DNA samples were wholeexome sequenced on the Agilent/Illumina platform at the Wellcome Trust Sanger Institute generating a median depth of 60x coverage (after duplicate removal and read clipping).

5.2.2. The somatic mutational landscape of adnexal tumors

DNA sequencing data from 52 tumor/germline pairs was subjected to somatic variant calling (see Methods) resulting in the identification of 1124 somatic point mutations in exons of which 817 were protein altering and 307 were silent mutations. The number of somatic single nucleotide variants (SNVs) varied markedly between individual tumor samples (mean 21.6 mutations, range 2-144) (Fig. 1, Supplementary Table 2a). In addition to SNVs, we also called 219 small insertion/deletions between 1bp to 314bp (Supplementary Table 2b). Recurrently altered cancer driver genes included CYLD (12 cases), NRAS (p.Q129E, p.Q61K in the same sample), AKT1 (p.E17K in three cases), TP53(p.E286K, p.G266E, p.R248Q) and DNMT3A (p.R556M, p.R320*, E213_splice, E585_splice)(Fig. 1). All mutations shown were validated using high-depth (median depth of coverage 117x) targeted exome capture across all samples where DNA was available (Agilent design ID: S3065404)(Supplementary Table 1). To further validate our variant calls and to determine the accuracy of our whole-exome sequence capture analysis, we used our targeted exome data to assay a further 119 randomly selected somatic variants revealing an overall validation rate of 82%. For indels the validation rate was 73%. A pan-cancer analysis revealed that in comparison to cancers sequenced by The Cancer Genome Atlas (TCGA), the tumors sequenced here have a low somatic point mutation burden in the exome and fall within the range of 0.04-2.88 mutations/Mb (Supplementary Fig. 1), a frequency similar to thyroid cancer and uveal melanoma. Generally, cylindromas were found to carry more mutations than the other tumor types (Wilcoxon test P-value 0.0153). Potential associations between the number of somatic mutations and age, sex of the patient, and tumor site were examined using a generalised linear model. No significant relationships with any individual clinical feature were observed. An overview of the genomic landscape including all available clinical characteristics for these cases can be found in Supplementary Fig. 2.

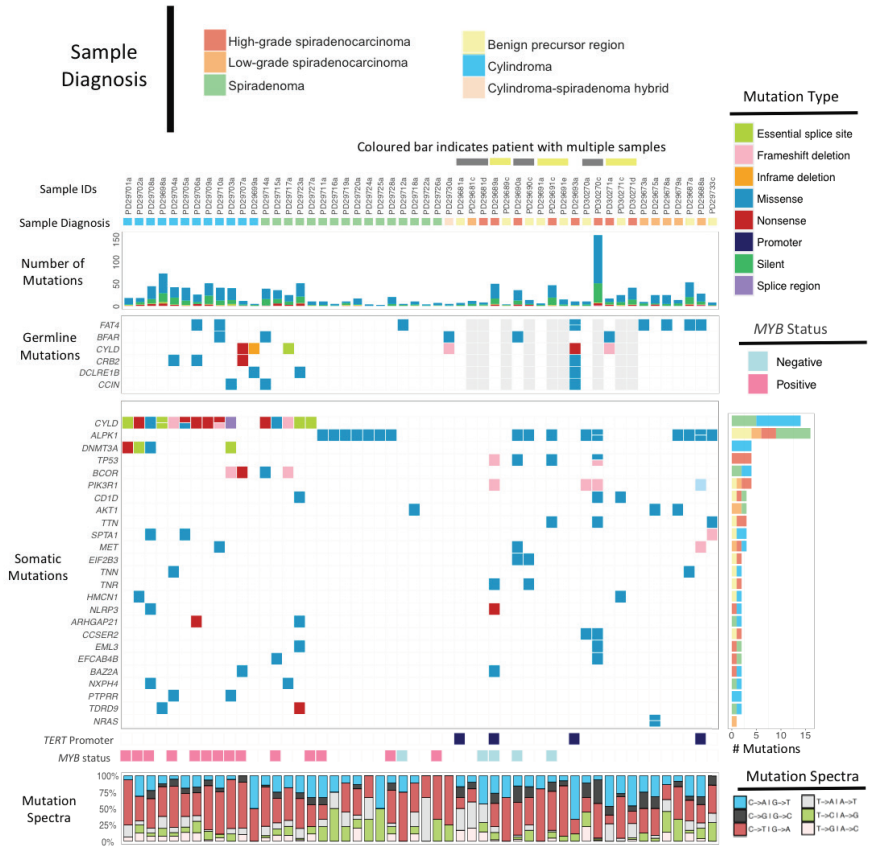


Figure 5.1: The driver gene landscape of adnexal tumours. Genetic data for the 52 cases where matched tumor/normal DNA sequencing data was available. Additional cases are shown in Supplementary Fig. 7. The germline and somatic mutations in this plot were all validated by high-depth targeted exome sequencing. Only mutations in coding regions are shown except for TERT promoter variants. Germline variants are also displayed.

5.2.3. Identification of driver genes in adnexal tumors

A typical tumor cell may contain tens to a few hundred somatic mutations distributed across hundreds of genes. Only a handful of these genes when mutated confer a selective growth advantage and thus may facilitate the promotion of tumor growth(Martincorena et al. 2018). We applied two independent driver gene discovery tools: IntOGen and dNdScv to detect potential driver genes in our adnexal tumor cohort (Gundem et al. 2010, Martincorena et al. 2017). The IntOGen driver gene prioritization framework combines scores from SIFT, PolyPhen2 (PPH2) and MutationAssessor (MA), to calculate the functional impact bias (FM bias) of mutations in genes against a background distribution (Gundem et al. 2010, Reva,

Antipin and Sander 2011, Ng and Henikoff 2003, Adzhubei et al. 2010). Using this approach genes computed to have a significant functional impact score (OncodriverFM q value) are reported as drivers. dNdScv on the other hand is a maximum likelihoodbased method used to quantify positive selection of genes mutated in cancer. We performed driver gene analyses using both of the aforementioned workflows using somatic mutations from the cylindromas, spiradenomas and highgrade and lowgrade spiradenocarcinomas. A consensus of these two approaches is reported here. CYLD and DNMT3A were identified as statistically significant driver genes in cylindroma. CYLD was also reported as a driver gene in spiradenoma, while the tumor suppressor gene TP53 was found to be significantly enriched with mutations in highgrade spiradenocarcinoma. Notably, a recurrent mutation of ALPK1 was also reported as a driver event both in spiradenoma (both methods) and spiradenocarcinoma (only by IntOGen), and is discussed in detail below. This mutation was absent from cylindromas. A complete list of the driver genes and significance values for each adnexal tumor type can be found in Supplementary Table 3.

5.2.4. Recurrent ALPK1 mutations in spiradenoma and spiradenocarcinoma

The ALPK1 (αkinase 1) gene is a member of the αkinase family and is located on chromosome 4q25 (Liao et al. 2016). Recent studies have indicated that the expression of ALPK1 during infection/inflammation can result in the activation of nuclear factorkappaB (NFκb) signalling and downstream gene expression (Ko et al. 2013, Wang et al. 2011). Somatic mutation of ALPK1 in 32/1397 lung cancer samples (2.29%) and 29/ 781 colorectal cancer samples (3.71%) has recently been reported (Liao et al. 2016).

We discovered a recurrent somatic hotspot mutation in the alpha kinase domain of the ALPK1 gene (p.V1092A) in 7/16 spiradenomas, 2/8 highgrade and 2/6 lowgrade spiradenocarcinomas (Fig. 2, Supplementary Table 1a) from our discovery cohort. All mutations were validated using targeted gene panel sequencing (see Methods). The hotspot mutation (p.V1092A) was also validated via Sanger sequencing in 8/11 samples tested. Interestingly, in several cases (5) we observed the ALPK1 p.V1092A mutation in the adjacent morphologically normal tissue (in addition to the tumour) from which the normal/germline DNA for somatic variant calling was extracted. The mutant allele fraction of the mutation in these samples was 0.32 suggesting that they are clonal or present in a significant proportion of cells. Since none of the other somatic mutations in the corresponding tumor sample were found in the sequence data from the adjacent morphologically normal tissue we can exclude the possibility of tumor to normal contamination (see Methods). This raises the possibility that the ALPK1 p.V1092A mutation is associated with a field change, as has been widely reported for other cancers, particularly skin(Martincorena et al. 2015). Interestingly, mutation of ALPK1 was mutually exclusive from mutation of CYLD (Curtius, Wright and Graham 2018, Martincorena et al. 2015) (Fig. 1). To further confirm the presence of the ALPK1 p.V1092A mutation a further 10 spiradenoma tumor/normal pairs were tested via Sanger sequencing and the p.V1092A mutation was observed in six tumors.

5.2.5. Mutation of CYLD in adnexal tumors and patients

CYLD (CYLD Lysine 63 Deubiquitinase) encodes a cytoplasmic protein with three cytoskeletalassociated proteinglycineconserved (CAPGLY) domains and functions as a deubiquitinating enzyme and tumor suppressor(Kovalenko et al. 2003). CYLD regulates the NFκb pathway, which plays important roles in cell growth and survival (Alameda et al. 2010, Sun 2010). Mutations in CYLD are associated with Brooke-Spiegler syndrome, which may present with cylindroma, cylindromatosis, trichoepithelioma and/or spiradenoma (Young et al. 2006). Eleven of the twelve cylindroma patients we sequenced carried either germline or somatic protein altering mutations of CYLD. One case (PD29703a) was found to carry a somatic splice region mutation (16_50815325_A_G) located three bases away from the splice junction. CYLD mutations were also found in 31% (5/16) of the spiradenomas (Fig. 1). All five patients with a prior BrookeSpiegler syndrome diagnosis whose germline we sequenced carried a germline CYLD mutation. The protein altering mutations in CYLD are shown in Fig. 2a.

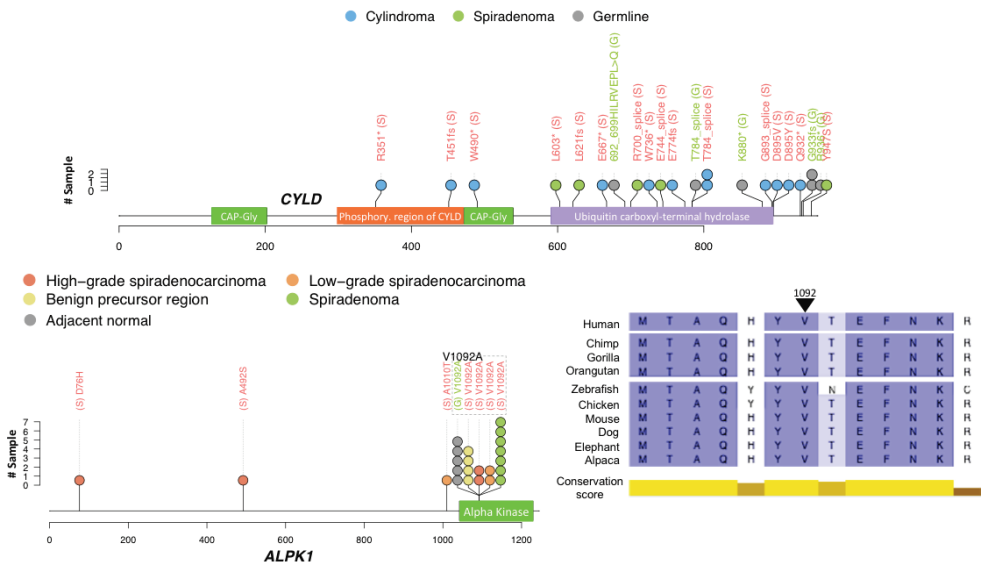


Figure 5.2: **Mutations identified in CLYD and ALPK1** Variants in CYLD (A) and ALPK1 (B) against the translation of the longest transcript of these genes. Protein domains are from UniProt. All of the variants shown were validated by highdepth targeted exome sequencing. Adjacent normal represents morphologically normal tissue from the same block as the tumor which was used as a germline sample for somatic variant calling. Variants in red were called somatically. Variants in green were called from the adjacent normal sequence. C. Protein alignment of ALPK1 across vertebrates. The conservation score represents constrained elements in multiple alignments by quantifying substitution deficits. The arrow indicates the position of the p.V1092 residue in humans.

5.2.6. Promoter and regulatory mutations

Cis and trans regulatory elements impact the transcription of many genes and mutations in these regions can potentially lead to aberrant protein production and tumorigenesis. Exome sequencing is not well equipped to detect cis regulatory element mutations as it is designed to capture proteincoding regions. However, sufficient coverage (>10x read coverage) around exon boundaries allowed us to investigate the status of proximal regulatory elements such as promoters. Detected noncoding mutations were scored for pathogenicity weighting them with a CADD (Combined Annotation Dependent Depletion) variant deleteriousness score(Mather et al. 2016) (see Methods). Mutations were also annotated in the regulatory regions of known cancer driver genes. In this way, we identified mutations in the TERT promoter region (C228T and C250T) in four spiradenocarcinomas, known hotspot positions in other cancers (Horn et al. 2013). Recurrent somatic mutations in the proximal regulatory region of SPTA1, HMCN1 were also detected (Supplementary Table 2c).

5.2.7. Mutational processes in adnexal tumors

Somatic mutations in tumor cells may be the consequence of aberrant endogenous processes such as defective DNA repair or due to exogenous factors such as exposure to carcinogens. The imprint of a mutational process on DNA sequence is commonly referred to as a mutational signature (Alexandrov et al. 2013). Analysis of mutational signatures has led to a better understanding of the underlying biological processes associated with a number of cancers and has also allowed patient stratification for therapy (NikZainal et al. 2012).

To assess the presence of published human cancer mutational signatures in the catalogue of somatic mutations from adnexal tumors we used deconstructSigs (see Methods)(Rosenthal et al. 2016). This approach computes the weighted contributions of the 30 published COSMIC signatures and one additional unknown signature to the mutational catalogue of each sample. The heatmap in Fig. 3a represents the contribution of these signature across all adnexal tumor subtypes. In more than a quarter (26.92%) of tumors the contribution of signature 1 was greater than 0.5. Signature 1, an endogenous mutational process associated with spontaneous deamination of 5methylcytosine, which is often correlated with age(Alexandrov et al. 2015). The mutation catalogue from cylindromas was also enriched for signature 7, which is predominantly found is skin cancers as a result of ultra violet (UV) light exposure. The predilection of cylindromas to form on the head and neck is likely to explain this signal. We also performed an analysis combining mutations for each tumour type together and again identified signature 1 and signature 7 in cylindromas, while lowgrade spiroadenocarcinomas where enriched for signature 26, which is thought to be associated with DNA mismatch repair (Alexandrov et al. 2013).

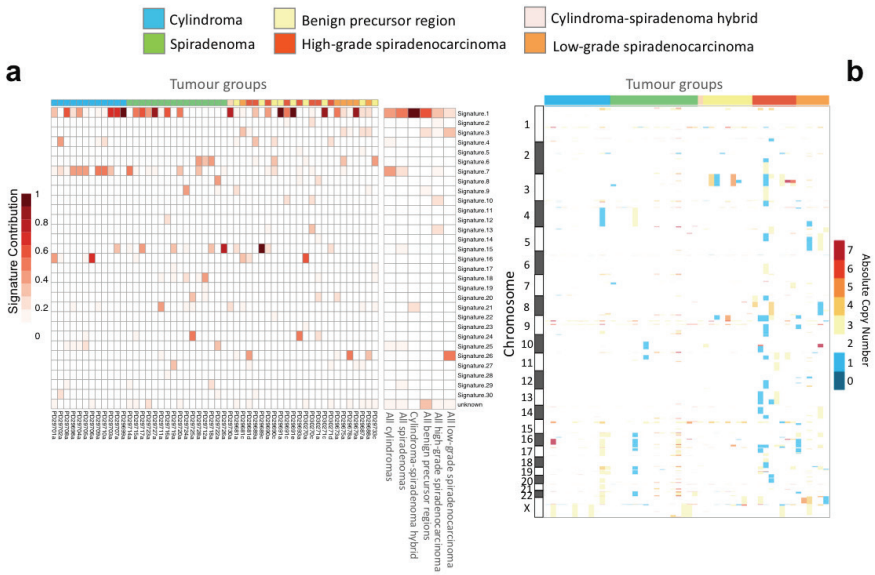


Figure 5.3: The somatic genetic landscape of adnexal tumors. A). Contribution of published mutational signatures in adnexal tumors detected using deconstructSigs (Rosenthal et al. 2016). Total contribution per sample adds up to one. For this analysis we used all variants including those in noncoding regions such as 5′ and 3′ UTRs. B). The copy number landscape of adnexal tumors. This analysis was performed using Sequenza to define the absolute copy number for chromosomal segments. These analyses were performed using the tumors shown in Fig. 1

5.2.8. Somatic DNA copy number alterations

The copy number status of our adnexal samples was assessed using Sequenza, an allele specific copy number analysis algorithm that uses matched tumornormal pairs (Favero et al. 2015). Sequenza reported a total of 1577 somatic copy number changes (1350 gains and 227 losses) across 52 tumors. Several highgrade spiradenocarcinomas showed large copy number changes, while lowgrade spiradenocarcinomas demonstrated a comparably lower number of copy number events. Cylindromas and spiradenomas generally showed few copy number changes, as did morphologically benign precursor regions of highgrade spiradenocarcinomas. Genomewide copy number profiles across all subtypes are reported in Fig. 3b.

5.2.9. The MYB-NFIB fusion in adnexal tumorigenesis

Previous reports have suggested a role for MYBNFIB fusions in the pathogenesis of both adenoid cystic carcinoma and cylindroma (Persson et al. 2009). Using multicolour FISH we analysed 21 cases including 13 cylindromas, 7 spiradenomas and 1 cylindromaspiradenoma hybrid tumor in addition to an adenoid cystic carcinoma case known to carry the MYBNFIB fusion as a control case. This analysis revealed that, despite previous reports, none of the cylindromas were found to carry the

fusion event (Fehr et al. 2011). The MYBNFIB fusion was also absent from the spiradenoma and cylindromaspiradenoma hybrid tumor (Fig. 4a, Supplementary Figs. 3 & 4). Overexpression of MYB was, however, confirmed in cylindroma and spiradenoma cases using immunohistochemistry (Fig. 1 & Supplementary Table 4) suggesting other mechanisms of gene overexpression are operative.

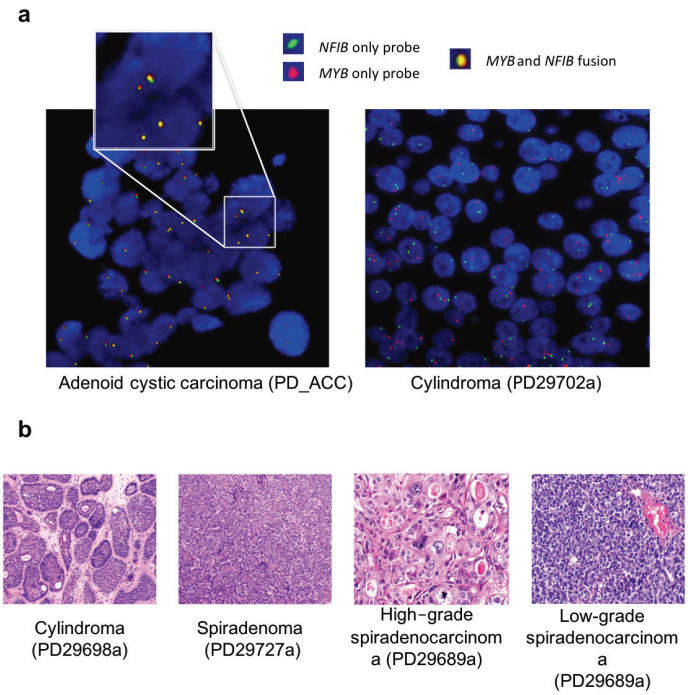


Figure 5.4: Assessment of the MYBNFIB fusion in adnexal tumors. A). Fluorescence in situ hybridization (FISH) imaging of the MYBNFIB fusion in an Adenoid cystic carcinoma and assessment in cylindroma samples. Previous reports have suggested that adnexal tumors such as cylindomas carry MYBNFIB fusions which have been associated with MYB overexpression (Fehr et al. 2011). Left panel shows an adenoid cystic carcinoma, which is positive control for the fusion event. Yellow signal is a blend of green NFIB probe and red MYB probe. Right panel: a representative cylindroma which was fusion negative. B). Representative histopathological images of a cylindroma at 100x magnification, spiradenoma at 100x magnification, highgrade spiradenocarcinoma at 200x magnification and a lowgrade spiradenocarcinoma 200x magnification.

5.2.10. Analysis of the benign precursor component of high-grade spiradenocarcinoma

To further understand the genesis of skin adnexal tumors we sequenced benign precursor regions immediately adjacent to malignant regions of highgrade spiradenocarcinomas (Fig. 1). Malignant samples demonstrated a significantly higher mutation rate compared to their benign counterparts (Wilcoxon test pvalue = 0.0055).

There was little overlap in somatic mutation calls between benign precursor regions and their highgrade spiradenocarcinomas (Supplementary Figs. 5 and 6). Two benign samples (PD30270a, PD29690c) shared the ALPK1 p.V1092A mutations with their highgrade spiradenocarcinomas. Collectively, these findings suggest that the malignant component of these tumors often arises from within a heterogeneous pool of cells.

5.2.11. Germline analysis of adnexal tumor patients

As mentioned above, we identified germline CYLD mutations in all five patients previously diagnosed with BrookeSpiegler syndrome. Germline CYLD mutations were also detected in two additional patients with no prior BrookeSpiegler syndrome diagnosis (Fig. 1). To extend the analysis of germline variation in patients from our cohort we used samtools mpileup and the bcftools variant genotyping strategy (Li et al. 2009). We assessed the mutation burden per gene using a Fisher's exact test (see Methods) (Supplementary Table 2de) (Lek et al. 2016) using variants from the 43 cases where adjacent normal/germline exome sequence had been generated. From this analysis CYLD was found to carry significantly more deleterious mutations than expected (BenjaminiHochberg (BH) adjusted pvalue 3.53e4), reconfirming its wellestablished role as an adnexal tumor predisposition gene. We also detected a significantly high number of deleterious mutations in BFAR and FAT4 (BH adjusted pvalue 6.96e4). FAT4 is a member of human FAT gene family which encodes a large transmembrane protein consisting of multiple extracellular cadherin domains and a cytoplasmic domain that can interact with signalling molecules (Katoh 2012). This gene is homologous to fat in Drosophila, a known tumor suppressor gene (Mahoney et al. 1991). It should be noted, however, that FAT4 has been reported as disease associated in several studies, which might suggest a high rate of polymorphism (Cappello et al. 2013, Ivanovski et al. 2018, Sebio et al. 2016). BFAR, the bifunctional apoptosis regulator, plays a role in the regulation of cell death and in this way could contribute to tumorigenesis (Roth et al. 2003). Notably for both FAT4 and BFAR we did not identify somatic mutations in the cases carry germline mutations in these genes suggesting that if they are contributing to tumour formation they probably don't function as classical tumour suppressors. Further, several samples with germline FAT4 and BFAR variants also had germline or somatic lossoffunction alleles of CYLD, making these a less likely candidates. We next asked if mutations of known pathogenicity were found in the germline of any of our adnexal cases. In this way we found 14 pathogenic or likely pathogenic variants including variants in PTEN and NSD1 (Supplementary Table 2f) (ClinVar database (dbSNP build 144). Thus, of the 43 adnexal patients analysed here we have shown that seven patients carry germline mutations in CYLD and propose several other candidate genes as mediators of germline susceptibility for followup studies.

5.2.12. Analysis of tumors without matched germline DNA

For 52 of the samples in our cohort we had matched tumor/adjacent normalgermline pairs (as described above). Matched germline DNA was not available for a further 23 samples (15 patients; three cylindromas, one spiradenoma, one cylindroma-

spiradenoma hybrid, three lowgrade spiradenocarcinomas, seven highgrade spiradenocarcinomas) and thus we used the tumor sequences from these cases as a validation cohort to look for variants in genes identified from the abovementioned analyses. We first called variants against an unmatched normal sample (Supplementary Table 1b) and then filtered these data using variants in the ExAC database(Lek et al. 2016) (Allele frequency > 0.0001) and from an inhouse panel of 100 normal germline exomes. We next focused on genes identified from our analysis of the discovery cohort (see Methods) revealing ALPK1 p.V1092A mutations in one cylindromaspiradenoma hybrid, two lowgrade spiradenocarcinomas and one spiradenoma. Lossoffunction mutations were also detected in CYLD in three cylindroma cases (PD29695, PD29696, PD29700) and in one lowgrade spiradenocarcinoma (PD29676a). Two highgrade spiradenocarcinoma patients were found to carry frameshift deletions in TP53 (p.P191fs*54 and p.T329fs*8). For each patient, the respective changes were present in all collected tumor samples indicating these changes maybe of germline in origin or occur early in tumor development. An overview of the driver gene landscape and clinical characteristics of all 75 tumors/samples can be found in Supplementary Fig. 7.

5.2.13. Functional studies of the ALPK1 p.V1092A variant

Given the role of ALPK1 in the regulation of the NFkB pathway in infection we next asked if the ALPK1 p.V1092A variant we identified could activate NFkB signalling and thus substitute for mutation of CLYD. To do this we generated fulllength wildtype and ALPK1 p.V1092A mutant cDNA constructs and transfected them together with a NFkB luciferase reporter construct into the cell lines MCF7 and T47D and HCT8R [Supplementary Fig. 9]. Analysis in this way showed that the mutant construct activated NFkB reporter activity considerably higher than wildtype construct in a range of epithelial cell lines consistent with a role in driving tumour growth akin to mutation of CYLD.

5.3. Discussion

The analysis of adnexal tumors in this study yielded several remarkable results. Firstly, we identified a recurrent somatic missense ALPK1 mutation (p.V1092A) in the kinase domain of this alphakinase and demonstrated that this mutation activates NFκb signalling in cell reporter systems. Since kinases can be readily inhibited this mutation represents a potential therapeutic target, which might be particularly advantageous in the metastatic setting, where it could be targeted to control tumor growth. Secondly, we find new driver genes not previously associated with adnexal tumors. For example, statistical analyses reveal significant enrichment of mutations in DNMT3A in cylindromas, a gene previously linked to haematopoietic malignancies, where it plays a role in the regulation of methylation (Feng et al. 2010, Guillamot, Cimmino and Aifantis 2016). Mutations in genes such as AKT1, BCOR and PIK3R1 were also observed and these genes may also contribute to tumor development. In keeping with previous studies, we found frequent mutation of the CYLD gene. Somatic or germline CYLD mutations were found in 12/12 cylin-

droma patients with mutations also being observed in spiradenoma and highgrade spiradenocarcinoma cases. Notably, these mutations were mutually exclusive from the abovementioned ALPK1 variant. As the aetiology of adnexal tumors is unknown we performed a mutational signatures analysis. This revealed the presence of signature 1 across all tumour types, which is ageassociated, but also the UVassociated signature 7 in cylindromas, presumably because these tumors are generally found on the head and neck. There was also some suggestion of signature 26, associated with mismatch repair, in lowgrade spiradenocarcinomas but no other recurrent signature was observed. Tumors in our adnexal collection were not only low in terms of their somatic mutation burden but also appeared to lack significant copy number alterations, the exception being highgrade spiradenocarcinomas which, compared to other adnexal tumors, were replete with copy number gains. Finally, we identified germline variants in CYLD that have not been described previously, and thus represent new pathogenic alleles, we also found cases with pathogenic variants in the ClinVar database including in PTEN and NSD1, suggesting potential adnexal tumor predisposition alleles. The identification of a patient with an NSD1 mutation, which is associated with Sotos syndrome, is of particular interest because previous case reports suggest adnexal tumours in some patients. These insights need to be explored in larger series (Gilaberte et al. 2008).

In summary, our paper reports the most comprehensive picture of the genomic landscape of adnexal tumors including driver genes, copy number alterations and a potentially actionable kinase mutation and mutational signatures. We hope these studies will help inform the management of patients with these malignancies.

5.4. Materials and Methods

5.4.1. Patients and samples

Samples for whole exome sequencing (WES) and targeted gene panel sequencing (TGPS) were collected from 58 patients and divided in to discovery (tumor/adjacent normalgermline pairs) and a validation cohort (tumor only). The discovery cohort contained 52 tumors and matched adjacent normal/germline DNA from 43 patients. This cohort was used for the initial genomic profiling and driver gene analyses. Mutations from 23 additional samples (15 patients) from the validation cohort were also reported. A detailed description of each case/sample can be found in Supplementary Table 1ab. All diagnoses were confirmed by two independent pathologists. Ethical approval was obtained from the West Lothian Tissue bank.

5.4.2. Wholeexome sequencing

Exonic DNA was captured using the Agilent wholeexome capture kit (SureSelect All Exon V5). Captured material was indexed and sequenced on the Illumina Hiseq2500 platform at the Wellcome Trust Sanger Institute to a median depth of 60x. Raw 75 bp pairend sequencing reads were aligned with BWA (v0.7.12) to the GRCh37 human reference genome producing a single Binary Alignment Mapping (BAM) file for each sample (Li and Durbin 2009). Duplicated reads resulting from PCR were

marked with BioBamBam (v2.0.54) (Tischler G 2014, Li et al. 2009, Li and Durbin 2009).

5.4.3. Targeted gene panel resequencing

To confirm our findings from whole exome sequencing we validated mutations in the top recurrently mutated genes using panel sequencing (Supplementary Table 5). Genomic regions for 550 genes were captured using Agilent custom pulldown baits. Captured material was indexed and sequenced on the Illumina Hiseq4000 platform to a median depth of 117x. Raw 75 bp pairend sequencing reads were processed using the same pipeline as used for wholeexome sequencing described above.

5.4.4. Somatic variant detection

Somatic variants were detected using CaVEMan, an expectation maximization–based somatic substitution detection algorithm (David Jones 2016). Candidate somatic variants were then filtered for quality and to remove common population variants (ExAC allele fraction > 1e04). Small insertion and deletion (indel) detection was performed using the cgppindel pipeline (v0.2.4w) (Ye et al. 2009). Detected indels were then filtered for quality, sequence coverage in both tumour and normal, strand bias and for overlap with known simple repeats or indels in the inhouse normal panel.

5.4.5. Germline mutation burden analysis

We applied an exomewide Fisher's exact test to assess the significance of observing n mutations in gene X in our 43 germline samples, given gene X has a mutation rate of Y in a control population. To select an appropriate control population, we performed a principal component analysis using 2504 individuals across multiple populations from the 1000 Genomes Project phase3(Abecasis et al. 2010). We randomly selected 2000 single nucleotide polymorphic variants (SNPs) and to mitigate the impact of population specific rare variants we only selected SNPs with a population allele frequency between 0.1 and 0.7. PCA analysis revealed that all 43 patients with tumorgermline pairs were of European descent (Supplementary Fig. 12). Therefore, benign polymorphic variants from the ExAC database from individuals of nonFinnish European descents were used as a negative control. To eliminate the impact of confounding variants, a Combined Annotation Dependent Depletion (CADD) score filter was applied and only variants with a CADD score above or equal to 15 were taken forward for the burden test. We also ensured that only variants that have a minimal sequence coverage of 10x in both case and control data sets were used. Finally, we applied a Fisher's exact test on every gene to estimate the likelihood of observing n deleterious mutations given the background mutation rate of that gene in the control population. The BenjaminiHochberg method was used to correct for multiple testing and only genes with an adjusted pvalue less or equal to 0.01 were reported as significant.

5.4.6. Variant quality control for FFPE artefact

Formalin fixation of tumor biopsies can have a detrimental impact on DNA and introduce C>T/G>A sequencing artefacts, which are primarily found at low allelic fractions (Wong et al. 2014). These artefacts are more frequently observed at a 0.010.10 mutant allele fraction (MAF) (Wong et al. 2014). To remove these variants, we used the following filters:

- Tumor read depth (TRD) and adjacent normal/germline read depth (NRD) greater than or equal to 10.
- Mutation with MAF <= 0.10 is kept only if TRD and NRD is greater than equal to 30.
- Mutation with MAF <= 0.05 is kept only if TRD is greater than or equal to 100.

After filtering our somatic point mutation validation rate from the whole exome sequencing data was 82% as confirmed by targeted sequencing.

5.4.7. Mutational signatures analysis

To alleviate the impact of artefacts from 5methycytocine deamination and degradation in our FFPE samples, low allelic fraction mutations (mutant allele fraction below 0.10 and read depth below 10) were removed from the signature delineation process (as outlined above). Somatic point mutations were then mapped to the 96 possible trinucleotide contexts taking into account the probability of each mutation occurring in each trinucleotide within the human genome. We then applied deconstructSigs, a multiple linear regressionbased algorithm to reconstruct the mutation profile of each tumor sample using a linear combination of predefined mutational signatures (Rosenthal et al. 2016). Thirty human cancer signatures as defined in Alexandrov et.al, were used for the reconstruction and one "unknown" signature (Alexandrov et al. 2013).

5.4.8. DNA copy number analysis

To estimate allelespecific copy number profiles we used the Sequenza software package (v2.1.0), a probabilistic modelbased algorithm applied to segmented average depth ratio (tumor versus normal) and B allele frequency(Favero et al. 2015). Preprocessing and analysis with Sequenza were performed as described in the Sequenza documentation and fitted models were manually examined. For four tumournormal pairs default fitted model suggested very high ploidy. However, after manual inspection of the depth ratio and BAllele fraction data an alternative solution closer to ploidy 2 was adapted due to lack of evidence for high ploidy.

5.4.9. Gene fusion analysis

Fusion gene analysis was performed using the MYBNFIB fusion/translocation FISH probe kit from CytoTest. The MYB 5′ probe covers the entire MYB gene along with upstream (5′) and some downstream (3′) genomic sequences. The NFIB 3′ probe

covers the 3′ (end) portion of the NFIB gene along with some adjacent genomic sequence. An adenoid cystic carcinoma (PD_ACC) case known to carry the fusion was used as a positive control.

5.4.10. ALPK1 hotspot validation using Sanger sequencing

DNA was extracted as above for exome sequencing. The region of interest of ALPK1 was amplified using ThermoFisher Platinum HiFi Taq DNA polymerase (following manufacturer's instructions) using the oligos shown below. Amplified products were sequenced by Sanger Sequencing (Eurofins) using the same oligos. Sequence traces were analysed by visual inspection.

ALPK1 Forward: 5′ TTGATCTCCTCTCTCTTACTCCA 3′
ALPK1 Reverse: 5′ ATGCTAGCCTGATTATGTGGAA 3′

5.4.11. Functional analysis of ALPK1 mutation

HCT8R, MCF7 and T47D cells were maintained in DMEM supplemented with 10% foetal calf serum and 2% glutamine and cultured at 370C, 5% CO2. NFκB transcriptional activity was determined using a pNFκBluciferase reporter containing four tandem copies of the κ enhancer (κB4) site in a pUC vector. Fulllength ALPK1 cDNAs (wildtype and mutant) were synthesized by GeneArt and cloned into the pcDNA3.1 expression vector. Briefly, cells were seeded on 24 well plates in such a density as to obtain a confluent monolayer. After 24 h the medium was removed and cells were transfected with Lipofectamine 2000 (Invitrogen, CA, USA) in OPTIMEM (Gibco, NY USA) following the recommendations of the manufacturer. Transfection efficiency was normalized by cotransfection with a TKrenilla luciferase plasmid (Promega, Madison, USA) together with the NFκB reporter and the wild type or mutant ALPK1 expression vectors. Cells were transfected overnight and the dual luciferase reporter assay kit from Promega (Madison, WI USA) was used to measure transcriptional activity in a FLUOstar Omega luminometer (BMG Labtech, Aylesbury UK).

5.4.12. MYB expression by immunohistochemistry

MYB overexpression in cylindromas has been reported in several earlier works(Rajan et al. 2016). We attempted to assessed MYB expression status in 29 samples (11 cylindromas, 6 spiradenomas and 12 highgrade spiradenocarcinomas) using immunohistochemistry (IHC) (Supplementary table 4). IHC was performed on 4-µmthick formalinfixed paraffinembedded wholetissue sections following antigen retrieval with Target Retrieval solution (pH 6.1; Dako, Carpinteria, CA, USA) in a pressure cooker using a rabbit monoclonal antiMYB monoclonal antibody (1:200 dilution; clone EP769Y; Abcam, Cambridge, MA, USA) and the Envision+ polymer detection system (Dako).

5.5. Acknowledgements

This work was supported by Cancer Research UK, the Wellcome Trust and by the ERC Combat Cancer Project. We also wish to thank Dr. J.W.R. Meijer, Rijnstate Hospital, Arnhem, the Netherlands.

5.6. Supplementary materials:

Supplementary materials can be found at : https://www.nature.com/articles/s41467-019-09979-0Sec31

6

Predicting cancer driving mutations in the non-coding genome

Mamunur Rashid, David J. Tax and Jeroen De Ridder

Distinguishing the driver mutations from the passenger ones is one of the most important challenges in cancer research and the research landscape was dominated by mutations in the protein-coding genome. Driver mutations in the noncoding genome offer an orthogonal perspective to cancer development. Despite recent advancements in the field, we identified several areas with a considerable scope of improvements such as dealing with the absence of gold standard mutations and striking imbalance in the number of driver and passenger mutations. We proposed a variable genomic window-based enrichment test approach to identify a reasonable substitution for putative, gold-standard cancer drivers (positive class). Finally, we developed an asymmetric loss function based random forest classifier to tackle serious class imbalance problem posed by a huge number of passenger (negative class) mutations. A model trained and validated on a pan-cancer data-set identified several novel as well as previously reported cancer-driving mutations.

6.1. Background

6.1.1. Driver mutations in cancer

'Driver' mutations are somatically acquired nucleotide changes that confer a selective growth and survival advantage to the tumour cells (Stratton et al., 2009; Pleasance et al., 2009). Somatic mutations are the consequence of faulty endogenous process in an aging human cell such as defective DNA repair or exposures to carcinogens such as the by-products of smoking or ultraviolet light. An aging cell accumulates these mutations and their numbers vary from few hundreds to hundreds of thousands. Only a handful are considered driver mutations because of their ability to driver tumourgenesis, while the vast majority of these mutations are classified as 'passenger'.

Distinguishing between drivers and passenger mutations is one of the most important challenges in cancer research, as this could unlock the potential to tailor therapeutic interventions based on the a patient's own tumour DNA sequence (Pleasance et al., 2009). A key reason for why this has proven to be such a challenge is that the identification of driver events relies on signals of positive selection in the genome, i.e. the recurrence of mutations across more independent tumours than expected by chance. Problematically, such "mutation mountains" are inherently scarce due to limited statistical power (Vogelstein et al., 2013). This is because, even with current sequencing capabilities, sample sizes of homogeneous tumour populations are small while the number of ways in which genomic variation can affect downstream cellular pathways is far greater. In fact, it has been found that many "mutation hills", i.e. recurrent but infrequent mutations, point to bona fide driver events, but fail to exceed the noise level or background mutation rates.

A wide range of computational approaches have been proposed to overcome this limitation. For instance, burden tests aim at improving statistical power by merging genomically dispersed events at the gene level. Conceptually, this results in the merging of several mutation hills into a single mutation mountain, which should have more chance of reaching the significance threshold to be called a driver mutations. Alternatively, it has been found that the local distribution of driver mutations within the gene body appears to be non-random and can thus be leveraged for identifying driver genes. For instance, tumour-suppressor genes can be functionally silenced by truncating mutations throughout the gene body, whereas oncogenes are often affected by a limited number of mutations at specific amino acid positions (Vogelstein et al., 2013). A number of methods have also taken the functional impact of variants in to account such as those predicted by SIFT and Polyphen under consideration while assessing tumour driving potential of genes (Ng and Henikoff, 2003; Adzhubei et al., 2013; Gonzalez-Perez et al., 2013; Gonzalez-Perez and Lopez-Bigas 2012). Finally, burden tests at the pathway level, or those that explicitly considering patterns of mutual exclusivity, are also employed, either by performing enrichment tests at the gene-set level or using more advanced network smoothing approaches (e.g. Hotnet) (Leiserson et al., 2014).

6.1.2. Non-coding driver mutations in cancer

Large-scale regulatory and epigenome characterisation projects (e.g. ENCODE) claim that as much as 40% of the human genome is estimated to carry regulatory elements such as transcription factor binding sites (TFBS), promoters, enhancers, silencers, insulators and Topologically Associated Domains (TAD) boundary elements [Fig. 1a]. The mutational load on these elements in a cancer genome are akin to that of the coding region [Fig. 1b-c] (Gonzalez-Perez et al 2012). Not surprisingly, evidence is mounting that non-coding mutations have ample opportunity to confer a selective growth advantages to tumour cells and should thus be considered as potential novel cancer driving events (Horn S et al 2013, Weinhold N et al 2014, Puente XS 2015). Indeed, several studies have shown strong enrichment of such mutations in a number of diseases and several recent studies have report recurrent mutations in the UTR and promoter region of NOTCH1 and TERT in different cancer types (Horn S et al., 2013, Puente XS et al., 2015, Epstein D et al., 2009, Vinagre J et al., 2013).

However, current computational approaches that rely on collapsing variants at the gene and/or pathway level, gene-centred mutation effect prediction method or inspection of within-gene distribution of variants are no longer effective in distinguishing drivers from passengers for non-coding variants (Martincorena I et al., 2017, Lawrence M et al., 2013, Porta-Pardo, E. et al., 2017).

6.1.3. Non-coding driver mutation prioritization

In recent years, a wide range of computational approaches have been developed that aim to distinguish non-coding driver mutations from benign non-coding passenger mutations (18-26). These tools typically leverage the wealth of genomic and epigenomic data generated by comprehensive epigenome profiling studies such as the ENCODE and the Roadmap Epigenomics project to provide a rich characterization of the genomic contexts in which mutations occur mutations in the non-coding genome. An initial, annotation step generally produces a $N \times M$ data matrix, where N is the total number of non-coding mutations for a set of cancer samples and M represents the number of regulatory features, such as overlap with promoter or enhancer elements or TFBSs and epigenomic markers in the genome (e.g. histone modification). This annotation step is then followed by a prioritization step that aims to rank variants based on their potential to regulate tumour formation and progression. This step consists of either (i) a rule-based scoring strategy e.g. FunSeq2 (Khurana E et al 2013) SuRFing (Ryan MN et al., 2014) (ii) unsupervised learning e.g. GenoCanyon (Lu et al 2015) or (iii) supervised learning e.g. GWAVA (Ritchie GRS et al., 2014), CADD (Kircher M et al., 2014), DANN (Quang D et al., 2015), FATHMM (Shihab HA et al., 2015), DeepSEA (Zhou J et.al. 2015).

These studies have provided the initial path for non-coding mutation prioritisation by demonstrating that properties of driver mutations can indeed be learned amid millions of passenger ones using a comprehensive source of annotation. For example, Combined Annotation–Dependent Depletion (CADD) trained it's Support Vector Machine (SVM) model on 29.4 million SNVs (simulated mutations and ob-

served SNPs in human genome) annotated with 63 distinct features. The trained model was then used to predict deleterious capability of 8.6 billion possible substitutions in the human genome. FATHMM and GWAVA used curated heritable germ-line mutations from the Human Gene Mutation Database (HGMD) (Stenson PD et al., 2014) and SNPs to train a Random forest and multiple kernel learning algorithm respectively and predicted using a cross validation approach. DeepSEA, a convolutional neural network based framework, learns the regulatory sequence pattern from large scale chromatin-profiling data (e.g. the Epigenome Roadmap). Learned regulatory sequence pattern then enables the framework to predict the effects of altered sequence on chromatin with single-nucleotide sensitivity

These methods collectively gathered a large compendium of curated annotation sources and made significant contributions to our understanding of the role of non-coding mutations in human diseases in general. An overview of these tools can be found in Additional file 2: Table S1.

Nonetheless in several areas - particularly in prioritizing non-coding variants there remains considerable scope for improvement, which if explored will further enhance our understanding of regulatory mutation prioritization. Here we will briefly discuss a few of these areas and explored potential avenues for improvement.

A well-characterized positive and negative set is essential for supervised learning and creating reliable predictions. While there are many examples of validated cancer-causing mutations and genes in the protein coding region there are only handful examples in the non-coding genome. Several computational approaches have been adopted to circumvent this problem and they can be further improved by taking in to account various biological characteristics of known driver mutations.

Annotations used by many existing tools often contain a mixture of coding and non-coding features, where the feature set is dominated by protein coding features (Kircher M et al., 2014, Quang D et al., 2015). As a result, they often make excellent predictions for mutations in the protein coding regions but perform poorly in prioritizing non-coding mutations (Kim K et al., 2016.)

Germline and somatic variants both play critical roles in cancer development and progression but their impact is manifested through two distinct paths (Milholland B et al., 2017, Vogelstein B 2013). Distinguishing driver somatic mutations from the benign ones poses a fundamentally different problem from that of germline mutations. Several existing tools have shown acceptable predictive performance discriminating germline pathogenic variations (ClinVar, HGMD) from benign ones but their performance on prioritizing cancer driving non-coding mutations remains inadequate (Stenson PD et al., 2014, Landrum MJ et al., 2014) [Additional file 1: Figure S1].

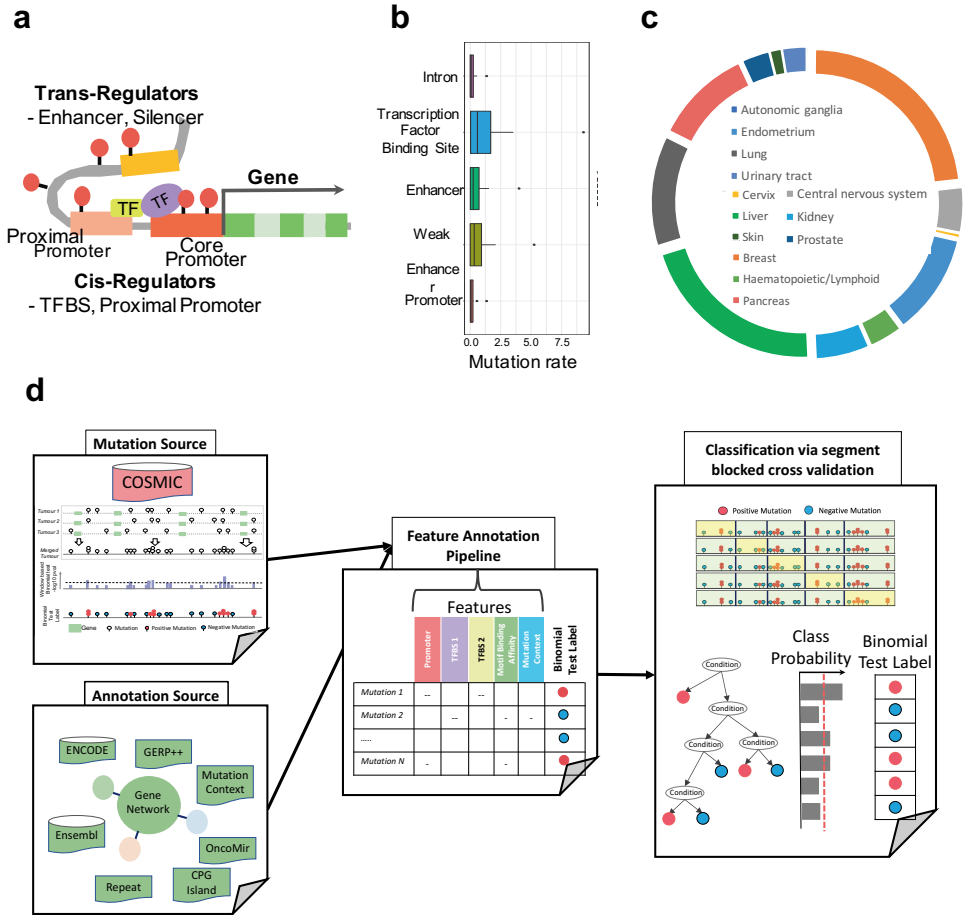


Figure 6.1: (a) Cartoon diagram demonstrating the non-coding elements of the human genome and non-coding mutations (b) Doughnut chart showing proportion of tumour samples from various tissue of origin (data source: COSMIC V70) (c) Boxplot showing mutation rate in various regulatory elements of human genome (data source: COSMIC V70) (d) Overview of our regulatory mutation prediction workflow.

Due to the scarcity of known, validated regulatory mutations and the abundance of passenger mutations, any prioritization tool aiming to distinguish between these two classes faces a serious class imbalance challenge. Under-sampling of the major class and oversampling of the minority class have previously been shown to offer some improvement however there are still ample opportunities to improve in this area. (Ramezankhani A et al., 2016, Chawla N et al., 2002).

In this paper, we propose a burden test based strategy to address the absence of gold standard positive non-coding driver mutations problem and explore a machine-

learning based approaches for prioritizing regulatory mutations. To this end, we redefine the potential positive (non-coding driver) and negative (non-coding passenger) mutation labels by harnessing the power of local mutation density, which provides a good approximation of potential regulatory mutations in the absence of functionally validated ones [Fig. 1]. Relabelled positive and negative mutations were annotated for 292 regulatory, genomic and epigenomic features using a custom feature annotation pipeline [Additional file 2: Table S2]. We show that for an imbalance classification problem such as this, a class dependent loss function can offer subtle yet consistent improvements over standard Random Forests (Breiman L 2001) and Support Vector Machine (Cortes & Vapnik 1995). A pan-cancer analysis of mutation from 1218 whole genome screened samples from COSMIC (Forbes S. et al., 2017) identifies number of novel regulatory somatic mutations including previously reported TERT promoter mutations.

6.2. Results and Discussion

6.2.1. Refined definition of regulatory mutations in non-coding genome

Functionally validated mutations are the optimal choice for training a classifier to predict regulatory mutations. In their absence various computational approaches have been developed taking advantage of prior biological insights in to account. For example, site specific recurrence across multiple patients is a strong indicator of a driver mutation and often observed in many tumour activating mutations (e.g. TERT promoter mutation) [Additional file 1: Figure S2]. A handful of studies have successfully trained prediction tools using site specific recurrent mutations and shown that they score recurrent non-coding cancer mutations significantly higher than random mutations. While site-specific recurrence provides a good approximation to capture activating mutations, loss of function mutations, which are often dispersed over a region (e.g. transcription factor binding sites) remains unexplored. To capture them instead of simply looking for vertically stacked site recurrence, we also need to look horizontally across the genome to identify if a particular region of the non-coding genome (e.g. promoters, TFBSs) are getting mutated across different individuals [Additional file 1: Figure S2].

Instead of targeting regulatory mutations, Lee and colleagues in their recent work prioritized hotspots across the non-coding regions of cancer genomes using a sliding window-based approach (Weinhold N & Sander C 2014). These hotspot regions showed significantly higher mutations than background mutation rate and harboured many mutations with potential tumour driver capabilities. We adopted a similar binomial test-based mutation prioritization strategy to identify mutations with significantly higher neighbouring mutations within a particular genomic window compared to the background mutation rate. [Materials and method; Additional file 1: Figure S3]. In the absence of a validated gold standard regulatory mutation set this approach provides us with a reasonable approximation of potential regula-

tory mutation for classifier training.

After applying the technique, our burden test based labelling strategy was applied on 1.81 million non-coding somatic mutations from 1218 whole genome screened tumour samples [Materials and method]. We identified 1160 potential regulatory mutations (positive training set) distributed in 200 hotspots or clusters across the genome. Mutation hotspots were determined by a 100bp single linkage neighbourhood clustering and mutations from the same cluster were labelled with the same identical hotspot id.

This positive set approximation approach described above excluded 56 site specific recurrent mutations (mutated in more than two samples) due to high background mutation rate in the corresponding tumour samples. Considering the importance of site-specific recurrence, we incorporated them in our positive set taking the total positive set size to 1216. Remaining mutations were labelled as negative (negative training set).

One caveat to this approach is that it identifies positive mutations in compact clusters across the genome, which violates the assumption of independence between observations during cross validation - a popular way to assess classification performance. To mitigate this, we used a single representative mutation from each cluster in classifier performance comparison in sections 2.4 & 2.5 while all mutations were used in section 2.6 to assess their driver potential via a cluster blocked cross validation approach [Materials and Methods].

6.2.2. Mutation annotation

Both positive and negative mutations were annotated for 292 genomic features. Features could be binary or real valued ranging in scope to indicate mutation overlap with regulatory elements (e.g. promoters, TFBSs), tri-nucleotide context of the nucleotide change, overlap with accessible genomic regions (e.g. histone marks), genome conservation (Davydov E.V. 2010) and to a mutations ability to alter transcription factor binding affinity. The final feature annotated data is a large data matrix with 1.81 million rows (mutations) and 292 columns (features). A detailed description of the features, their source and processing can be found in Materials and Methods and Additional file 2: Table S2.

Mutations with missing annotation values were excluded. The final positive mutation training set contained 914 potential regulatory mutations from 308 independent tumours. This included previously reported mutations in the TERT promoter region (C228T, C250T) [Fig. 1a; Additional file 2: Table S34].

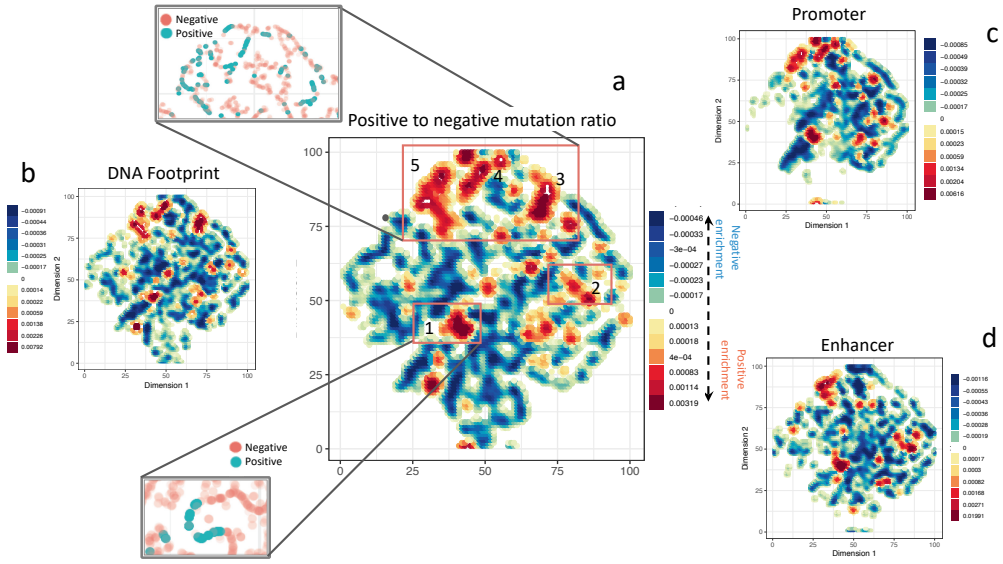


Figure 6.2: (a) Gaussian smoothed positive to negative ratio of two dimensional tSNE projection. Red indicates higher concentration of positive mutations in the block while deep blue indicates the opposite. Red dotted areas indicate enrichment of regulatory and epigenomic markers for positive mutations. Original data points from two-dimensional t-SNE projection are shown in areas highlighted with orange rectangle. (b-e) Positive mutation enrichment for various regulatory elements e.g. transcription factor binding sites, promoters.

6.2.3. Mutation cluster analysis

To identify underlying clusters of positive or negative mutations we deployed t-Distributed Stochastic Neighbourhood Embedding (t-SNE), an unsupervised manifold learning technique (Maaten et.al 2017). For computational feasibility we created a smaller data set including all positive mutations and randomly sampled ten thousand (10K) negative mutations. Points in the two dimensional (2D) representation were further smoothed by applying a 10*x*10 2D Gaussian smoothing window (Fig. 6.2a). For a number of regulatory features such as promoters, enhancers we computed the ratio between feature counts associated with positive and negative mutations in any 10*x*10 window. The Gaussian smoothed representation of these ratios can be found in Fig. 2b-d. Red or yellow shade indicates either enrichment of positive mutations (Fig. 2a) or enrichment of regulatory feature count for positive mutations and blue or green indicates otherwise (Fig. 2b-d).

We identified three different positive mutation clusters. Cluster 1, predominantly breast cancer mutations from a large segment of chromosome 6 was previously described by (Nik-Zainal et al., 2012) and has significant overlap with proximal regulatory elements. Cluster 2 is composed of mutations originating from a large

segment in chromosome 14 and is significantly enriched for repressive elements and HK3K27me3 (Yip et.al 2012). The small positive mutation island (red dot) at the top of in cluster three is significantly enriched for several histone modification elements (e.g. H3K4me2, H3K4me3, H3K27ac) associated with promoter and enhancer activity indicating their potential role in transcription regulation. Non-smoothed data points from two-dimensional t-SNE projection for these clusters are shown in two highlighted windows. Non-smoothed t-SNE figure can found in Additional file 1: Figure S4.

The intertwined nature of positive and negative mutations in complex structures highlighted the necessity to apply a complex classifier rather than a simple rule-based method to distinguish the handful of driver mutations.

6.2.4. Class imbalance and its impact on classification

In any cancer genome, regulatory mutations are a small minority among thousands of benign passenger mutations. This huge class imbalance presents a serious challenge for any learning algorithm trying to distinguish these mutations from passenger ones. Standard classification algorithms assume balanced class distributions and uses an equal misclassification costs per class. Hence, when presented with complex imbalanced data sets such as in this scenario, these algorithms fail to represent the distributive characteristics of the underlying data and provide misleading accuracy often biased towards the majority class (He et al., 2009).

Oversampling of the minority class (Chawla et al., 2002) or under sampling of the majority class (Tomek 1976) has been a popular choice for many imbalance classification problems. The combination of these two approaches have been reported to show improved performance versus their individual application (Batista et al., 2016). We combined Synthetic Minority Over Sampling Technique (SMOTE), a popular over sampling technique with Tomek link, an under-sampling technique to circumvent the class imbalance problem. We tested classic random forest and support vector machine classifiers on both original and re-sampled data sets. Interestingly both classifiers performed better on the former than the latter [Fig. 3A; Additional file 1: Figure S5]. This indicates that the positive mutations lie in a compact neighbourhood clusters with the negative mutations as suggested by the t-SNE analysis and re-sampling fails to generate significant performance improvement.

Class imbalance also has a substantial impact on classification performance evaluation. Receiver Operator Curve (ROC) is a highly popular metric to assess classification performance. However, in scenarios with large class imbalance such as here, ROC curves do not reflect the true classification performance on the desired class (Powers DMW 2015). Due to the large size of the negative training set these mutations are predicted more accurately than the positive ones and the ROC curve stays reasonably stable even with increased negative size giving an ambiguous sense of performance. Precision recall curves on the other hand provides a better insight in

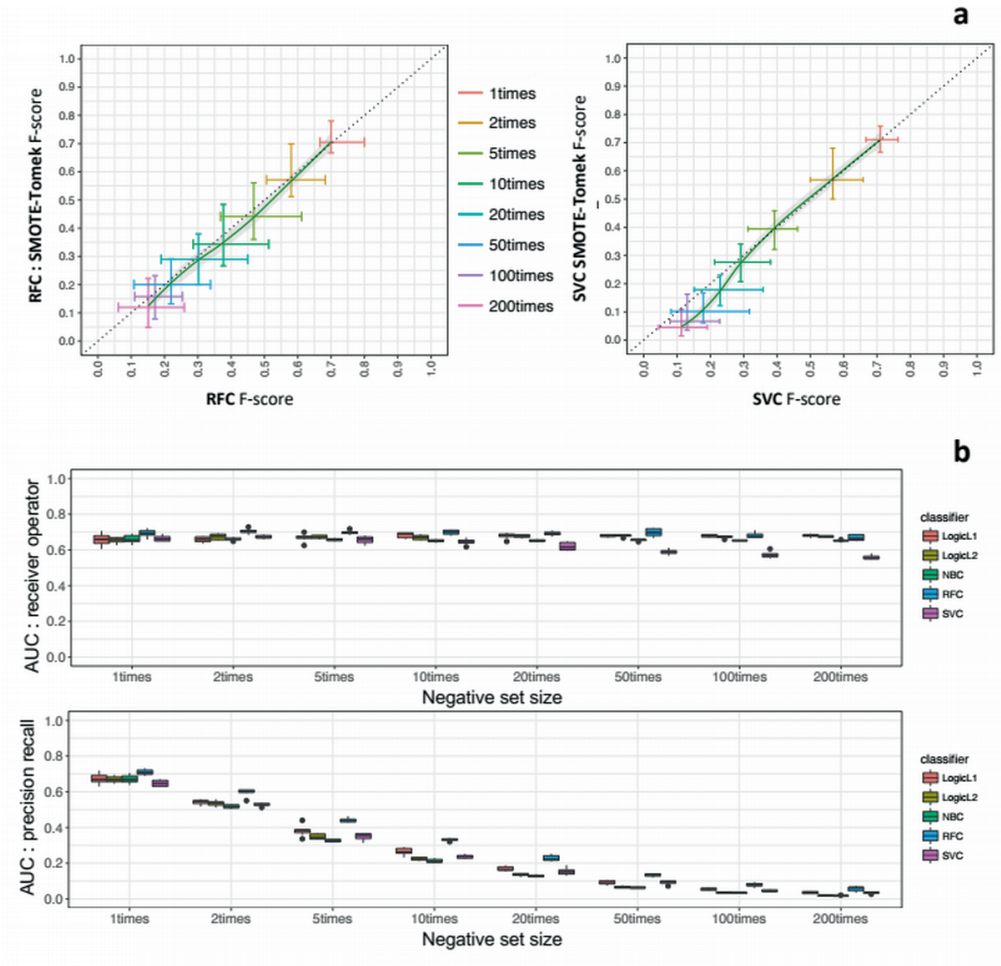


Figure 6.3: (a) Class imbalance problem (a) Maximum F-score comparison between over and under sampled vs original data using random forest classifier and support vector classifier across various negative set sizes. Each data point shows F-score comparison at respective negative set sizes. Error bar indicates variation in maximum F-score between multiple runs of the same negative set size. (b) Area under the receiver operator curve (ROC) and precision recall curve demonstrate how AUC of ROC remain invariant to increasing class imbalance while precision recall reflects the extent of performance decline.

to true classification performance in imbalance scenarios [Fig. 3b] (Davis J. 2006). In the subsequent analysis we have predominantly used precision-recall and F-score (harmonic mean of precision-recall) for classifier performance evaluation.

6.2.5. Asymmetric loss based random forest classifier

Considering our current classification problem, decision tree-based classifier such as the random forest classifier (RFC) offers a favourable solution compared to other classifiers because of their ability to better handle mixture of binary and continu-

ous feature values (Diaz-Uriarte R 2006). Random forest also regulates over-fitting by random sub-sampling of sample and feature space (Breiman et.al 2001). A systematic evaluation of several classifiers (both complex non-linear and simple density-based ones) revealed a classic random forest outperforms other classifiers and is demonstrated comparatively robust performance against increasing class size imbalance [Additional file 1: Figure S6]. We focused on classic random forest framework in an effort to further optimize for class imbalance scenarios.

A classic random forest uses the Gini impurity measure as a node splitting criterion treating both classes with equal weights (Breiman L 2001). This is not, however, suitable for large class imbalance where the desired class (e.g. regulatory mutations) is extremely rare. Typically this limitation is tackled by defining misclassification cost per class. This in itself is not sufficient in scenarios where the positive and negative class has strong overlap and it is acceptable to miss a fraction of positive observations in order to identify the purest set of positive observations. A robust loss function for the positive objects is thereby required.

We developed an asymmetric loss-based tree splitting criteria (asymRFC) that aims to optimize precision. It uses a class dependent loss function to obtain the purest possible split for the positive class and hence minimizes false positive prediction in the case of the large, negative passenger mutation set [materials and methods]. We evaluated asymRFC alongside traditional random forest and several other classifiers and it demonstrates subtle yet consistent performance improvement in larger class imbalance scenarios.

We created eight data sets by randomly selecting one representative positive mutation per positive mutation hotspots identified by our positive label approximation approach and a random selection of negative mutations; gradually increasing the class imbalance (negative set size 1,2,5,10,20,50,100 and 200 times of positive mutations). To test the performance of asymRFC we compared it against a classic/standard RFC and a SVC. All classifiers were trained and tested via a five-fold cross validation. The entire process is then repeated seven times, each time with a different set of positive mutations (randomly chosen representative of a hotspot) and a set of negative mutations. asymRFC shows equal or better performance against all tested classifiers [Fig. 4a]. When we compared the averaged maximum F-score values between asymRFC and RFC across all these independent runs, a small but statistically significant (Wilcoxon rank sum test p value <= 0.05) performance gain was observed in the larger class imbalance setting (negative class 50, 100 and 200 times of positive mutations) [Figure 4b]. A pairwise maximum F-score comparison between asymRFC, RFC and SVC and asymRFC is shown in Additional file 1: Figure S7. These findings suggest that the class dependent loss function approach shows slight improvements over gini impurity-based loss function in larger class imbalance scenarios and might have the potential for further optimization for performance enhancement.

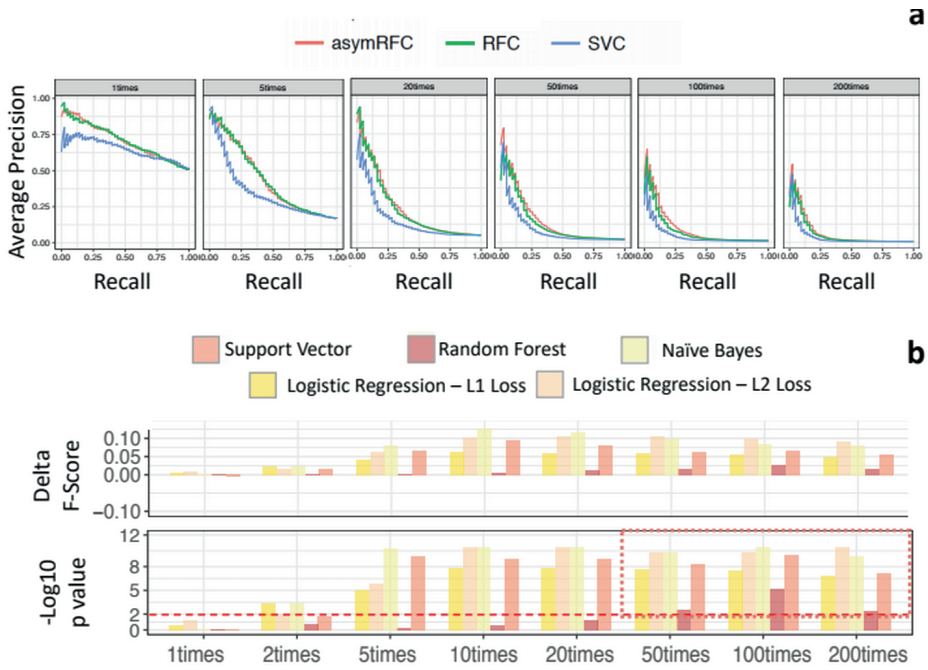


Figure 6.4: Various performance measures averaged from 7 independent runs (a) average precision-recall curve comparing asymmetric loss-based RFC, traditional RFC and SVC with increasing class imbalance. (b) Maximum F-score differences (Δ max F-score) between asymmetric loss-based RFC and five other classifiers and asymRFC shows minor (delta F-score) but significant (dotted red line indicates -log10 p value threshold) performance improvement over other classifiers in larger class imbalances (50, 100, 200 times).

6.2.6. Prioritization of non-coding driver mutations

To assess the cancer driving potential of all our labelled positive mutations, we used asymmetric random forest classifier via a cluster blocked cross validation approach. Positive mutations within 100 bp window of each other are kept entirely within either in training or test set to preserve the independence of observations principle [Materials and Methods]. All mutations were scored via five-fold cross validation and finally mutations with posterior probability score higher than 0.2 (classifier operating point at max F-score) were labelled as positive. We detected 509 mutations as non-coding driver mutation at an average precision of 75% [figure 5a].

More than half (54%) of the predicted non-coding driver mutations fall within proximal regulatory regions such as promoters, repressor elements, proximal enhancers. About 56% of the mutations overlap with transcription factor binding sites active in more than one ENCODE cell type.

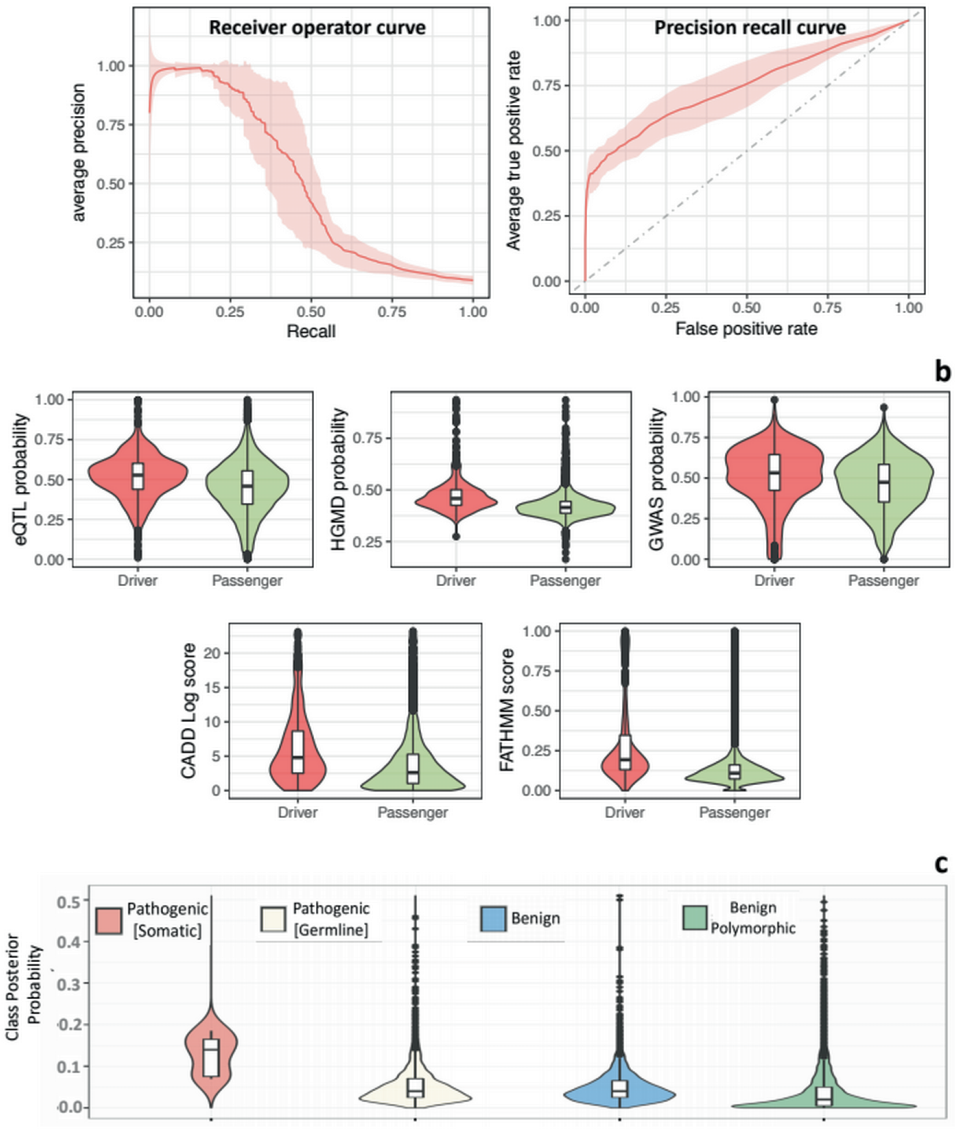


Figure 6.5: (a) Mean receiver operator curve and precision recall curve showing 100 bp cluster blocked five-fold cross validation of an asymmetric random forest classifier on all positive and 10K randomly selected negative mutations. (b) Venn diagram on the left shows overlap between our non-coding driver predictions and CADD and FATHMM likely pathogenic mutations. Venn diagram on the right indicates overlap between these mutations and e-QTL regions. (c) Violin plot showing the prediction score of an asymmetric random forest classifier trained using positive mutations and randomly selected 10K mutations on ClinVar mutations and a set of polymorphic SNPs from 1000Genome.

Promoter mutations can uniquely disrupt or attract the transcription binding machinery hence inhibiting or inducing the downstream transcription process. We identified three mutations in the promoter of TERT including the two well known (228 G>A/T; 250 G>A) cancer driving mutations (Horn et.al 2013). Potential driver mutations were also detected in the promoter regions of several cancer associated genes such as FDFT1, HIST1H2AE, CCDC3, IRF5 (Liao et.al 2017, Nieminen et.al 2014, Kinyamu et.al 2008, Fukuma et.al 2012). FDFT1 (Farnesyl-Diphosphate Farnesyltransferase) encodes a membrane-associated enzyme located at a branch point in the mevalonate pathway and its promoter mutation has previously been associated with aggressive prostate cancer phenotype (Fukuma 2012). The HIST1H2AE gene encodes a member of replication-dependent histone H2A family that regulates DNA accessibility and ranks within the top 10% in respect of the degree of centrality in various interaction networks. Its transcription inhibition has previously been described in breast cancer cell lines (Kinyamu et al., 2008). IRF5 (Interferon regulatory factor 5) encodes a group of transcription factors with diverse roles, including modulation of cell growth, differentiation, apoptosis, and immune system activity. Loss of IRF5 expression in human ductal carcinoma correlates with disease stage and contributes to metastasis (Bi et al., 2011). A complete list of predicted driver mutations annotated with all features and other external prioritization tools can be found in Additional file 2: Table S5.

Predicted driver mutations scores were compared with three established non-coding variant prioritization methods: CADD, FATHMM & DeepSea. Out of the 509 predicted mutations 102 (20%) have CADD score >=10, pathogenicity cut off suggested by CADD authors while 94 (18.5%) mutations have FATHMM score >= 0.5, deleteriousness cut off suggested by authors. Predicted driver mutations have significantly higher CADD (p-value 2.2e-16), FATHMM (p-value 2.2e-16) and DeepSea GWAS (p-value 4.314e-08), HGMD (p-value 2.2e-16), e-QTL (p-value 2.2e-16) probability than passenger mutations suggesting predicted drivers are more likely to influence cancer development [Figure 5b] (Zhou et.al 2015).

We also trained an asymmetric random forest classifier using all 914 positive mutations as well 10 thousand negative mutations and tested its predictive power on several published dataset, such as ClinVar pathogenic mutations and benign polymorphic mutations from the 1000 Genome consortium (1000 Genomes consortium). A Wilcoxon test showed somatic pathogenic mutations have significantly higher posterior probability compared to pathogenic germline (p-value: 3.454e-05) benign (p-value: 1.339e-05), polymorphic SNPs (p-value 2.144e-05) Figure 5c. This demonstrates despite being trained on a cancer specific somatic mutation data set our classifier shows a decent discrimination power on diverse disease associated mutation sets.

6.2.7. Prioritization of germline regulatory mutations

We have previously described the distinct roles germline and somatic mutations play in cancer development [Section 1.3]. Germline mutations associated to can-

cer are convoluted with millions of benign polymorphic variants. To identify these mutations in the non-coding genome, we trained a separate random forest classifier. We used disease associated non-coding variants from the HGMD (Human Gene Mutation Database) as the positive set and unmatched polymorphic variant set from (Ritchie et.al 2014) as negative mutations. The classifier shows a strong performance in discriminating diseases associated variants from population variations with an average AUC score of 0.97 and an average precision score of 0.94 [Additional file 1: Figure S1].

6.3. Discussion

Being better able to identify non-coding mutations has tremendous potential to unlock novel therapeutic approaches and add new perspective towards understanding of cancer genomes. In this paper we aimed to provide a comprehensive overview of several limitations with existing prioritization approaches and explored number of strategies for improvement. Machine learning based approaches have already demonstrated remarkable improvement over rule-based mutation prioritisation methods and provided sufficient examples of validated non-coding drivers significantly advancements can be achieved in the non-coding genome as well (Kircher et.al 2014, Shihab et.al 2015). Recent large-scale cancer genome profiling studies will surely expand the non-coding driver landscape in near future but mutation density-based hotspot/enrichment detection approaches remain the sole option at the moment. Taking spatial organization of the DNA using chromatin confrontation data such as Hi-C can considerably improve the hotspot prediction (Kim K et al., 2016). They used breast and lung Hi-C sequencing data to identify a set of mutations arising in individual samples and altering different cis-regulatory elements that converge on a common gene via chromatin interactions. However, many recent works have shown that chromatin interactions are highly tissue specific (e.g. Yeung J et al., 2018) and interaction data is only available for a small number of cell types, which limits the application of this approach. With increasing number of studies profiling chromatin structure of DNA across different cell types, in near future we should be able to leverage these data and develop a better understanding of mutation hotspot formation in cancer genomes.

Class imbalance remains another fundamental problem in non-coding mutation prioritisation task. In several dataset we tested over and under-sampling strategies such as SMOTE and Tomek-link does not seem to produce any improvements in classification performance [Additional file 1: Figure S5]. Tomek-link under samples by removing pair of examples that belongs to different classes but have the shortest distance between themselves than with any other data points. As demonstrated in the tSNE analysis in section 2.3, positive and negative mutations are in close entanglement in our non-coding data set and this under-sampling approach while removing negative example have most likely removed a large number of positive mutations that are in tomek-link with negative examples causing classifier performance depletion.

The reliable estimation of classifier performance is very crucial for selecting the right classifier. In imbalanced data sets such as this, the importance of accurately predicting the positive mutations significantly outweighs the prediction accuracy of the negative class. Using several imbalance data sets we demonstrated that receiver-operator curves seriously fall short in illustrating the true impact of increasing class imbalance on classification performance. Precision, on the other hand, by comparing false positives to true positives rather than true negatives, provides a more reliable performance comparison between various classifiers. In imbalance scenarios such as this we identified metrics such total/partial area under the precision-recall curve, precision at various recall points or F-score provide more reliable performance estimation than ROCs.

Following an extensive comparison across number of data sets with varying degree of class imbalance, the random forest classifier showed better classification performance compared to the other tested classifiers. Using a class specific asymmetric loss function we have only managed to obtain very subtle performance gain. However, this small yet consistent improvement hinted that optimizing class specific loss function might hold the key for further performance enhancement in non-coding mutation prioritization where class imbalance and entanglement of observations both hinder the learning process.

The predictive performance of any classification model is assessed via stratified cross validation. A stratified data split for training and test subsections works perfectly for a data set where observations are independent (Burman P 2016). Cancer mutations are often closely interrelated to its neighbours for various reasons such as DNA interactions, linkage and genome accessibility. Our window based binomial approach for generating a set of positive and negative mutations for the purpose of model training identified positive mutations in compact clusters across the genome. In a simple stratified cross validation approach these interrelated mutations will be split between training and test folds. Consequently, any classification routine will learn from the split examples in the training set and produce an erroneously inflated performance based on the remaining mutations of the same cluster in the test fold. Numerous previous works have outlined techniques to tackle the dependent observation problem (Telford RJ 2005). We adopted a cluster blocked cross validation approach where mutations from each single linked cluster are kept entirely within training or test set to get a generalized predictive performance [Material and methods; Additional file 1: Figure S8]. Every single mutation was scored via a cluster blocked five-fold cross validation.

Using F-score derived cut-off point we identified 509 non-coding mutations with potential cancer regulating ability. Majority of these mutations overlap with proximal regulatory elements. A number of epigenetic markers associated with transcription activation (e.g. H3K9ac, H4K20me1, H3K4me3) and repression (e.g. H3K27me3) have been significantly enriched for the predicted positive mutations suggesting the possible role of these mutations in disrupting normal transcription

process. The impact of mutations in cis-regulatory elements are arbitrated to downstream transcription process in a more directed fashion compared to the same in trans-elements simply due their close proximity to transcribed elements. We observed the predicted positive mutations are in significantly closer genomic proximity to the transcription start site than the predicted negative (passenger) ones and have significantly higher overlap with proximal regulatory elements than their negative counterparts. We adopted a uniform weight for all features during the prioritization process but a carefully devised feature weighted approach taking prior biological insights into account might result in better classification.

6.4. Methods

6.4.1. Classifier data curation

We downloaded 6463360 non-coding somatic variants from the COSMIC database (version 70) (Forbes SA 2017). A large fraction of these mutations come from exome sequencing studies, which focus on exonic regions. We used somatic mutations from whole genome screened samples. After filtering for exonic mutations, indels (Insertions and Deletions), known population variations (SNPs) we acquired a set of 1.81 million somatic non-coding mutations from 1218 whole genome screened samples [Additional file 1: Figure S9].

6.4.2. Binomial test based positive set selection

To circumvent the problem of gold standard data we have redefined the potential positive mutations leveraging the power of regional recurrence. For every single point mutation, we applied the binomial test shown equation (ii) to compute the likelihood of detecting more than or equal to the observed mutations n_0 within that genomic window given a background mutation rate. The background mutation rate is estimated by averaging the genome-wide mutation rate of the samples contributing to that particular window. This way we have manage to decreases the impact of hypermutated samples. The same test is then repeated across different flanking genomic windows (+/- 1, 2, 4, 6, 10, 20, 30, 50 base pair) for every single mutation observed in the data set. [Additional file 1: Figure S3].

$$p - value(w_i) = B(n >= n_0 | n_0 >= 1; P, W * S) \quad \text{(i)}$$

Where,

$$P = \frac{Total\ Number\ of\ mutation\ in\ non - codinggenome}{Total\ Samples\ \times\ Size\ of\ non - coding\ Genome}$$

W=Window Size

S=Total Number of Samples

Mutations with a significantly (Bonferroni adjusted p-value <= 0.01) high num-

ber of neighbouring mutations in any genomic window were labelled as potential regulatory mutation.

6.4.3. Gaussian smoothing of tSNE result

We applied t-SNE on all positive mutations and randomly sampled ten thousand (10K) negative mutations. To find an optimal two-dimensional (2D) representation of the data we tested several perplexity values (10,20,30) and with minor variation in orientation, the underlying 2D representation remained stable. Both t-SNE dimensions were scaled between 1 to 100 to produce a 100x100 pixel image where each pixel contains a discrete value representing the number of mutations at the co-ordinate. From this image we created two distinct images; positive and negative image only using corresponding mutation counts as pixel values. Each image was then normalised by respective total mutation count to generate a proportionate representation of mutations per pixel bin. We computed a pixel by pixel ratio between the normalised positive and negative image to identify regions enriched for positive (red or yellow) or negative (green or blue) enrichment. Finally, the ratio image smoothed a with a two-dimensional gaussian kernel (standard deviation 1.5) [Figure 2] (Pau G et al., 2010).

6.4.4. Tackling dependent observations problem

Under this approach mutations lying within 100bp genomic window of each other are labelled as a member of the same genomic cluster [Additional file 1: Figure S8]. During the cross-validation process mutations belonging to the same genomic cluster are kept entirely within training or test set to prevent information leak between dependent observations and unbiased estimation of classifier performance. Fig. 4a demonstrates the average performance difference between a tradition cross validations and cluster blocked one. Classifier trained using classic cross validation (red line) gives an inaccurate sense of performance because of ability to easily predict neighbouring mutations when a fraction of mutations from one cluster is used on training and remaining mutations were used in test set. Cluster blocked cross validation however trains using all mutations in a cluster and aims to predict regulatory potential of distal mutations, giving a generalized genome wide predictability.

6.4.5. Class dependent asymmetric loss function

As described in section 2.5, in large class imbalance scenario average precision (AP) provides a better alternative than global loss function. In the absence of knowledge about the true distribution of the data, an empirical average precision estimation for $n+$ sorted positive objects $x^{+}_{(i)}$ and $n-$ negative objects $x^{-}_{(j)}$, can be achieved via:

$$\hat{AP} = \frac{1}{n_+} \sum_{i=1}^{n+} \frac{i}{i + \frac{n_+}{n_-} \sum_{j=1}^{n-} \prod(x^{-}_{(j)} > x^{+}_{(i)})} \qquad \text{(iii)}$$

This allows us to empirically determined the influence of single objects on the AP-performance. An example of positive and negative curve for a simulated two class data set can be seen in Additional file 1: Figure S10 . The shapes of these

curves are remarkably similar to the sigmoid and the exponential function. Based on this observation we propose a novel asymmetric loss-based tree splitting criteria that aims to optimize precision. We use a sigmoid loss for positive class and an exponential loss for negative class to penalize false positives far more severely than false negatives.

In a two-class classification problem, with a positive $y = +1$ and a negative $y = -1$, at every node of the decision the feature space is split in to two discrete regions and for each region an output $\hat{y}$ is predicted. When $\hat{y} < 0$ the object is classified to the negative class, and when $\hat{y} > 0$ the object is classified to the positive class.

Assume that n^+ positive objects and n^- objects fall into a region/node for which the output $\hat{y}$ is predicted. The (average) sigmoid-exponential loss incurred on these objects is:

$$l(\hat{y}) = \frac{1}{n}(n^+ l^+(\hat{y}) + n^- l^-(\hat{y})) \tag{iv}$$

where, $n = n^+ + n^-$

We define the loss on the positive objects:

$$l^+(\hat{y}) = \frac{2}{(1 + exp(\hat{y})} \tag{v}$$

and the loss on the negative objects:

$$l^-(\hat{y}) = exp(\hat{y}) \tag{vi}$$

The optimal prediction y^* is found by setting the derivative of l to 0:

$$\frac{d}{dy} l(y) = \frac{1}{n}(n^+ . - 2.(1 + exp(y))^{-2} \, . \, exp(y) \; + \; n^- exp(y) = 0$$

$$n^- exp(y^*) = 2n^+ . \frac{1}{(1 + exp\,(y^*)} . \frac{1}{(1 + exp\,(y^*)} . exp\,(y^*)$$

$$n^- = 2n^+ . \frac{1}{(1 + exp\,(y^*))} . \frac{1}{(1 + exp\,(y^*))}$$

$$n^- = 2n^+ . \frac{1}{(1 + exp\,(y^*))} . \frac{1}{(1 + exp\,(y^*))}$$

$$\frac{2n^+}{n^-} = (1 + exp(y^*))^2$$

Because n^+ , n^- and $exp(y^*)$ are all positive:

$$1 + exp(y^*) = \sqrt{\frac{(2n^+)}{(n^-)}}$$

$$y^* = log\sqrt{\frac{2n^+}{n^-}}$$

For the optimal prediction value y^*, we see that:

$$exp(y^*) = \sqrt{\frac{2n^+}{n^-}} - 1 \quad \text{(vii)}$$

The loss becomes:

$$l(y^*) = \frac{1}{n}(n^+ \frac{2}{1 + exp(y^*)}) + n^- exp(y^*)$$

$$= \frac{1}{n}\left(2n^+ \frac{1}{\sqrt{\frac{2n^+}{n^-}}} + n^-(\sqrt{\frac{2n^+}{n^-}} - 1)\right)$$

$$= 2\frac{n^+}{n}\sqrt{\frac{n^-}{2n^+}} + \frac{n^-}{n^+}\sqrt{\frac{2n^+}{n^-}} - \frac{n^-}{n} \quad \text{(viii)}$$

This can be further simplified to:

$$l(y^*) = 2\sqrt{2.\frac{n^+}{n}.\frac{n^-}{n}} - \frac{n^-}{n} \quad \text{(ix)}$$

Indeed, the asymmetry of two classes can be seen in equation (ix). Now when we train a tree, at each node of the decision tree we choose one feature and try to decide if splitting the node in two is advantageous. The asymmetric loss between the master node and child nodes are compared.

6.4.6. Additional Data Sets

ClinVar pathogenic mutations

The National Center for Biotechnology Information (NCBI) ClinVar database is a

public archive reporting the relationship between human variants and diseases (Landrum M et al., 2014). We obtained 119,602 variants from ClinVar build v144. We discarded any known population variation using 1000 genome phase 1 SNPs and mutations common with training data set. Using ClinVar defined categories we extracted 5153 benign mutations, 4370 pathogenic germline variants and 9 pathogenic somatic variants.

Chromatin recurrence cohort
Potential regulatory mutations identified by Jung Kyoon et.al based on chromatin recurrence in Breast and Lung cancer data sets (Kim K et al., 2016). Different negative sets (1, 2, 5, 10, 20, 50 times of the positive mutations) were chosen.

HGMD disease causing mutations
Disease associated non-coding mutations from Human Gene Mutation Database as described in Ritchie et.al 2014.

Polymorphic SNPs
Unmatched benign polymorphic mutations from 1000 genome project as described in Ritchie et.al. 2014 and 1000 Genomes consortium.

6.4.7. Feature Data

We compiled a collection of 292 genomic, epigenomic, regulatory and cancer associated data sources for feature annotation. Feature data is a mixture of continuous numeric values (e.g. Conservation score, TF binding affinity), binary values (e.g. mutation overlaps with a known repeat Region or cpg island) and some composite features (sum or average over a flanking region). A brief description of each feature group and source is described below.

Genome Segmentation
We used integrated annotation of chromatin elements (ChromHMM and SegWay) from ENCODE, which grouped regulatory elements in seven different states (ENCODE Consortium 2012).

Transcription Factor Binding Sites (TFBS)
We obtained ChIP-seq peak calls for 124 transcription factors from ENCODE, JASPAR via Ritchie G et.al. 2014

Chromatin Status
We have used formaldehyde-assisted isolation of regulatory elements followed by sequencing and DNaseI hypersensitivity assay followed by sequencing and peak calls and DNase footprints (DNASE_FPS) from ENCODE. Each data track is informative independently but it has been shown that a combined approach produces more meaningful results (ENCODE consortium 2012).

Histone modifications
We used Peak call data from ChIP-seq experiment of 12 different modifications from ENCODE obtained via Ritchie et.al. 2014.

Gene annotation context
We used the gene annotation regions (e.g. utr, intron, exonic) and distance from transcription start site, splice site from GENCODE (v16) (Harrow J et al., 2012).

Distal and proximal regulatory regions
We used proximal (promoter) and distal (enhancer) regulatory information from two different sources as described via Khurana E et al., 2013.

Network gene features
From three gene networks (regulatory gene network, protein-protein interaction network and phosphorylation networks) reported by Khurana et al., 2014, we identified genes placed within top 10% and 25% of the degree of centrality rank. Gencode (v16) promoter and enhancer regions of these genes were used to annotate the network feature as described in Khurana E et al., 2013.

Genome conservation
We used Genomic Evolutionary Rate Profiling (GERP) Rejected Substitution (RS) scores and neutral scores from Sidow lab (Davydov E et al., 2010).

Trinucleotide Mutation Context
Mutations were annotated for 96 trinucleotide mutation context features.

Cancer Associated Features
Cancer mutations are often seen clustered around particular genomic loci because only a handful of mutations will give the cell a selective survival advantage. We used mutation rate per megabase (Mb)) and Oncogenic MicroRNA (OncoMir) information (Wang K et al., 2014).

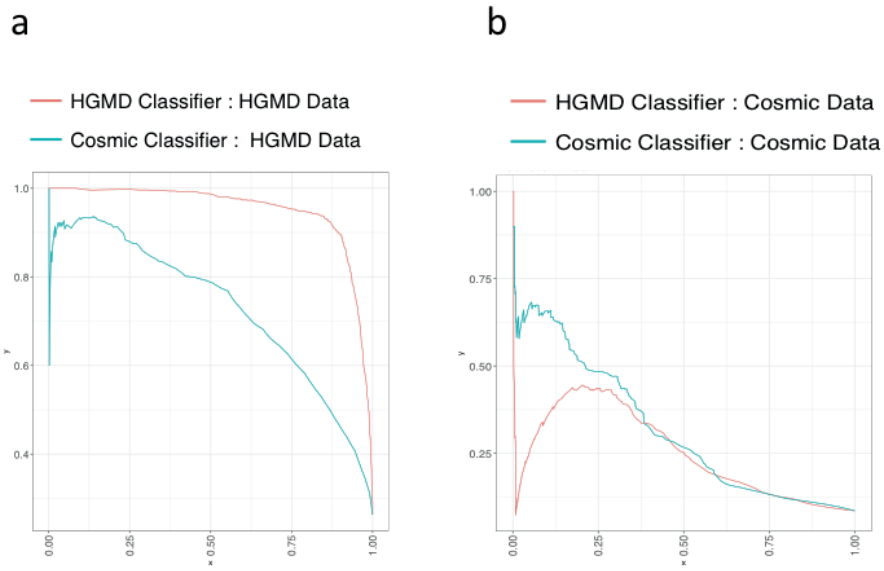
A description of individual features, their category and encoding types can be found in Additional file 2: Table S2.

6.5. Acknowledgements

We sincerely Dr. Alistair G. Rust for his during relevant literature review and helpful comments. We also thank Dr. David J. Adams for his insightful comments about the prediction results.

6.6. Supplementary Materials:

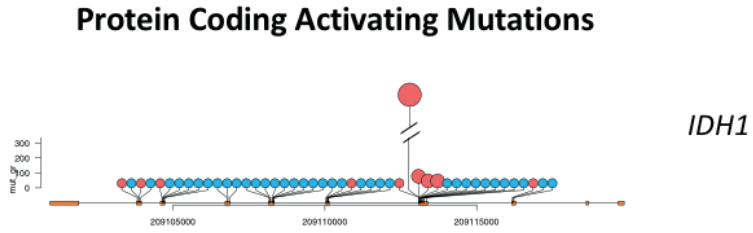
Supplementary Figure 1



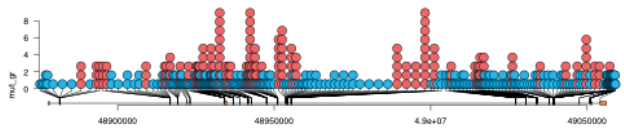
Supplementary Figure 1 : Performance of somatic and germline mutation trained classifier. Two classifiers trained on HGMD pathogenic variant and benign SNPs and COSMIC binomial based positive and negative data; **(a)** Prediction on HGMD pathogenic variant and benign SNPs **(b)** Prediction on COSMIC binomial based positive and negative data. This indicates somatic and germline noncoding mutations have distinct patterns and it is difficult to achieve a generalised prediction performance.

Supplementary Figure 2

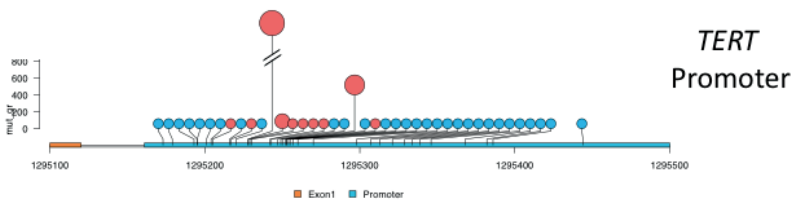
a



b

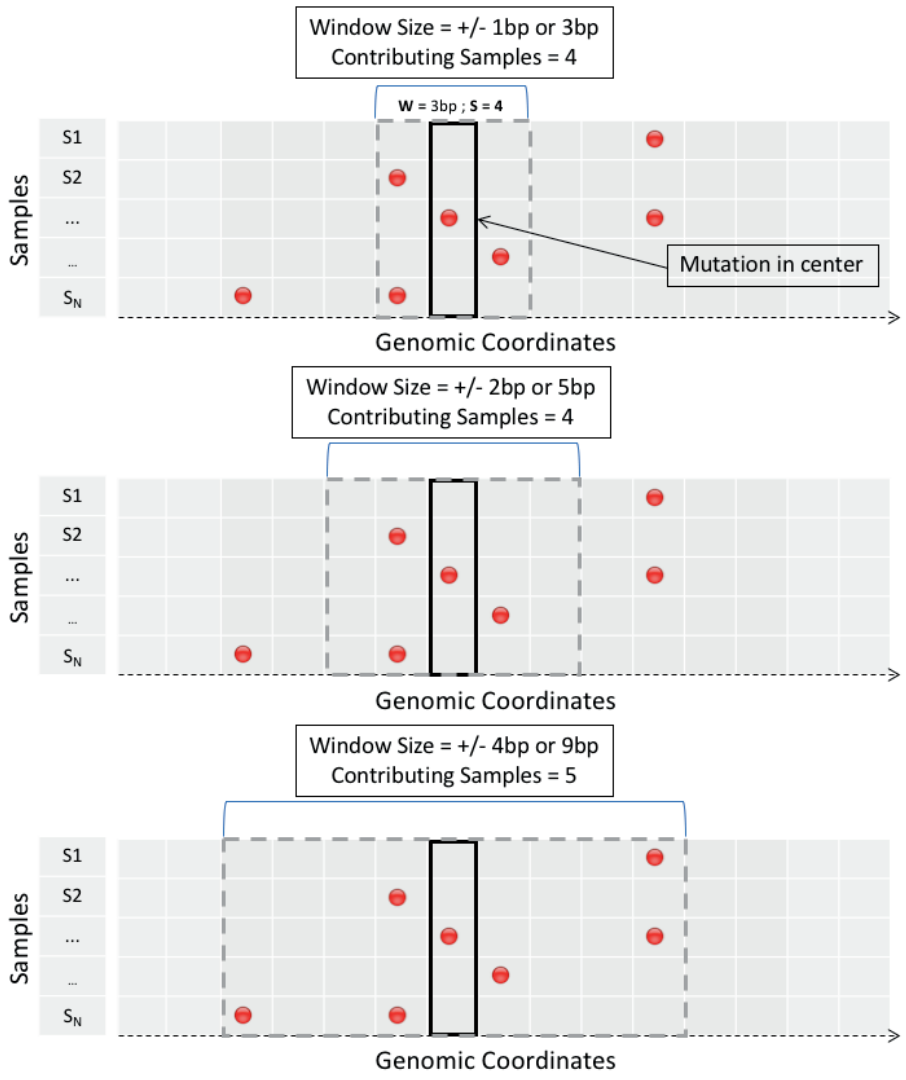


c



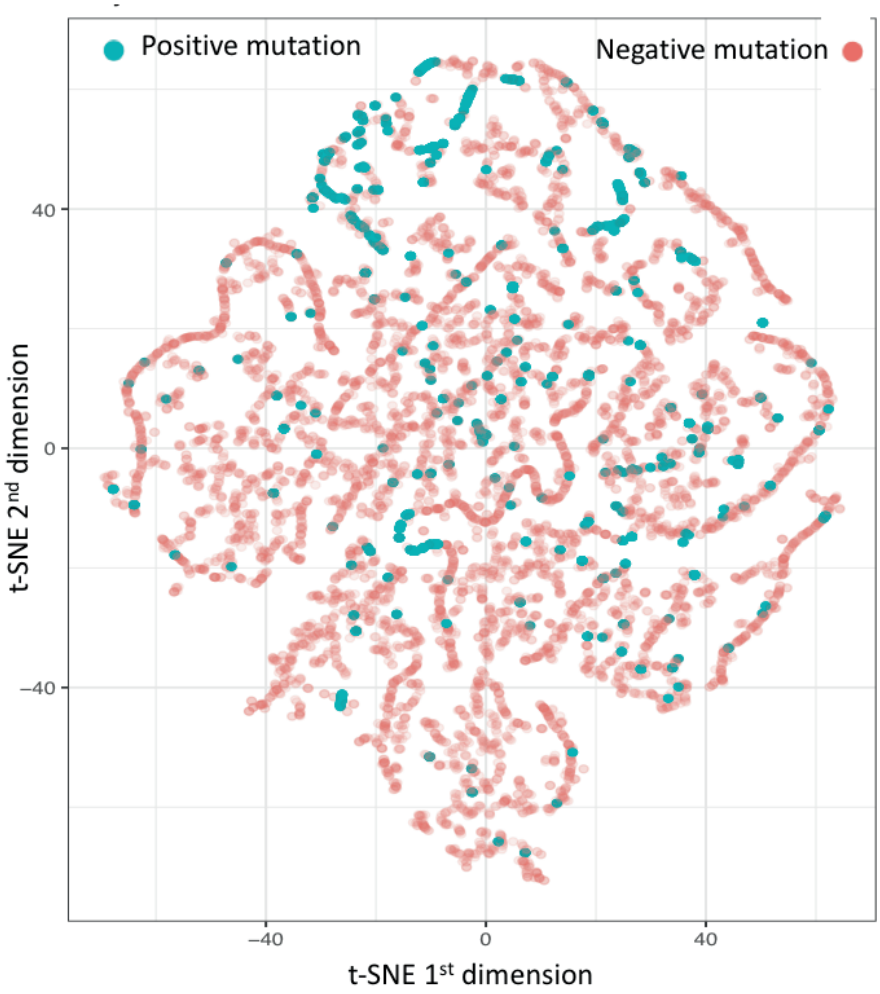
Supplementary Figure 2 : Activating and disrupting mutations : (**a**) Activating mutations in oncogene *IDH1* are cantered R132H locus while (**b**) disrupting mutations in tumour suppressor gene *RB1* is spread across the gene body. (**c**) Activating mutations in *TERT* promoter are also concentrated at two loci.

Supplementary Figure 3



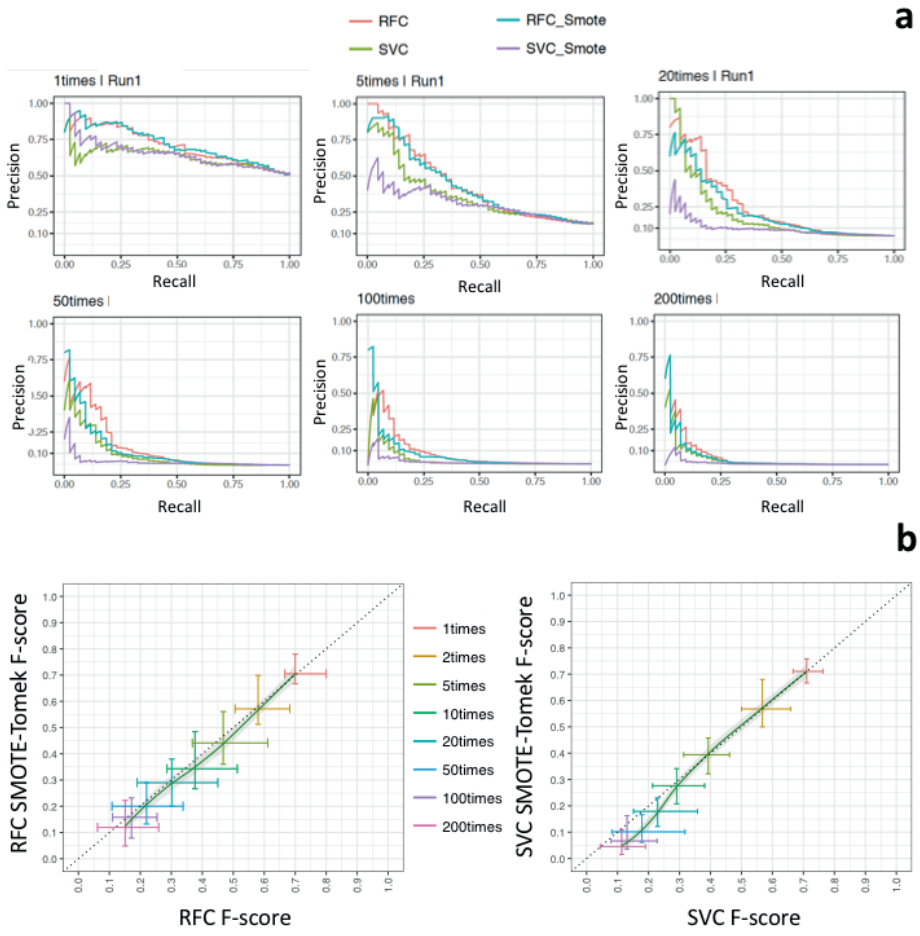
Supplementary Figure 3 : For every point mutation we applied binomial test to compute the likelihood of observing more than or equal to (>=) ***n*** mutations in a window and the test is repeated for genomic windows (+/- 1, 2, 4, 6, 10, 20, 30, 50 bp) around the mutation

Supplementary Figure 4

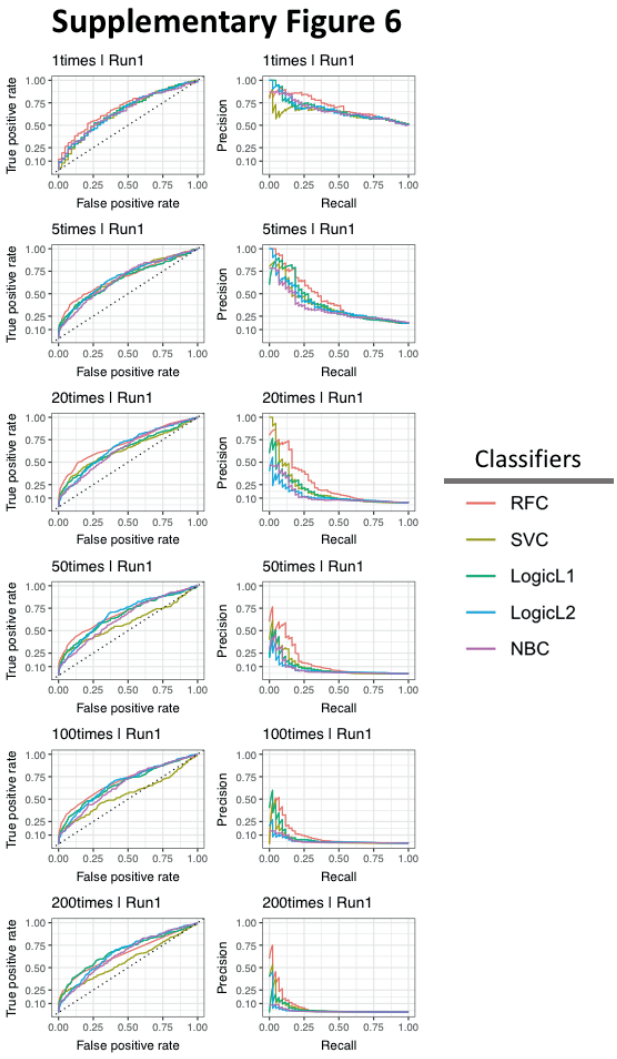


Supplementary Figure 4 : Two dimensional non-smoothed t-SNE projection of positive and 10,000 randomly selected negative mutations.

Supplementary Figure 5

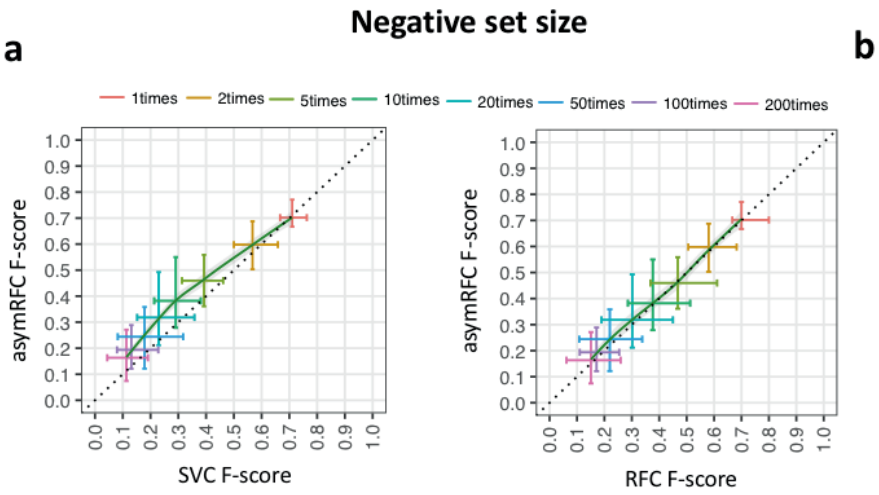


Supplementary Figure 5 : Oversampling (SMOTE) of the minority class and under-sampling (Tomek-link) of the majority class (**a**) Average precision recall curve showing performance random forest and support vector classifier with sampling and without sampling. **(b)** Maximum F-score comparison between over and under sampled vs original data for both aforementioned classifiers across. Each data point shows F-score comparison at respective negative set sizes. Error bar indicates variation in F-score between multiple runs at the same negative set size.

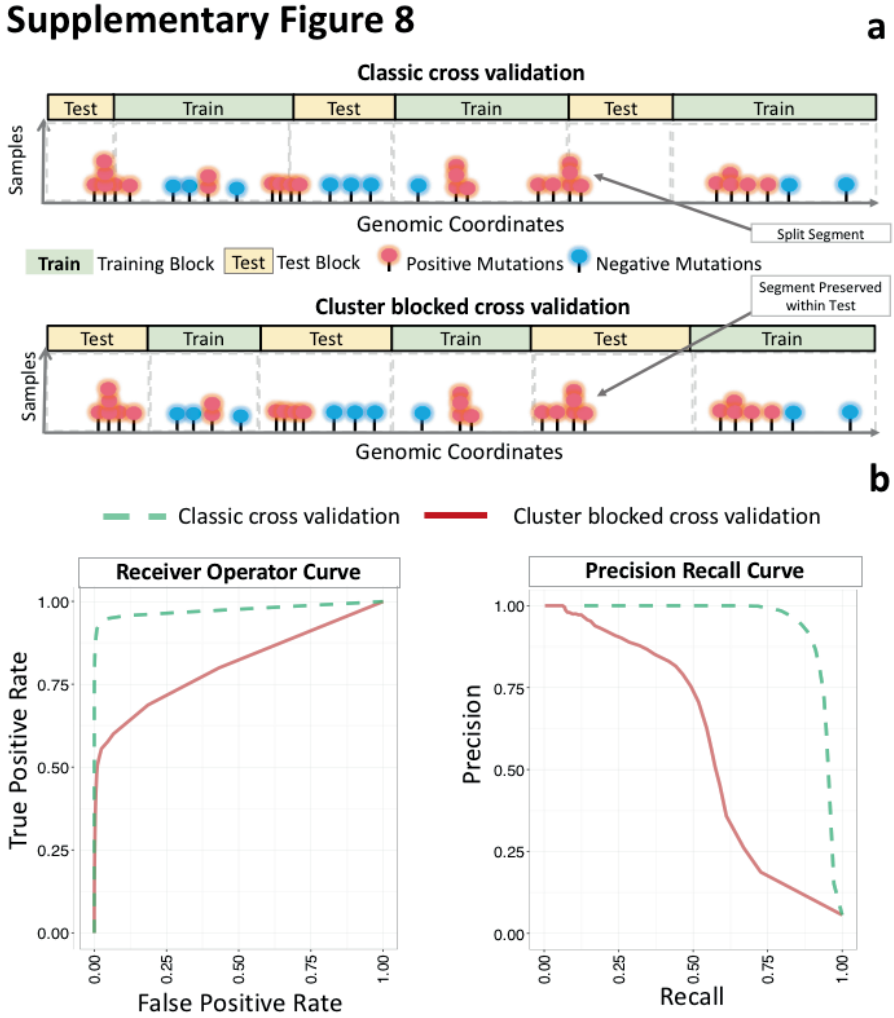


Supplementary Figure 6 : as Comparison of random forest classifier with other linear and non-linear classifiers. Random forest shows robust performance with increasing class imbalance.

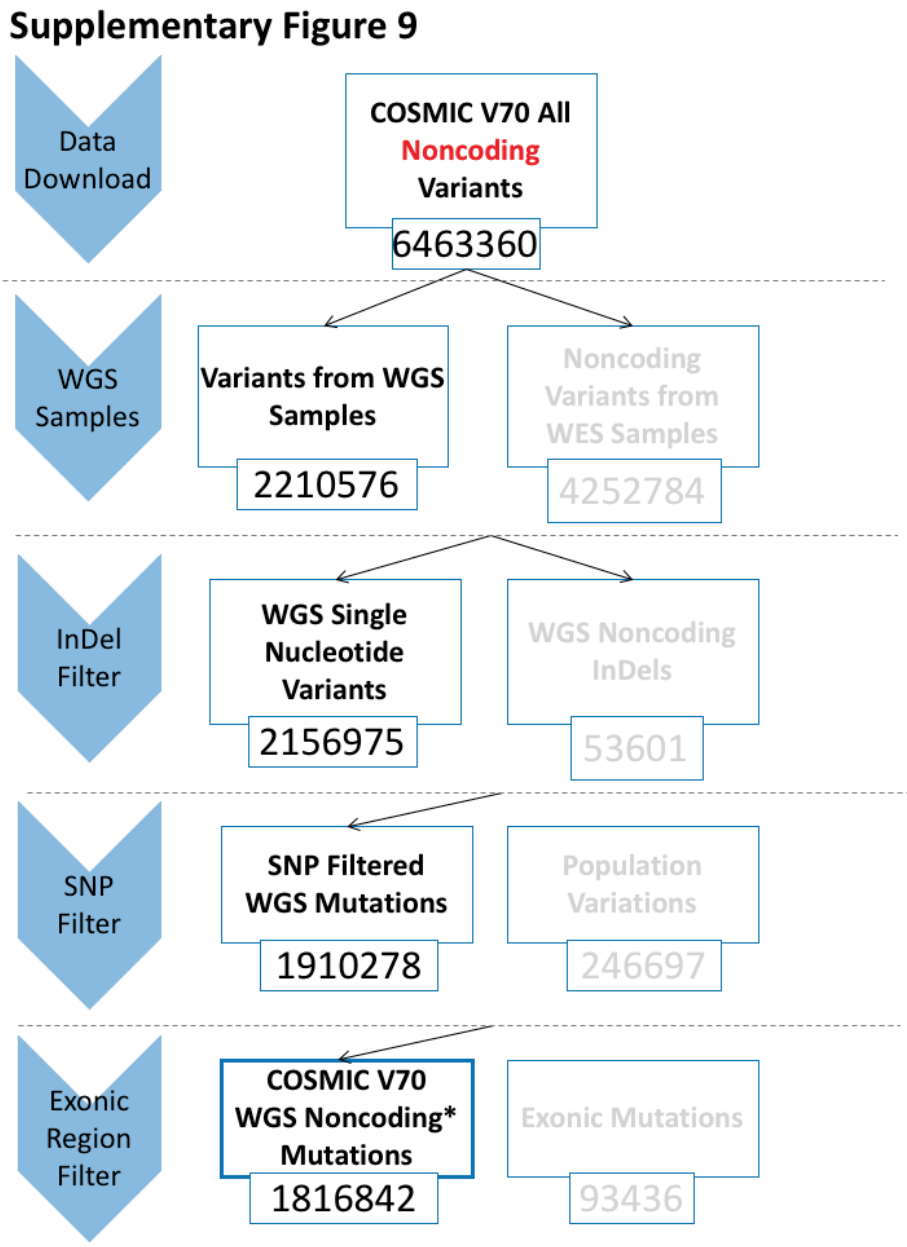
Supplementary Figure 7



Supplementary Figure 7 : Comparing average of maximum F-score from each run between two classifiers. Left side panel (**a**) shows comparison between asymmetric loss based random forest and support vector classifier and right hand panel (**b**) shows comparison between asymmetric loss based random forest and a traditional random forest. For each negative set size max F-score from each independent run was average to plot the line. Error bar indicates minimum and maximum max F-score from those runs.

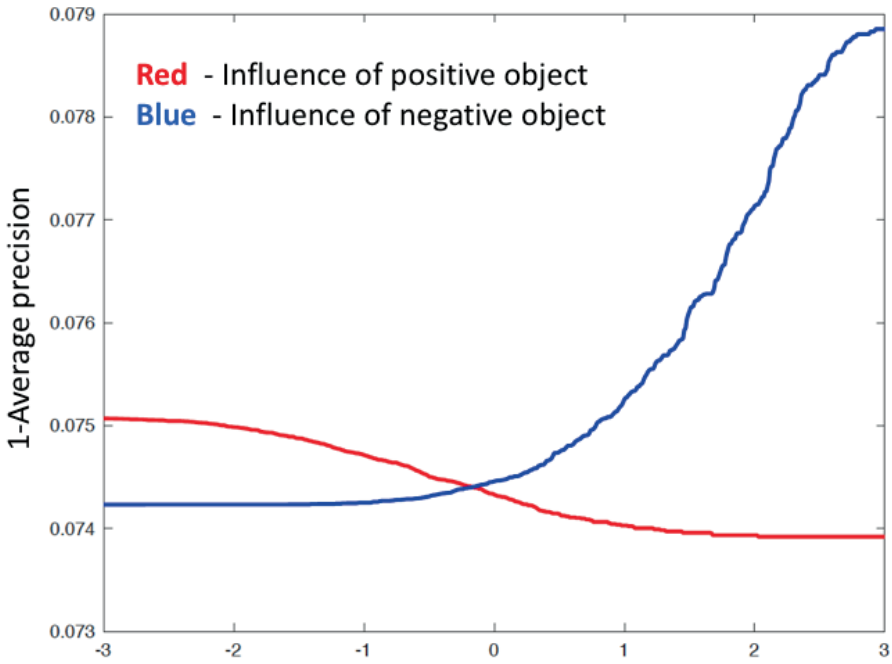


Supplementary Figure 8 : Dependent observation problem and segment blocked cross validation (**a**) Classic cross validation randomly splits the data maintaining the class label balance (***top panel***) while cluster-blocked cross validation approach (***bottom panel***) uses a 100bp single linkage cluster to bin mutations lying within **100bp genomic window** of each into same cluster and mutations with identical cluster ids are kept entirely within training or test set to prevent information leaking between dependent observations. (**b**) ROC and PRC curve shows the artificially inflated performance (green line classic cross validation) vs cluster blocked cross validated result.



Supplementary Figure 9 : Data pre-processing workflow.

Supplementary Figure 10



Supplementary Figure 10 : Empirical precision recall. We empirically determine the influence of single objects on the AP-performance. A two-class, one-dimensional dataset is created, where 50 objects per class are drawn from two Gaussians, with $\mu_+ = +1$ and $\mu_- = -1$.

One additional object x is added, once with a label y = +1 and once with a label y = -1. For each situation the AP-performance is recorded. 1-AP, called AP-loss is shown in the figure. The **red** curve shows the influence from the positive objects, the **blue** line the influence from the negative objects.

Note that the shape of these curves are remarkably similar to the sigmoid and the exponential function.

7

Discussion

Cancer is a complex heterogeneous disease with many subtypes each with their own unique genomic characteristics. Despite remarkable advancements in cancer patient care and treatment, a World Health Organization (WHO) estimation predicted around 9.6 million cancer-related death worldwide in 2018 alone. Because it is a disease caused by aberrant changes in our genome, genomic research is at the forefront of the ongoing battle against cancer. A complete genomic characterization of these sub-types will lead to better patient stratification and treatment.

In this thesis, we presented several novel approaches. This included a method to improve the reliable detection of somatic mutations in cancer genomes, in-silico and in-vitro framework for improving the detection accuracy somatic mutations and a prioritization method to distinguish driver mutations both within and beyond the protein-coding genome. We began by introducing several fundamental concepts & challenges in cancer genomics and their present solutions as a foundation for the understanding of the remainder of the thesis. In chapter 2, we presented a novel framework combining several existing somatic single point mutation detection tools and a novel post-processing work-flow to improve the detection accuracy. We exploited this newly developed workflow in chapter 3 to study early-stage adenomas from two distinct sub-types of human colorectal cancer patients, their mutational profiles, inter-tumour heterogeneity and driver genes. In chapter 4, we explored the genetic profiles of paediatric/adolescent melanoma patients and compared those with that of adult melanoma patients in the context of therapeutic intervention and clinical management of these patients. We searched for driver genes and investigated the transition of a benign skin adnexal tumour to malignancy in chapter 5. Finally, in chapter 6, we ventured beyond the boundaries of protein-coding genes and performed a pan-cancer analysis to identify novel cancer-driving noncoding mutations. In this chapter, we shall reflect on our findings and explore the possibilities of utilizing the described methods beyond the demonstrated examples, their limitations and possible future improvements.

7.1. Somatic mutation detection: challenges and prospects

Recent advancements in sequencing techniques and significant reduction in the DNA re-sequencing cost have reinforced the call for precision cancer medicine derived from patient genomic data. Fast and accurate detection of somatic variations, together with distinguishing cancer driving events from benign passenger ones, will play pivotal roles in this endeavour.

7.1.1. Genome-wide vs high-depth targeted perspective

Despite recent reduction in cost, Whole Genome Sequencing (WGS) cancer samples is significantly more expensive than targeted DNA sequencing and one of the frequent conundrums faced by many cancer sequencing studies is to select between these two options (Schwarze et al. 2018). Both approaches come with their respective advantages and challenges. WGS is more suitable in detecting large scale genomic variations such as copy number changes, structural variations that do not

require high sequencing depth, while targeted sequencing is ideal for detecting low allelic sub-clonal variations. In chapter 3, we deployed a combination of Whole Exome Sequencing (WES) and targeted gene sequencing to investigate multiple early stage polyps collected from a set of colorectal adenoma patients. WES data allowed us to detect novel driver genes, mutational signatures and tumour clonal composition, while using the high-depth targeted gene sequencing data we investigated the discovered novel driver genes at a much finer resolution. In recent years, methods such as FREEC (Boeva et al. 2012) and Sequenza (Favero et al. 2015) have shown reliable performance on estimating copy number profile from exome sequencing data. However, whole genome sequencing remains the primary choice for accurately detecting large scale genomic variations. In chapter 4, therefore we profiled tumour-germline pair of a melanoma patient using WGS and reported tumour specific copy number changes and novel inter-chromosomal translocations (Rabbie et al. 2017).

7.1.2. Analysis of formalin fixed tumour samples

Fresh frozen samples are the optimal choice for DNA/RNA sequencing but due to the unavailability of this capability across many cancer centres, decade old Formalin-Fixed Paraffin-Embedded (FFPE) tissue preservation technique is still the predominant source of storing cancer samples. Despite recent improvements in sequencing techniques the concordance between somatic mutation detected from FFPE and fresh frozen sample remains poor (Robbe et al. 2018). In line with the findings reported by Wong et al.(2014), a systematic evaluation of **119** randomly selected adnexal tumour mutations revealed that the majority of these artifacts occur at low depth (**Supplementary Fig. 5.11**). We have also shown how using a rule-based filtering strategy the number of artefacts can be reduced significantly. Nonetheless, it is important to remember that computational approaches alone are not sufficient to clean up all artefacts. DNA quality control steps (e.g. fragment size analysis) and treatment procedures such as the one suggested by Do and Dobrovic (2012) prior to sequencing, can significantly reduce the number of false positives.

7.1.3. Mutation validation strategy: necessity or extravagance

In chapter 1, we briefly discussed the impacts of artefacts to distinguish true somatic mutations from false ones and how orthogonal validation can significantly improve our ability to reduce false positives. Here we further elaborate on the scope, advantages and drawbacks of some of the available technologies and techniques. Traditional Sanger sequencing, the mass spectrometry based sequenom platform, KASP genotype assays and high-depth targeted loci sequencing are among some of the popular techniques. Throughout chapters 2 to 5 we used these techniques independently or in conjunction with each other to validate genetic variations. Sanger sequencing is probably the cheapest of the spectrum but it is a tedious sequential process and often not efficient to validate more than a handful genomic loci. Additionally, Sanger sequencing technique suffers from a lack of sensitivity at low allelic fraction and often fails to detect mutations present below 10% allele fraction. Sequenom or KASP genotype assays offer a comparatively better parallelization but

are significantly more expensive. Finally, high-depth targeted loci sequencing approaches (e.g. Illumina MiSeq) are the most expensive in this spectrum and are more suitable for scenarios where the aim is to sequence multiple samples across many genes/loci. Use of a different aliquot of DNA is highly recommended for orthogonal validation to eliminate potential artefacts that may have been introduced during NGS sample preparation. Orthogonal validation requires considerable time and resources and sometimes can be considered unnecessary. False discoveries can have disastrous clinical implications (Tandy-Connor et al. 2018). A large proportion of the published mutations available in various known databases such as COSMIC, cBioPortal were never validated orthogonally and thousands of users around the world use these data everyday completely oblivious to the dangers of false positive rates in these data sets. We strongly recommend that every large scale genome profiling study should undergo some degree of orthogonal validation for a better understanding of their false discovery rate (Beck et al. 2016). Expanding the sample cohort during the validation process will also consolidate findings.

7.2. Therapeutic insight through better understanding of tumour heterogeneity

Heterogeneity within cancer cells (spatial heterogeneity) at any given time or throughout various stages of its evolution (temporal heterogeneity), plays a crucial role in developing several key tumour properties such as drug resistance and metastasis. Collecting multiple samples from the same individuals, spatially or longitudinally, is a non-trivial process and requires substantial resources. However, these studies give us unique insight about the disease evolution. Another conundrum faced by investigators is whether to perform high-depth sequencing of targeted loci to identify sub-clonal mutations or to sequence whole genomes to get a genome wide perspective from larger genomic variations such as copy number changes. In chapter 3, we used a combination of whole exome sequencing and high-depth targeted gene sequencing to explore both end of this spectrum. Whole exome sequencing allowed us to identify mutational signatures and sub-clones while through targeted gene sequencing, we managed to verify low allele somatic mutations in a novel driver gene (Rashid et al. 2016). In chapter 5, our comparison of benign and malignant components of skin adnexal tumours contrasted the current hypothesis that malignant spiradenocarcinomas almost always arise from benign spiradenomas. We investigated the benign and malignant tumours of six individuals and found only malignant tumour to carry all the mutations from its benign precursor, indicating that the malignancy might be initiated independently of their benign counterpart. These findings surely give us a unique insight into these tumours, their progression and novel therapeutic intervention angles.

While the bulk tissue sequence based techniques used in the aforementioned analysis can give us a decent approximation of cancer heterogeneity, their resolutions are often not adequate to delineate the complete heterogeneous landscape at a single cancer cell resolution. In recent years droplet based single cell sequencing technologies have shown tremendous capacity in characterizing tumour heterogeneity

(Tsoucas and Yuan, 2017; Levitin et al. 2018). For example, Janiszewska et al. (2015) have reported two distinct subpopulation of cells, one with *PIK3CA* mutation and the other with *HER2* (*ERBB2*) amplification within HER2-positive breast cancer during neoadjuvant therapy. Recent work by Zheng et al. (2018) on hepatocellular carcinoma patients reported the presence of a distinct set of cancer stem cell populations with distinct molecular features and potentially contribute to differences in treatment response. Future tumour profiling studies should, therefore, take the heterogeneous tumour landscape into account and consider exploiting these new technologies before diving into this complex multidimensional challenge.

7.3. Towards personalized cancer treatment

Genome sequencing studies have elevated our understanding of the cancer genome, driver genes, tumour heterogeneity as well as evolution and the relevance of these phenomenons concerning treatment and drug resistance. Recent advancements in cancer immune checkpoint blockade therapies in several cancer types have bolstered the case for patient tumour data derived personalized cancer treatment. In chapter 4, we compared the genomic profiles of paediatric/adolescent patients and adult skin cutaneous melanoma patients. Our analysis revealed a subset of paediatric/adolescent patients that carried somatic mutation burden and driver mutations similar to adult cutaneous melanoma patients. High mutational load, a common indicator of higher neo-antigen expression and is strongly associated with better response to immunotherapies such PD-1 inhibitor and anti-CTLA4. Interestingly, despite the high mutational burden, paediatric patients are not routinely considered for immunotherapy, mostly due to the risk of side effects of the immunotherapy itself. A comparative analysis of an extended paediatric melanoma cohort previously published by Lu et al. (2015) and adult cutaneous melanoma patients published by the TCGA consortium, revealed that many adolescent melanoma patients exhibit molecular features similar to their adult counterparts (chapter 4 and Rabbie et al. 2017). In light of these findings reported in chapter 4 and Rabbie et al. (2017), we argued that genomic data should be routinely taken in to account during clinical course assessment of paediatric patients.

7.4. Driver mutations detection: potentials, pitfalls and future directions

Driver mutations in the protein-coding genome and particularly driver genes have been the primary focus of cancer studies for decades. Many pan-cancer driver genes such as *TP53,BRCA1/2* and BRAF and cancer specific driver mutations such as *CDKN2A* deletion in melanoma, *APC* mutation in colorectal cancer, have allowed scientists to develop therapeutic interventions targeting these genes. In chapter 3, we used a novel Monte Carlo simulation-based technique to assess the significance of the abundance of truncating mutations in *AMER1*(APC Membrane Recruitment Protein 1), a gene commonly inactivated in wilms tumours, in colorectal cancer pa-

tients. Although we observed a significant enrichment of loss of function mutations, due to the absence of the any functional validation, we were unable however to pinpoint it's role as a driver mutation in these tumours (Rashid et al. 2016). Driver gene discovery methods reported in chapter 5, reliably detect genes under positive selection in the tumour genomes but these findings should always be supported by additional analysis such as pathway or interaction network analysis and in-vitro or in-vivo functional experiments. In chapter 5, we identified a novel hotspot mutation in the alpha kinase domain of *ALPK1*(Alpha Kinase 1) gene defining a sub-population of skin adnexal tumours. The hotspot mutation was orthogonally validated using Sanger sequencing and using fluorescent cell imaging technique we confirmed that in its mutant form this gene increases the activity of the Nfkβ pathway, a pathway responsible for transforming inflammation into malignancy.

Although not as frequently studied as somatic driver gene searching studies, hunting for germline cancer-predisposing gene also occupies a significant area of cancer research. In chapter 5, we also searched for genes that may dispose of individuals to adnexal tumour development. Mutation of *CYLD* is a known germline predisposing gene in these tumour types. As discussed in chapter 5, this is mostly done by assessing the mutation burden with respect to a control population and selecting an appropriate control population plays a vital role in this analysis. Our analysis revealed a number of interesting genes carrying significantly more deleterious mutations in the adnexal tumour genomes compared to the control population. These include previously described *CYLD* and *FAT4*, a gene previously implicated as a tumour suppressor gene in Drosophila. Detection of germline cancer risk alleles requires a large number of samples and family pedigree to identify if mutation(s) in a particular gene is recurrently observed within affected members of a family. Despite our best efforts to mitigate the impact of various confounding factors by focusing only on deleterious mutations (CADD score >= 15), the small cohort size (only 43 individuals) and the absence of family pedigree severely restricted our ability to confidently ascertain the risk associated with the mutations in these genes. In the case of many cancer profiling studies, the genomic characteristics of family members are overlooked but this information plays a vital role in germline risk allele identification (Robles-Espinoza et al. 2014). Health care services in many developed countries are gradually moving towards this practice as a routine diagnostics measure and in near future, we should, therefore, be able to detect germline cancer risk-associated loci more accurately. Finally, even in the presence of a larger data cohort and the family pedigree, serious caution should be taken with respect to variant quality, the control population, balancing sequencing depth in both cases as well as control data sets [Chapter 5 : material and method].

Until very recently our understanding of the genome was more or less limited to the protein-coding regions and as a result, the discovery of cancer-driving genes dominated the central stage of cancer research for decades. Further, it is safe to say that they still have a significant role to play in cancer diagnosis and treatment. The rising demand for multi-faceted genomic screening for personalized genomics will

produce an unprecedented volume of genomic data and the biggest challenge lies in fast and accurate detection of somatic mutations and correctly associating these mutations with disease progression through functional analysis and experimental studies.

7.5. Noncoding driver mutations: a new hope beyond the coding genome

The publication of large scale epigenome regulation data by the ENCODE consortium and the Roadmap Epigenome project have revolutionized our perspective of the noncoding genome. Epigenetic modifications, such as DNA methylation and histone modification, regulatory elements, such as transcription factors, play crucial roles in regulating the transcription and translation of many cancer-driving genes (Esteller 2007; Sandoval and Esteller 2012). The study of cancer epigenome and mutations that perturb the epigenetic landscape has gained momentum in recent years (Kircher et al. 2014; Weinhold et al. 2014; Zhou and Troyanskaya 2015).

In chapter 6, we attempted to predict noncoding driver mutations using a supervised machine learning approach and we tried to address several challenges in the field. As described in chapter 6, the absence of a gold standard (experimentally validated) noncoding driver mutations pose the most significant hindrance to the prediction accuracy of any pattern-based approach. Experimentally validating the cancer-driving properties of a mutation requires significant resource and time. We are still decades away from establishing any such large collection of noncoding driver mutations. We presented several strategies to improve existing in-silico approaches. Our proposed window-based enrichment analysis to identify mutations within genomic regions under positive selection offers a reasonable substitution for putative, gold-standard cancer drivers. It detected a number of cancer-associated hotspots including the well known and validated TERT promoter mutations. Enrichment based approaches such as this are prone to erroneous prediction mostly due to hyper-mutated samples. We addressed this issue by correcting for the genome wide mutation rate of each sample contributing to the region under question. However, tumour genomes are not uniformly mutated and taking regional mutation rate (e.g. per chromosome) in-to account is likely to produce a better approximation of true positive selection (Weinhold et al. 2014). Kim and others (2016) have shown modelling the spatial structure of the chromatin can considerably refine the output of the hotspot detection approach. With a rapid increase in chromatin interaction profiling across various cell types this approach offers a better alternative in the absence of validated cohort of driver mutations.

In any tumour genome, driver mutations are a small minority while benign passenger mutations dominate the landscape. This imbalance in the size of driver and passenger mutation sets severely affects the learning and prediction processes. In chapter 6, we used a fixed driver (positive) mutation set against a gradually increasing passenger (negative) mutation set to demonstrate the impact of class imbalance

on prediction performance. We showed how popular classifier performance evaluation metric such as the Receiver Operator Curve (ROC) can produce misleading impressions in imbalanced scenarios (figure 6.3). An evaluation metric that does not rely on true negative sets such as precision-Recall, F-measure, area under the precision recall curve or average precision are some of the better alternatives to ROC in imbalanced learning scenarios.

A comprehensive comparison of five different classifiers presented in chapter 6 revealed that the random forest classifier provided the most reliable classification performance in most scenarios, most likely due to its ability to better handle mixed feature types. However, all classifiers suffered significant performance depletion as the imbalance increased. Imbalance data sets restrict our ability to learn the true properties of the minority class, however important it might be. This is most likely because the intrinsic design of classic loss functions aim to optimize the cost by uniformly minimizing both false positive and false negative and misclassification of a handful of minority objects does not impact the global loss. Under-sampling of the majority class as shown in Kim et al. (2016) can improve the scenario but increases the chance of under-fitting. We improved over the performance of the classic random forest, by developing a class dependent loss function where false positives were penalized at a significantly higher rate than false negatives. Our classifier bench-marking exercise presented in figure 6.4 showed a random forest classifier with class-dependent loss function produces significantly better results in larger class imbalance scenarios than classic random forest and other tested classifiers.

The analysis of pan-cancer data sets allows us to harness the power of large data sets and to search for noncoding driver mutations active across different cancer types. However, one caveat of this approach is that regulatory and epigenomic elements used to annotate mutations, to generate the feature matrix for supervised learning, are highly tissue-specific. We strongly believe cancer specific noncoding driver mutation prediction strategy, where mutation from a certain cancer type will only be annotated with regulatory and epigenetic elements active in that tissue type, will produce more meaningful results.

7.6. Concluding remarks

In this thesis, our primary focus was to address two fundamental challenges in cancer genomic. First, the accurate detection of somatic mutations from cancer genome sequencing data and second, to identify the driver potential of detected somatic mutations. We also invested significant efforts in verifying the clinical implications of these discoveries via downstream experimental validation. In the limited scope of this thesis we also explored a number of other important genetic events such as copy number variations and mutational signatures. We explored beyond the traditional boundaries of protein coding genome to look for novel noncoding driver mutations and reported a number of exciting candidates. Experimental follow-up of these novel cancer driving candidates will significantly enrich our knowledge of

cancer genomes.

Summary

Cancer is an umbrella terminology that binds hundreds of complex genetic diseases based on a set of common phenotypic hallmarks. Each cancer and their sub-types have their unique genomic profiles. The common factor that binds them all together is that they all arise from changes in the DNA. Theses changes range from single nucleotide levels variation to large scale chromosomal aberrations. The consequences of these changes also have distinct impacts on disease development and progression depending on their ability to alter the protein function. Changes in the DNA of a protein-coding gene might have a directly quantifiable impact while quantifying the impact of a change in the regulatory DNA (viz. noncoding) element is a non-trivial task. A better understanding of the complex interplay between coding and noncoding genetic variation will lead to a better understanding of the diseases and improve diagnostics and patient care.

This thesis proposes a novel framework for reliable prediction of somatic point mutations in cancer genomes. The framework was applied to several whole-genome and exome sequenced cancer datasets. Our findings suggested that a consensus-based approach produces a more reliable result than individual mutation detection tools. We also proposed an in-silico post-processing workflow and in-vitro validation guideline to improve the detection accuracy of using orthogonal techniques. Different cancers have distinct mutational burden and profile and understanding these genomic sub-types will lead to better patient stratification and clinical management. Using mutational signature analysis we investigated the inter- as well as intra-tumour heterogeneity in colon adenomas and skin adnexal tumours. By comparing the mutational signature as well as mutation burden between adult and paediatric patients, we identified striking genomic similarities between them. Based on these findings, we recommend that like many adult patients, genomic profiles of paediatric patients should also be routinely taken into consideration while deciding the therapeutic course.

Mutations that give selective survival advantages to cancer cells are commonly referred to as driver mutations. These mutations can occur both in the protein-coding region of the genome or beyond it. This thesis reviewed several available driver mutation detection tools and identified a few areas with a considerable scope of improvement. We proposed a novel machine learning-based framework to prioritize noncoding driver mutations in cancer genomes.

Samenvatting

Kanker is een overkoepelende term die honderden complexe genetische ziekten beschrijft via een set gedefinieerde eigenschappen. Elke soort kanker heeft een eigen karakteristieke genoom profiel, maar wat al deze ziekten met elkaar gemeen hebben is dat ze de gevolgen van mutaties in het DNA zijn. Deze mutaties gaan van een enkele base substitutie tot verandering op chromosoom schaal en als een mutatie het proteïne van een gen verandert kan dat gevolgen hebben voor de progressie van de ziekte. Maar voor mutaties die buiten genen vallen (in niet-coderend DNA) is het kwantificeren van de gevolgen niet triviaal. Een beter begrip van de wisselwerking tussen coderende en niet-coderende mutaties zal daarom bijdragen aan betere behandelingsmogelijkheden en zal leiden tot een beter begrip van kanker als ziektebeeld.

In deze thesis wordt een framewerk voor het accuraat identificeren van somatische punt mutaties in kanker genomen gepresenteerd. Het framewerk is toegepast op verscheidene whole genome en whole exome sequencing datasets. Analyse van de resultaten geeft aan dat een consensus van meerdere individuele detectie algoritmen het meest betrouwbaar is. Verder presenteren we een in-silico nabewerkingstap en een in-vitro validatie richtlijn om de detectie nauwkeurigheid van detectie algoritmen te verbeteren. Verschillende soorten kanker bevatten specifieke mutationele patronen en het beter begrijpen van deze eigenschappen kan leiden tot betere behandelingskeuzen. Via het voorgestelde framewerk bestuderen we de intra- en inter-tumor heterogeniteit van adenocarcinoom van de dikke darm en adnexal huid tumoren. Verder vinden we sterke overeenkomsten tussen vormen van pediatrische kanker in kinderen en volwassenen. Op basis van deze bevindingen stellen we voor dat het routinematig afnemen van een genomisch profiel kan bijdragen bij het kiezen van de juiste behandelingsstrategie van kinderkankers.

Mutaties die een kankercel een selectief voordeel geven worden doorgaans drivers genoemd. Deze mutaties kunnen zowel binnen het coderende en niet-coderende genoom voorkomen. In deze thesis worden verscheidene driver detectie algoritmen besproken en worden een aantal aspecten genoemd waar verbetering mogelijk is. We stellen een nieuw framewerk voor waarbij via een machine learning algoritme driver mutaties in het niet-coderende genoom geprioriteerd kunnen worden.

Acknowledgements

Jeroen, I probably wouldn't be writing this thesis if not for your constant support throughout my PhD. I am not aware any PhD supervisor who persistently allocate time for his/her PhD student even when they are sitting literally outside their doors. You went great lengths to support me academically as well as emotionally through some difficult times. Machine learning has been a area of keen interest for me since my masters in 2008 but after studying your work with Alistair and Dave, inspired me to pursue a PhD on this topic.

Alistair, without your support I might not be able to put together the courage to start this challenging endeavour in the first place. You not only introduced me to the murky world of next-generation sequencing data analysis but are also my golf teacher (Still trying to improve my swing). Helping me to organise the very first meeting with Jeroen and Marcel, enthusiastically sending me relevant research papers and in so many other aspects of this work is indebted to you. I am sincerely grateful for the fact that despite not working at the same place for more than five years and not physically meeting each other for almost three years you are continuous effort to support me through this.

Marcel, I would like to begin by thanking you for your eternal patience with me. Due to my visa complications I was not able to fly to Delft for my interview and You and Jeroen flew in to Cambridge to meet me. I was genuinely humbled by that gesture. Your encouraging words during every year end evaluation has always been a big source of motivation.

Shyama, even if I managed to successfully complete my Phd, convincing you to spend rest of your life with me will probably remain my biggest achievement in my life. No one else supports every crazy steps of my life as you do. PhD took both of us through some serious rough patches in last few years. During the first couple of years of my PhD, I literally took the back seat from managing life and you had to take whole lot more responsibility then you were prepared for. I am very proud that despite an ongoing battle outside and within you have successfully finished your PhD and continued to support me. I am sure we will both make through this.

Dave, I joined your lab in 2011 as a naive computer scientist trying to learn bioinformatics. From the very beginning you encouraged and allowed me to take responsibility for several large scale bioinformatics projects. It helped me to grow as a scientist. I have always admired you scientific integrity and your dedication to science will always be an inspiration to me. With your ever helpful and friendly personality you have managed to build a wonderful research team and I am proud

that for almost seven and half years the experimental cancer genetics team has been my second home. I have met so many wonderful people there and some made great friends.

David, I have met you first time during the pattern recognition workshop and both your passion and your depth of the knowledge just blew my mind. I am really proud that for a part of my PhD, I had the opportunity to work closely with you. I would like to take this opportunity to sincerely thank you for all the trips you made from Delft to Utrecht to meet myself and Jeroen to discuss the one class classification problem.

Sofia, your dedication to science and resilient attitude in difficult times have to truly inspired me. In a way we both ended up supporting each other during some tough times of our PhD. Your meticulous attention to details to the targeted pull-down bait design greatly helped the ALPK1 story to see the daylight. I am glad our paths crossed because of our common interests in science and through our shared scientific struggles we became good friends.

Curriculum Vitæ

Mamunur Rashid was born in Noahkali, Bangladesh on the 7th of August, 1985. He grew up in Dhaka and completed his higher secondary school in 2002 from Notredame college, Dhaka. In 2003, he received the prestigious Indian Council for Cultural Relations scholarship and moved to Pune, India to study Computer Science. He finished is bachelor degree with a distincion in 2006 and moved back to Bangladesh. For a brief period of four months he worked for BRAC, largest non govt. organisation as a network rollout officer. In 2007, he moved to London to study M.Sc. in Advanced Methods of Computer Science. During his masters, he picked up interest in the application of machine learning in digital image processing. As part of his M.Sc. thesis he developed a mobile appication for facial beauty assessment using partial Marquardt beauty mask. He completed his M.Sc. degree with merit in November 2008. He started working as a computer scientist in Kings College, London as part of Anurist, an European collborative project to identify biomarkers for early detection. While working for this project he developed a keen interest in bioinformatics and application of pattern recognition in biomedical research. To further pursue this interest he next took the position of a genome data analyst at Professor Andrew Tutt's lab investigating the genomic and transcriptomic markers of triple negative breast cancer. Finally, in 2011, he moved to the Sanger Institute and under the supervision of Dr. David Adams he has been working on a number cancers subtypes to pinpoint the underlying genetic causes. The opportunity to work alongside leading scienttists in the field allowed him to develop vast experience in the field of cancer genomics. In 2014, he started a PhD under the supervision of professor Marcel J.T. Reinders and professor Jeroen De. Ridder to identify the cancer driving mutations in the noncoding genome. One particular aspect of his research was to apply machine learning algorithms to mine large sclae genomic and epigenomic data to develop better insight about cancer casuing noncoding mutations. As part of Sanger Institute's public outreach programme he actively participated in a number of public engagement activities to aware the impact of genomic reserch in healthcare. He has also founded Shouharda Youth Foundation, an non-profit organization to educate and inspire youths from deprived background in Bangladesh.

List of Publications

1. Rashid M., van der Horst M., Mentzel T., Butera F., Ferreira I., Pance A., Rütten A., Luzar B., Marusic Z., de Saint Aubain N., Ko S. J., Billings S. D., Chen S., Abi Daoud M., Hewinson J., Louzada S., Harms PW., Cerretelli G., Robles-Espinoza C.D.., Patel RM., van der Weyden L., Bakal C., Hornick J. L., Arends M. J., Brenn T. and Adams D. J., 2019. ALPK1 hotspot mutation as a driver of human spiradenoma and spiradenocarcinoma. Nature Communications. ISSN 20411723., (URL: http://doi.org/10.1038/s41467-019-09979-0).

2. Rashid M., Robles-Espinoza C. D., Rust A.G. and Adams D. J. (2013) Cake: A bioinformatics pipeline for the integrated analysis of somatic variants in cancer genomes. Bioinformatics. ISSN 14602059, (URL: http://doi.org/10.1093/bioinformatics/btt371).

3. Rashid M., Fischer A., Wilson C. H., Tiffen C. H., Rust A. G., Stevens P., Shelley I., Maynard J., Williams G. T., Mustonen V., Sampson J. R. and Adams D.J. (2016). Adenoma development in familial adenomatous polyposis and MUTYH -associated polyposis: Somatic landscape and driver genes. Journal of Pathology. ISSN 10969896, (URL: http://doi.org/10.1002/path.4643).

4. Rabbie R., Rashid M., Puig S., Arance A. M., Sánchez M., Tell-Marti G., Potrony M., Conill C., van Doorn R., Dentro S., Gruis N. A., Corrie P., Iyer V., Robles-Espinoza C. D., Puig-Butille J. A. and Adams D. J. et al. (2017). Genomic analysis and clinical management of adolescent cutaneous melanoma. Pigment Cell and Melanoma Research. ISSN 1755148X, (URL: http://doi.org/10.1111/pcmr.12574).

5. Wong C. C., Martincorena I., Rust A. G., Rashid M., Alifrangis C. Alexandrov L. B., Tiffen J C, Kober C, Chronic Myeloid Disorders Working Group of the International Cancer Genome Consortium; Green A. R., Massie C. E., Nangalia J., Lempidaki S., Döhner H., Döhner K., Bray S. J., McDermott U., Papaemmanuil E. Campbell P. J. and Adams D. J. (2014). Inactivating CUX1 mutations promote tumorigenesis. Nature Genetics . ISSN 15461718, (URL: http://doi.org/10.1038/ng.2846).

6. Westcott P. M., Halliwill K. D., To M. D., Rashid M. , Rust A. G., Keane T. M., Delrosario R., Jen K., Gurley K. E., Kemp C. J., Fredlund E., Quigley D. A., Adams D. J. and Balmain A. (2015). The mutational landscapes of genetic and chemical models of Kras-driven lung cancer. Nature. ISSN 14764687, (URL: http://doi.org/10.1038/nature13898)

7. Van der Weyden L., Papaspyropoulos A., Poulogiannis G., Rust A. G., Rashid M., Adams D. J., Arends M. J. and O'Neill E. (2012). Loss of Rassf1a synergizes with deregulated Runx2 signaling in tumorigenesis. Cancer Research. ISSN 15387445, (URL: http://doi.org/10.1158/0008-5472.CAN-11-3343).

8. Pierfrancesco M., Mathew M., Grigoriadis A., Wu Y., Kyle-Cezar F., Watkins J., Rashid M. De Rinaldis E, Hessey S., Gazinska P., Hayday A., Tutt A. (2014). IL15RA drives

antagonistic mechanisms of cancer development and immune control in lymphocyte-enriched triple-negative breast cancers. Cancer Research . ISSN 15387445, (URL: http://doi.org/10.1158/0008-5472.CAN-14-0637).

9. Gonzalez-Meljem M., Haston S., Carreno G., Apps J. R., Pozzi S., Stache C., Kaushal G., Virasami A., Panousopoulos L., Mousavy-Gharavy S. N., Guerrero A., Rashid M., Jani N., Goding C. R., Jacques T. S., Adams D. J., Gil J., Andoniadou C. L. Martinez-Barbera J. P. (2017). Stem cell senescence drives age-attenuated induction of pituitary tumours in mouse models of paediatric craniopharyngioma. Nature Communications. ISSN 20411723, (URL: http://doi.org/10.1038/s41467-017-01992-5).

10. Din S., Wong K., Müller M. F., Oniscu A., Hewinson J., Black C. J., Miller M. L., Jiménez-Sánchez A., Rabbie R., Rashid M, Satsangi J., Adams D. J. and Arends M. J. (2018). Mutational analysis identifies therapeutic biomarkers in inflammatory bowel disease-associated colorectal cancers. Clinical Cancer Research . ISSN 15573265, (URL: http://doi.org/10.1158/1078-0432.CCR-17-3713).

11. Clipson A., Wai-In C., Sasca D., Yiangou L., Osaki H., Basheer F., Gallipoli P., Burrows N., Erdem A. , Sybirna A., Foerster S., Zhao W., Sustic T., Harrison A. P., Laurenti E., Okosun J., Hodson D., Wright P., Smith K. G., Maxwell P., Fitzgibbon J., Du M. Q., Adams D. J., Huntly B. J. P., (2017). Early loss of Crebbp confers malignant stem cell properties on lymphoid progenitors. Nature Cell Biology. ISSN 14764679, (URL: http://doi.org/10.1038/ncb3597).

12. Sherborne A. L., Davidson P. R., Yu K., Nakamura A. O., Rashid M. and Nakamura J. L, (2015). Mutational Analysis of Ionizing Radiation Induced Neoplasms. Cell Reports. ISSN 22111247, (URL: http://doi.org/10.1016/j.celrep.2015.08.015).

Bibliography

- Loewe, L. and Hill, W. G. (2010). The population genetics of mutations: Good, bad and indifferent. Philosophical Transactions of the Royal Society B: Biological Sciences. https://doi.org/10.1098/rstb.2009.0317

- Milholland, B., Dong, X., Zhang, L., Hao, X., Suh, Y. and Vijg, J. (2017). Differences between germline and somatic mutation rates in humans and mice. Nature Communications. https://doi.org/10.1038/ncomms15183

- Epstein, D. J. (2009). Cis-regulatory mutations in human disease. Briefings in Functional Genomics & Proteomics. https://doi.org/10.1093/bfgp/elp021

- Stratton, M. R., Campbell, P. J. and Futreal, P. A. (2009). The cancer genome. Nature, 458(7239), 719–724. url: http://dx.doi.org/10.1038/nature07943

- Konnick, E. Q. and Pritchard, C. C. (2016). Germline, hematopoietic, mosaic, and somatic variation: Interplay between inherited and acquired genetic alterations in disease assessment. Genome Medicine. https://doi.org/10.1186/s13073-016-0350-8

- Karki, R., Pandya, D., Elston, R. C. and Ferlini, C. (2015). Defining mutation and polymorphism in the era of personal genomics. BMC Medical Genomics. https://doi.org/10.1186/s12920-015-0115-z

- 1000 Genomes Project Consortium, Abecasis, G. R., Altshuler, D., Auton, A., Brooks, L. D., Durbin, R. M., … McVean, G. A. (2010). A map of human genome variation from population-scale sequencing. Nature, 467(7319), 1061–1073. https://doi.org/10.1038/nature09534

- The Cost of Sequencing a Human Genome | NHGRI. (n.d.). Retrieved August 11, 2019, from https://www.genome.gov/about-genomics/fact-sheets/Sequencing-Human-Genome-cost

- Landrum, M., Lee, J., Riley, G., Jang, W., Rubinstein, W., Church, D. and Maglott, D. (2013). ClinVar. In The NCBI Handbook. https://doi.org/10.4016/12837.01

- Hanahan, D. and Weinberg, R. A. (2011). Hallmarks of cancer: The next generation. Cell. https://doi.org/10.1016/j.cell.2011.02.013

- Parikh, K., Antanaviciute, A., Fawkner-Corbett, D., Jagielowicz, M., Aulicino, A., Lagerholm, C., … Simmons, A. (2019). Colonic epithelial cell diversity in health and inflammatory bowel disease. Nature. https://doi.org/10.1038/s41586-019-0992-y

- Casasent, A. K., Schalck, A., Gao, R., Sei, E., Long, A., Pangburn, W., … Navin, N. E. (2018). Multiclonal Invasion in Breast Tumors Identified by Topographic Single Cell Sequencing. Cell. https://doi.org/10.1016/j.cell.2017.12.007

- Siegel, R., Naishadham, D. and Jemal, A. (2013). Cancer statistics, 2013. CA: A Cancer Journal for Clinicians. https://doi.org/10.3322/caac.21166

- Needles in a genome haystack. Biochimica et Biophysica Acta (BBA) - Molecular Cell Research, 1843(11), 2698–2704. https://doi.org/10.1016/j.bbamcr.2014.08.001 Loewe, L. (2008). Genetic Mutation. Nature Education, 113.

- Cohen, A. J., Saiakhova, A., Corradin, O., Luppino, J. M., Lovrenert, K., Bartels, C. F., … Scacheri, P. C. (2017). Hotspots of aberrant enhancer activity punctuate the colorectal cancer epigenome. Nature Communications, 8, 14400. https://doi.org/10.1038/ncomms14400

- Vogelstein, B., Papadopoulos, N., Velculescu, V. E., Zhou, S., Diaz, L. A. and Kinzler, K. W. (2013). Cancer genome landscapes. Science. https://doi.org/10.1126/science.1235122

- Bernstein, B. E., Stamatoyannopoulos, J. A., Costello, J. F., Ren, B., Milosavljevic, A., Meissner, A., … Thomson, J. A. (2010). The NIH Roadmap Epigenomics Mapping Consortium. Nature Biotechnology, 28(10), 1045–1048. https://doi.org/10.1038/nbt1010-1045

- Gerstein, M. B., Kundaje, A., Hariharan, M., Landt, S. G., Yan, K. K., Cheng, C., … Dobrovic, A. (2015). Architecture of the human regulatory network derived from ENCODE data. Nature, 7(1), 91–100. https://doi.org/10.1038/nature11245

- Khurana, E., Fu, Y., Chakravarty, D., Demichelis, F., Rubin, M. A. and Gerstein, M. (2016). Role of non-coding sequence variants in cancer. Nat Rev Genet, 17(2). https://doi.org/10.1038/nrg.2015.17

- Tsoucas, D. and Yuan, G. C. (2017). Recent progress in single-cell cancer genomics. Current Opinion in Genetics and Development. https://doi.org/10.1016/j.gde.2017.01.002

- Fehr, A., Kovács, A., Löning, T., Frierson, H., van den Oord, J. and Stenman, G. (2011). The MYB-NFIB gene fusion-a novel genetic link between adenoid cystic carcinoma and dermal cylindroma. The Journal of Pathology, 224(3), 322–327. https://doi.org/10.1002/path.2909

- Horn, S., Figl, A., Rachakonda, P. S., Fischer, C., Sucker, A., Gast, A., … Kumar, R. (2013). TERT promoter mutations in familial and sporadic melanoma. Science, 339(6122), 959–961. https://doi.org/10.1126/science.1230062

- Polak, P., Karlić, R., Koren, A., Thurman, R., Sandstrom, R., Lawrence, M. S., … Sunyaev, S. R. (2015). Cell-of-origin chromatin organization shapes the mutational landscape of cancer. Nature, 518(7539), 360–364. https://doi.org/10.1038/nature14221

- Li, H. (2012). Exploring single-sample SNP and INDEL calling with whole-genome de novo assembly. Bioinformatics, 28. https://doi.org/10.1093/bioinformatics/bts280

- Ryan, N. M., Morris, S. W., Porteous, D. J., Taylor, M. S., & Evans, K. L. (2014). SuRFing the genomics wave: An R package for prioritising SNPs by functionality. Genome Medicine.

- He, B., Chen, C., Teng, L. and Tan, K. (2014). Global view of enhancer-promoter interactome in human cells. Proceedings of the National Academy of Sciences of the United States of America, 111(21), E2191-9. https://doi.org/10.1073/pnas.1320308111
- ENCODE Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. Nature. https://doi.org/10.1038/nature11247
- Pleasance, E. D., Cheetham, R. K., Stephens, P. J., Mcbride, D. J., Sean, J., Greenman, C. D., … Stratton, M. R. (2009). A comprehensive catalogue of somatic mutations from a human cancer genome. Nature. https://doi.org/nature08658 [pii]10.1038/nature08658
- Chawla, N. V, Bowyer, K. W., Hall, L. O. and Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research, 16, 321–357.
- Roadmap Epigenomics Consortium, Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., … Kellis, M. (2015). Integrative analysis of 111 reference human epigenomes. Nature. https://doi.org/10.1038/nature14248
- Beck, T. F., Mullikin, J. C. and Biesecker, L. G. (2016). Systematic evaluation of sanger validation of next-generation sequencing variants. Clinical Chemistry. https://doi.org/10.1373/clinchem.2015.249623
- Iakoubova, O. A., Olsson, C. L., Dains, K. M., Ross, D. A., Andalibi, A., Lau, K., … West, D. B. (2001). Genome-tagged mice (GTM): two sets of genome-wide congenic strains. Genomics, 74(1), 89–104. https://doi.org/10.1006/geno.2000.6497
- A Guide to the Exome Aggregation Consortium (ExAC) Data Set | MacArthur Lab. (n.d.). Retrieved May 12, 2017, from https://macarthurlab.org/2014/11/18/a-guide-to-the-exome-aggregation-consortium-exac-data-set/
- Li, R., Li, Y., Kristiansen, K., & Wang, J. (2008). SOAP: short oligonucleotide alignment program. Bioinformatics, 24. https://doi.org/10.1093/bioinformatics/btn025
- Rockman, M. V, & Wray, G. A. (2002). Abundant raw material for cis-regulatory evolution in humans. Molecular Biology and Evolution, 19(11), 1991–2004. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/12411608
- Din, S., Wong, K., Mueller, M. F., Oniscu, A., Hewinson, J., Black, C. J., … Arends, M. J. (2018). Mutational analysis identifies therapeutic biomarkers in inflammatory bowel disease-associated colorectal cancers. Clinical Cancer Research. https://doi.org/10.1158/1078-0432.CCR-17-3713
- Batista, G. E. A. P. A., Prati, R. C., & Monard, M. C. (n.d.). A Study of the Behavior of Several Methods for Balancing Machine Learning Training Data.
- He, Y., Gorkin, D. U., Dickel, D. E., Nery, J. R., Castanon, R. G., Lee, A. Y., … Ohler, U. (n.d.). Improved regulatory element prediction based on tissue-specific local epigenomic signatures. https://doi.org/10.1073/pnas.1618353114
- Rashid, M., Robles-Espinoza, C. D., Rust, A. G., & Adams, D. J. (2013). Cake: A bioinformatics pipeline for the integrated analysis of somatic variants in cancer genomes. Bioinformatics. https://doi.org/10.1093/bioinformatics/btt371

- Varley, J. M. (2003). Germline TP53 mutations and Li-Fraumeni syndrome. Human Mutation. https://doi.org/10.1002/humu.10185

- Slaughter DP, Southwick HW, S. W. (1953). Field cancerization in oral stratified squamous epithelium; clinical implications of multicentric origin. Cancer, (6), 963–968.

- Mansour, M. R., Abraham, B. J., Anders, L., Berezovskaya, A., Gutierrez, A., Durbin, A. D., … Look, A. T. (2014). Oncogene regulation. An oncogenic super-enhancer formed through somatic mutation of a noncoding intergenic element. Science (New York, N.Y.), 346(6215), 1373–1377. https://doi.org/10.1126/science.1259037

- Pinto, E. M., Chen, X., Easton, J., Finkelstein, D., Liu, Z., Pounds, S., … Zambetti, G. P. (2015). Genomic landscape of paediatric adrenocortical tumours. Nature Communications, 6, 6302. https://doi.org/10.1038/ncomms7302

- Ewing, I., Hurley, J. J., Josephides, E., & Millar, A. (2014). The molecular genetics of colorectal cancer. Frontline Gastroenterology, 5(1), 26–30. https://doi.org/10.1136/flgastro-2013-100329

- Kinzler, K., Nilbert, M., Su, L., Vogelstein, B., Bryan, T., Levy, D., … et, al. (1991). Identification of FAP locus genes from chromosome 5q21. Science, 253(5020), 661–665. https://doi.org/10.1126/science.1651562

- Lynch, H. T., & Lynch, J. F. (n.d.). 25 years of HNPCC. Anticancer Research, 14(4B), 1617–1624. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/7979196

- Sampson, J. R., Dolwani, S., Jones, S., Eccles, D., Ellis, A., Evans, D. G., … Cheadle, J. P. (2003). Autosomal recessive colorectal adenomatous polyposis due to inherited mutations of MYH. The Lancet, 362(9377), 39–41. https://doi.org/10.1016/S0140-6736(03)13805-6

- Palles, C., Cazier, J.-B., Howarth, K. M., Domingo, E., Jones, A. M., Broderick, P., … Tomlinson, I. (2013). Germline mutations affecting the proofreading domains of POLE and POLD1 predispose to colorectal adenomas and carcinomas. Nature Genetics, 45(2), 136–144. https://doi.org/10.1038/ng.2503

- Heitzer, E., & Tomlinson, I. (2014). Replicative DNA polymerase mutations in cancer. Current Opinion in Genetics & Development, 24, 107–113. https://doi.org/10.1016/j.gde.2013.12.005

- Grover, S., Kastrinos, F., Steyerberg, E. W., Cook, E. F., Dewanwala, A., Burbidge, L. A., … Syngal, S. (2012). Prevalence and Phenotypes of APC and MUTYH Mutations in Patients With Multiple Colorectal Adenomas. JAMA, 308(5), 485–492. https://doi.org/10.1001/jama.2012.8780

- Cancer Genome Atlas Network. (2012). Comprehensive molecular characterization of human colon and rectal cancer. Nature, 487(7407), 330–337. https://doi.org/10.1038/nature11252

- Levy, D. B., Smith, K. J., Beazer-Barclay, Y., Hamilton, S. R., Vogelstein, B., & Kinzler, K. W. (1994). Inactivation of both APC alleles in human and mouse tumors. Cancer Research, 54(22), 5953–5958. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/7954428

- Patel, S. G., & Ahnen, D. J. (2012). Familial colon cancer syndromes: An update of a rapidly evolving field. Current Gastroenterology Reports. https://doi.org/10.1007/s11894-012-0280-6

- Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Aparicio, S. A. J. R., Behjati, S., Biankin, A. V., … Stratton, M. R. (2013). Signatures of mutational processes in human cancer. Nature, 500(7463), 415–421. https://doi.org/10.1038/nature12477

- Segditsas, S., Rowan, A. J., Howarth, K., Jones, A., Leedham, S., Wright, N. A., … Tomlinson, I. P. M. (2009). APC and the three-hit hypothesis. Oncogene, 28(1), 146–155. https://doi.org/10.1038/onc.2008.361

- Nieuwenhuis, M. H., Vogt, S., Jones, N., Nielsen, M., Hes, F. J., Sampson, J. R., … Vasen, H. F. A. (2012). Evidence for accelerated colorectal adenoma–carcinoma progression in MUTYH -associated polyposis? Gut, 61(5), 734–738. https://doi.org/10.1136/gut.2010.229104

- Vijaya Chandra, S. H., Wacker, I., Appelt, U. K., Behrens, J., & Schneikert, J. (2012). A Common Role for Various Human Truncated Adenomatous Polyposis Coli Isoforms in the Control of Beta-Catenin Activity and Cell Proliferation. PLoS ONE, 7(4), e34479. https://doi.org/10.1371/journal.pone.0034479

- Major, M. B., Camp, N. D., Berndt, J. D., Yi, X., Goldenberg, S. J., Hubbert, C., … Moon, R. T. (2007). Wilms Tumor Suppressor WTX Negatively Regulates WNT/ -Catenin Signaling. Science, 316(5827), 1043–1046. https://doi.org/10.1126/science/1141515

- Jenkins, Z. A., van Kogelenberg, M., Morgan, T., Jeffs, A., Fukuzawa, R., Pearl, E., … Robertson, S. P. (2009). Germline mutations in WTX cause a sclerosing skeletal dysplasia but do not predispose to tumorigenesis. Nature Genetics, 41(1), 95–100. https://doi.org/10.1038/ng.270

- Moisan, A., Rivera, M. N., Lotinun, S., Akhavanfard, S., Coffman, E. J., Cook, E. B., … Bardeesy, N. (2011). The WTX Tumor Suppressor Regulates Mesenchymal Progenitor Cell Fate Specification. Developmental Cell, 20(5), 583–596. https://doi.org/10.1016/j.devcel.2011.03.013

- Rivera, M. N., Kim, W. J., Wells, J., Stone, A., Burger, A., Coffman, E. J., … Haber, D. A. (2009). The tumor suppressor WTX shuttles to the nucleus and modulates WT1 activity. Proceedings of the National Academy of Sciences, 106(20), 8338–8343. https://doi.org/10.1073/pnas.0811349106

- Kim, W. J., Rivera, M. N., Coffman, E. J., & Haber, D. A. (2012). The WTX Tumor Suppressor Enhances p53 Acetylation by CBP/p300. Molecular Cell, 45(5), 587–597. https://doi.org/10.1016/j.molcel.2011.12.025

- Rivera, M. N., Kim, W. J., Wells, J., Driscoll, D. R., Brannigan, B. W., Han, M., … Haber, D. A. (2007). An X Chromosome Gene, WTX, Is Commonly Inactivated in Wilms Tumor. Science, 315(5812), 642–645. https://doi.org/10.1126/science.1137509

- Behjati, S., Tarpey, P. S., Sheldon, H., Martincorena, I., Van Loo, P., Gundem, G., … Campbell, P. J. (2014). Recurrent PTPRB and PLCG1 mutations in angiosarcoma. Nature Genetics, 46(4), 376–379. https://doi.org/10.1038/ng.2921

- Cardoso, J., Molenaar, L., de Menezes, R. X., van Leerdam, M., Rosenberg, C., Möslein, G., … Fodde, R. (2006). Chromosomal Instability in MYH - and APC -Mutant Adenomatous Polyps. Cancer Research, 66(5), 2514–2519. https://doi.org/10.1158/0008-5472.CAN-05-2407

- Grasso, F., Giacomini, E., Sanchez, M., Degan, P., Gismondi, V., Mazzei, F., … Bignami, M. (2014). Genetic instability in lymphoblastoid cell lines expressing biallelic and monoallelic variants in the human MUTYH gene. Human Molecular Genetics, 23(14), 3843–3852. https://doi.org/10.1093/hmg/ddu097

- Coffey, A. J., Kokocinski, F., Calafato, M. S., Scott, C. E., Palta, P., Drury, E., … Palotie, A. (2011). The GENCODE exome: sequencing the complete human exome. European Journal of Human Genetics, 19(7), 827–831. https://doi.org/10.1038/ejhg.2011.28

- Fischer, A., Illingworth, C. J., Campbell, P. J., & Mustonen, V. (2013). EMu: probabilistic inference of mutational processes and their localization in the cancer genome. Genome Biology, 14(4), R39. https://doi.org/10.1186/gb-2013-14-4-r39

- Fischer, A., Illingworth, C. J., Campbell, P. J., & Mustonen, V. (2013). EMu: probabilistic inference of mutational processes and their localization in the cancer genome. Genome Biology, 14(4), R39. https://doi.org/10.1186/gb-2013-14-4-r39

- Timmermann, B., Kerick, M., Roehr, C., Fischer, A., Isau, M., Boerno, S. T., … Schweiger, M. R. (2010). Somatic Mutation Profiles of MSI and MSS Colorectal Cancer Identified by Whole Exome Next Generation Sequencing and Bioinformatics Analysis. PLoS ONE, 5(12), e15661. https://doi.org/10.1371/journal.pone.0015661

- Nikolaev, S. I., Sotiriou, S. K., Pateras, I. S., Santoni, F., Sougioultzis, S., Edgren, H., … Halazonetis, T. D. (2012). A Single-Nucleotide Substitution Mutator Phenotype Revealed by Exome Sequencing of Human Colon Adenomas. Cancer Research, 72(23), 6279–6289. https://doi.org/10.1158/0008-5472.CAN-12-3869

- Coffey, A. J., Kokocinski, F., Calafato, M. S., Scott, C. E., Palta, P., Drury, E., … Palotie, A. (2011). The GENCODE exome: sequencing the complete human exome. European Journal of Human Genetics, 19(7), 827–831. https://doi.org/10.1038/ejhg.2011.28

- Jones, S., Lambert, S., Williams, G. T., Best, J. M., Sampson, J. R., & Cheadle, J. P. (2004). Increased frequency of the k-ras G12C mutation in MYH polyposis colorectal adenomas. British Journal of Cancer, 90(8), 1591–1593. https://doi.org/10.1038/sj.bjc.6601747

- Thomas, R. K., Baker, A. C., DeBiasi, R. M., Winckler, W., LaFramboise, T., Lin, W. M., … Garraway, L. A. (2007). High-throughput oncogene mutation profiling in human cancer. Nature Genetics, 39(3), 347–351. https://doi.org/10.1038/ng1975

- Coffey, A. J., Kokocinski, F., Calafato, M. S., Scott, C. E., Palta, P., Drury, E., … Palotie, A. (2011). The GENCODE exome: sequencing the complete human exome. European Journal of Human Genetics, 19(7), 827–831. https://doi.org/10.1038/ejhg.2011.28

- Nikolaev, S. I., Sotiriou, S. K., Pateras, I. S., Santoni, F., Sougioultzis, S., Edgren, H., … Halazonetis, T. D. (2012). A Single-Nucleotide Substitution Mutator Phenotype Revealed by Exome Sequencing of Human Colon Adenomas. Cancer Research, 72(23), 6279–6289. https://doi.org/10.1158/0008-5472.CAN-12-3869

- Timmermann, B., Kerick, M., Roehr, C., Fischer, A., Isau, M., Boerno, S. T., … Schweiger, M. R. (2010). Somatic Mutation Profiles of MSI and MSS Colorectal Cancer Identified by Whole Exome Next Generation Sequencing and Bioinformatics Analysis. PLoS ONE, 5(12), e15661. https://doi.org/10.1371/journal.pone.0015661

- Al-Tassan, N., Chmiel, N. H., Maynard, J., Fleming, N., Livingston, A. L., Williams, G. T., … Cheadle, J. P. (2002). Inherited variants of MYH associated with somatic G:C□T:A mutations in colorectal tumors. Nature Genetics, 30(2), 227–232. https://doi.org/10.1038/ng828

- Slupska, M. M., Baikalov, C., Luther, W. M., Chiang, J. H., Wei, Y. F., & Miller, J. H. (1996). Cloning and sequencing a human homolog (hMYH) of the Escherichia coli mutY gene whose function is required for the repair of oxidative DNA damage. Journal of Bacteriology, 178(13), 3885–3892. https://doi.org/10.1128/jb.178.13.3885-3892.1996

- Segditsas, S., & Tomlinson, I. (2006). Colorectal cancer and genetic alterations in the Wnt pathway. Oncogene, 25(57), 7531–7537. https://doi.org/10.1038/sj.onc.1210059

- Lowe, S. W., Cepero, E., & Evan, G. (2004). Intrinsic tumour suppression. Nature, 432(7015), 307–315. https://doi.org/10.1038/nature03098

- Cheadle, J. P., Krawczak, M., Thomas, M. W., Hodges, A. K., Al-Tassan, N., Fleming, N., & Sampson, J. R. (2002). Different combinations of biallelic APC mutation confer different growth advantages in colorectal tumours. Cancer Research, 62(2), 363–366. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/11809680

- Clevers, H., Loh, K. M., & Nusse, R. (2014). An integral program for tissue renewal and regeneration: Wnt signaling and stem cell control. Science, 346(6205), 1248012. https://doi.org/10.1126/science.1248012

- Moon, R. T., Bowerman, B., Boutros, M., & Perrimon, N. (2002). The Promise and Perils of Wnt Signaling Through beta -Catenin. Science, 296(5573), 1644–1646. https://doi.org/10.1126/science.1071549

- Valdar, W. (2006). Genome-wide genetic association of complex traits in heterogeneous stock mice. Nature Genet., 38. https://doi.org/10.1038/ng1840

- Pope, B. D., Ryba, T., Dileep, V., Yue, F., Wu, W., Denas, O., … Gilbert, D. M. (2014). Topologically associating domains are stable units of replication-timing regulation. Nature, 515(7527), 402–405. https://doi.org/10.1038/nature13986

- Yates, L. R., Gerstung, M., Knappskog, S., Desmedt, C., Gundem, G., Van Loo, P., … Campbell, P. J. (2015). Subclonal diversification of primary breast cancer revealed by multiregion sequencing. Nature Medicine. https://doi.org/10.1038/nm.3886

- Levitin, H. M., Yuan, J., & Sims, P. A. (2018). Single-Cell Transcriptomic Analysis of Tumor Heterogeneity. Trends in Cancer. https://doi.org/10.1016/j.trecan.2018.02.003

- Huttlin, E. L., Bruckner, R. J., Paulo, J. A., Cannon, J. R., Ting, L., Baltier, K., … Harper, J. W. (2017). Architecture of the human interactome defines protein communities and disease networks. Nature Publishing Group, 545. https://doi.org/10.1038/nature22366

- Lu, C., Zhang, J., Nagahawatte, P., Easton, J., Lee, S., Liu, Z., … Bahrami, A. (2015). The genomic landscape of childhood and adolescent melanoma. Journal of Investigative Dermatology. https://doi.org/10.1038/jid.2014.425

- Kakimi, K., Karasaki, T., Matsushita, H., & Sugie, T. (2017). Advances in personalized cancer immunotherapy. Breast Cancer. https://doi.org/10.1007/s12282-016-0688-1

- Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics, 25. https://doi.org/10.1093/bioinformatics/btp324

- Melton, C., Reuter, J. A., Spacek, D. V, & Snyder, M. (2015). Recurrent somatic mutations in regulatory regions of human cancer genomes. Nature Genetics, 47(7), 710–716. https://doi.org/10.1038/ng.3332

- Robbe, P., Popitsch, N., Knight, S. J. L., Antoniou, P., Becq, J., He, M., … Schuh, A. (2018). Clinical whole-genome sequencing from routine formalin-fixed, paraffin-embedded specimens: pilot study for the 100,000 Genomes Project. Genetics in Medicine. https://doi.org/10.1038/gim.2017.241

- Gerstein, M. B., Kundaje, A., Hariharan, M., Landt, S. G., Yan, K. K., Cheng, C., … Snyder, M. (2012). Architecture of the human regulatory network derived from ENCODE data. Nature, 489(7414), 91–100. https://doi.org/10.1038/nature11245

- Zhang, J., Baran, J., Cros, A., Guberman, J. M., Haider, S., Hsu, J., Kasprzyk, A. (2011). International cancer genome consortium data portal-a one-stop shop for cancer genomics data. Database. https://doi.org/10.1093/database/bar026

- Hudson, T. J., Anderson, W., Aretz, A., Barker, A. D., Bell, C., Bernabé, R. R., … Wainwright, B. J. (2010). International network of cancer genome projects. Nature. https://doi.org/10.1038/nature08987

- Leiserson, M. D. M., Vandin, F., Wu, H.-T., Dobson, J. R., Eldridge, J. V, Thomas, J. L., … Raphael, B. J. (2014). Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. Nature Genetics, 47(2), 106–114. https://doi.org/10.1038/ng.3168

- bam2fastq. [http://gsl.hudsonalpha.org/information/software/bam2fastq]. (n.d.). Retrieved from http://gsl.hudsonalpha.org/information/software/bam2fastq

- Epstein, D. J. (2009). Cis-regulatory mutations in human disease. Brief Funct Genomic Proteomic, 8(4), 310–316. https://doi.org/10.1093/bfgp/elp021

- Lu, Y., Yi, Y., Liu, P., Wen, W., James, M., Wang, D., & You, M. (2007). Common human cancer genes discovered by integrated gene-expression analysis. PloS One, 2(11), e1149. https://doi.org/10.1371/journal.pone.0001149

- Vinagre, J., Almeida, A., Populo, H., Batista, R., Lyra, J., Pinto, V., … Soares, P. (2013). Frequency of TERT promoter mutations in human cancers. Nat Commun, 4, 2185. https://doi.org/10.1038/ncomms3185

- Weinhold, N., Jacobsen, A., Schultz, N., Sander, C., & Lee, W. (2014). Genome-wide analysis of noncoding regulatory mutations in cancer. Nature Genetics, 46(11), 1160–1165. https://doi.org/10.1038/ng.3101

- Gonzalez-Perez, A., Perez-Llamas, C., Deu-Pons, J., Tamborero, D., Schroeder, M. P., Jene-Sanz, A., … Lopez-Bigas, N. (2013). IntOGen-mutations identifies cancer drivers across tumor types. Nature Methods. https://doi.org/10.1038/nmeth.2642

- Chan, K. C. A., Jiang, P., Zheng, Y. W. L., Liao, G. J. W., Sun, H., Wong, J., … Lo, Y. M. D. (2013). Cancer Genome Scanning in Plasma: Detection of Tumor-Associated Copy Number Aberrations, Single-Nucleotide Variants, and Tumoral Heterogeneity by Massively Parallel Sequencing. Clinical Chemistry, 59(1), 211–224. https://doi.org/10.1373/clinchem.2012.196014

- Sherry, S. T. (2001). dbSNP: the NCBI database of genetic variation. Nucleic Acids Research. https://doi.org/10.1093/nar/29.1.308

- Larrayoz, M., Rose-Zerilli, M. J. J., Kadalayil, L., Parker, H., Blakemore, S., Forster, J., … Strefford, J. C. (2016). Non-coding NOTCH1 mutations in chronic lymphocytic leukemia; their clinical impact in the UK CLL4 trial. Leukemia. https://doi.org/10.1038/leu.2016.298

- Illumina. (2015). An introduction to Next-Generation Sequencing Technology.

- Stroud, H., Feng, S., Morey Kinney, S., Pradhan, S., & Jacobsen, S. E. (2011). 5-Hydroxymethylcytosine is associated with enhancers and gene bodies in human embryonic stem cells. Genome Biology, 12(6), R54. https://doi.org/10.1186/gb-2011-12-6-r54

- Balakrishnan, A., Bleeker, F. E., Lamba, S., Rodolfo, M., Daniotti, M., Scarpa, A., … Bardelli, A. (2007). Novel Somatic and Germline Mutations in Cancer Candidate Genes in Glioblastoma, Melanoma, and Pancreatic Carcinoma. Cancer Research, 67(8), 3545–3550. https://doi.org/10.1158/0008-5472.CAN-07-0065

- Bergmann, E. A., Chen, B. J., Arora, K., Vacic, V., & Zody, M. C. (2016). Conpair: Concordance and contamination estimator for matched tumor-normal pairs. Bioinformatics. https://doi.org/10.1093/bioinformatics/btw389

- Jackson, S. E., & Chester, J. D. (2015). Personalised cancer medicine. International Journal of Cancer. https://doi.org/10.1002/ijc.28940

- Heitzer, E., Ulz, P., Belic, J., Gutschi, S., Quehenberger, F., Fischereder, K., … Speicher, M. R. (2013). Tumor-associated copy number changes in the circulation of patients with prostate cancer identified through whole-genome sequencing. Genome Medicine, 5(4), 30. https://doi.org/10.1186/gm434

- Alioto, T. S., Buchhalter, I., Derdak, S., Hutter, B., Eldridge, M. D., Hovig, E., … Gut, I. G. (2015). A comprehensive assessment of somatic mutation detection in cancer using whole-genome sequencing. Nature Communications. https://doi.org/10.1038/ncomms10001

- Persson, M., Andren, Y., Mark, J., Horlings, H. M., Persson, F., & Stenman, G. (2009). Recurrent fusion of MYB and NFIB transcription factor genes in carcinomas of the breast and head and neck. Proc Natl Acad Sci U S A, 106(44), 18740–18744. https://doi.org/10.1073/pnas.09

- Li, W., & Liu, M. (2011). Distribution of 5-hydroxymethylcytosine in different human tissues. Journal of Nucleic Acids, 2011, 870726. https://doi.org/10.4061/2011/870726

- Davis, J., & Goadrich, M. (2006). The relationship between Precision-Recall and ROC curves. In Proceedings of the 23rd international conference on Machine learning - ICML '06. https://doi.org/10.1145/1143844.1143874

- Nature Education. (2018). Eukaryotic Cells Possess a Nucleus and Membrane-Bound Organelles. In Essentials of Cell Biology. Nature Publishing Group. Retrieved from https://www.nature.com/scitable/ebooks/essentials-of-cell-biology-14749010/118237915

- Koboldt, D. C., Zhang, Q., Larson, D. E., Shen, D., McLellan, M. D., Lin, L., … Wilson, R. K. (2012). VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Research. https://doi.org/10.1101/gr.129684.111

- Porta-Pardo, E., Kamburov, A., Tamborero, D., Pons, T., Grases, D., Valencia, A., … Godzik, A. (2017). Comparison of algorithms for the detection of cancer drivers at subgene resolution. Nature Methods, 14(8), 782–788. https://doi.org/10.1038/nmeth.4364

- Zheng, H., Pomyen, Y., Hernandez, M. O., Li, C., Livak, F., Tang, W., … Wang, X. W. (2018). Single-cell analysis reveals cancer stem cell heterogeneity in hepatocellular carcinoma. Hepatology. https://doi.org/10.1002/hep.29778

- Ramezankhani, A., Pournik, O., Shahrabi, J., Azizi, F., Hadaegh, F., & Khalili, D. (2016). The impact of oversampling with SMOTE on the performance of 3 classifiers in prediction of type 2 diabetes. Medical Decision Making. https://doi.org/10.1177/0272989X14560647

- Horn, S., Figl, A., Rachakonda, P. S., Fischer, C., Sucker, A., Gast, A., … Kumar, R. (2013). TERT promoter mutations in familial and sporadic melanoma. Science. https://doi.org/10.1126/science.1230062

- Zhou, J., Troyanskaya, O. G. (2015). Predicting effects of noncoding variants with deep learning–based sequence model. Nature Methods, 12(10), 931–934. https://doi.org/10.1038/nmeth.3547

- Fortin, J.-P., & Hansen, K. D. (n.d.). Reconstructing A/B compartments as revealed by Hi-C using long-range correlations in epigenetic data. https://doi.org/10.1101/019000

- Araya, C. L., Cenik, C., Reuter, J. A., Kiss, G., Pande, V. S., Snyder, M. P., & Greenleaf, W. J. (2015). Identification of significantly mutated regions across cancer types highlights a rich landscape of functional molecular alterations. Nature Genetics, 48(2), 117–125. https://doi.org/10.1038/ng.3471

- Teer, J. K., Zhang, Y., Chen, L., Welsh, E. A., Cress, W. D., Eschrich, S. A., & Berglund, A. E. (2017). Evaluating somatic tumor mutation detection without matched normal samples. Human Genomics. https://doi.org/10.1186/s40246-017-0118-2

- Cibulskis, K., Lawrence, M. S., Carter, S. L., Sivachenko, A., Jaffe, D., Sougnez, C., … Getz, G. (2013). Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nature Biotechnology. https://doi.org/10.1038/nbt.2514

- Emami, N. C., Leong, L., Wan, E., Van Blarigan, E. L., Cooperberg, M. R., Tenggara, I., … Simko, J. P. (2017). Tissue Sources for Accurate Measurement of Germline DNA Genotypes in Prostate Cancer Patients Treated With Radical Prostatectomy. Prostate. https://doi.org/10.1002/pros.23283

- Kim, S. Y., & Speed, T. P. (2013). Comparing somatic mutation-callers: Beyond Venn diagrams. BMC Bioinformatics. https://doi.org/10.1186/1471-2105-14-189
- Maston, G. A., Evans, S. K., & Green, M. R. (2006). Transcriptional regulatory elements in the human genome. Annual Review of Genomics and Human Genetics, 7, 29–59. https://doi.org/10.1146/annurev.genom.7.080505.115623
- McLendon, R., Friedman, A., Bigner, D., Van Meir, E. G., Brat, D. J., Mastrogianakis, G. M., … Thomson, E. (2008). Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature. https://doi.org/10.1038/nature07385
- Pirooznia, M., Kramer, M., Parla, J., Goes, F. S., Potash, J. B., McCombie, W., … Sham, P. (2014). Validation and assessment of variant calling pipelines for next-generation sequencing. Human Genomics, 8(1), 14. https://doi.org/10.1186/1479-7364-8-14
- Brosens, R. P., Haan, J. C., Carvalho, B., Rustenburg, F., Grabsch, H., Quirke, P., … Ylstra, B. (2010). Candidate driver genes in focal chromosomal aberrations of stage II colon cancer. The Journal of Pathology, 221(4), n/a-n/a. https://doi.org/10.1002/path.2724
- Larson, D. E., Harris, C. C., Chen, K., Koboldt, D. C., Abbott, T. E., Dooling, D. J., … Ding, L. (2012). Somaticsniper: Identification of somatic point mutations in whole genome sequencing data. Bioinformatics. https://doi.org/10.1093/bioinformatics/btr665
- Ko, A. M., Tu, H. P., Liu, T. T., Chang, J. G., Yuo, C. Y., Chiang, S. L., … Ko, Y. C. (2013). ALPK1 genetic regulation and risk in relation to gout. Int J Epidemiol, 42(2), 466–474. https://doi.org/10.1093/ije/dyt028
- Brown, S. D., Warren, R. L., Gibb, E. A., Martin, S. D., Spinelli, J. J., Nelson, B. H., & Holt, R. A. (n.d.). Neo-antigens predicted by tumor genome meta-analysis correlate with increased patient survival. https://doi.org/10.1101/gr.165985.113
- Cibulskis, K., Lawrence, M. S., Carter, S. L., Sivachenko, A., Jaffe, D., Sougnez, C., … Getz, G. (2013). Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nature Biotechnology. https://doi.org/10.1038/nbt.2514
- Martincorena, I., Raine, K. M., Gerstung, M., Dawson, K. J., Haase, K., Van Loo, P., … Campbell, P. J. (2017). Universal Patterns of Selection in Cancer and Somatic Tissues. Cell. https://doi.org/10.1016/j.cell.2017.09.042
- Jon Kleinberg. (2003). An impossibility theorem for clustering. Nips. https://doi.org/10.1103/PhysRevE.90.062813
- Murphy, M. F. G., Bithell, J. F., Stiller, C. A., Kendall, G. M., & O'Neill, K. A. (2013). Childhood and adult cancers: Contrasts and commonalities. Maturitas. https://doi.org/10.1016/j.maturitas.2013.05.017
- Goya, R., Sun, M. G. F., Morin, R. D., Leung, G., Ha, G., Wiegand, K. C., … Shah, S. P. (2010). SNVMix: Predicting single nucleotide variants from next-generation sequencing of tumors. Bioinformatics. https://doi.org/10.1093/bioinformatics/btq040
- Lawson, D. A., Kessenbrock, K., Davis, R. T., Pervolarakis, N., Werb, Z. (2018). Tumour heterogeneity and metastasis at single-cell resolution. Nature Cell Biology. https://doi.org/10.1038/s41556-018-0236-7

- Schumacher, T. N., & Schreiber, R. D. (2015). Neoantigens in cancer immunotherapy. Science. https://doi.org/10.1126/science.aaa4971

- Wong, C. C., Martincorena, I., Rust, A. G., Rashid, M., Alifrangis, C., Alexandrov, L. B., … Adams, D. J. (2014). Inactivating CUX1 mutations promote tumorigenesis. Nature Genetics. https://doi.org/10.1038/ng.2846

- Perou, C. M., Jeffrey, S. S., van de Rijn, M., Rees, C. A., Eisen, M. B., Ross, D. T., … Botstein, D. (1999). Distinctive gene expression patterns in human mammary epithelial cells and breast cancers. Proceedings of the National Academy of Sciences. https://doi.org/10.2515/therapie:2008034

- Kim, S. Y. and Speed, T. P. (2013). Comparing somatic mutation-callers: Beyond Venn diagrams. BMC Bioinformatics. https://doi.org/10.1186/1471-2105-14-189

- O'Rawe, J., Jiang, T., Sun, G., Wu, Y., Wang, W., Hu, J. and Lyon, G. J. (2013). Low concordance of multiple variant-calling pipelines: Practical implications for exome and genome sequencing. Genome Medicine. https://doi.org/10.1186/gm432

- Xu, H., DiCarlo, J., Satya, R. V., Peng, Q., & Wang, Y. (2014). Comparison of somatic mutation calling methods in amplicon and whole exome sequence data. BMC Genomics. https://doi.org/10.1186/1471-2164-15-244

- Polak, P., Karlić, R., Koren, A., Thurman, R., Sandstrom, R., Lawrence, M. S., … Sunyaev, S. R. (2015). Cell-of-origin chromatin organization shapes the mutational landscape of cancer. Nature, 518(7539), 360–364. https://doi.org/10.1038/nature14221

- Li, H. (2012). Exploring single-sample SNP and INDEL calling with whole-genome de novo assembly. Bioinformatics, 28. https://doi.org/10.1093/bioinformatics/bts280

- Ryan, N. M., Morris, S. W., Porteous, D. J., Taylor, M. S., & Evans, K. L. (2014). SuRFing the genomics wave: An R package for prioritising SNPs by functionality. Genome Medicine. https://doi.org/10.1186/s13073-014-0079-1

- Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., … 1000 Genome Project Data Processing Subgroup. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics (Oxford, England), 25(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

- Bobisse, S., Genolet, R., Roberti, A., Tanyi, J. L., Racle, J., Stevenson, B. J., … Harari, A. (2018). Sensitive and frequent identification of high avidity neo-epitope specific CD8+T cells in immunotherapy-naive ovarian cancer. Nature Communications. https://doi.org/10.1038/s41467-018-03301-0

- Muse, S. V., & Gaut, B. S. (1997). Comparing patterns of nucleotide substitution rates among chloroplast loci using the relative ratio test. Genetics.

- Rabbie, R., Rashid, M., Arance, A. M., Sánchez, M., Tell-Marti, G., Potrony, M., … Adams, D. J. (2017). Genomic analysis and clinical management of adolescent cutaneous melanoma. Pigment Cell and Melanoma Research. https://doi.org/10.1111/pcmr.12574

- Guichard, C., Amaddeo, G., Imbeaud, S., Ladeiro, Y., Pelletier, L., Maad, I. Ben, … Zucman-Rossi, J. (2012). Integrated analysis of somatic mutations and focal copy-number changes identifies key genes and pathways in hepatocellular carcinoma. Nature Genetics. https://doi.org/10.1038/ng.2256
- Yip, K. Y., Cheng, C., Bhardwaj, N., Brown, J. B., Leng, J., Kundaje, A., … Gerstein, M. (2012). Classification of human genomic regions based on experimentally determined binding sites of more than 100 transcription-related factors. Genome Biology. https://doi.org/10.1186/gb-2012-13-9-r48
- Hodis, E., Watson, I. R., Kryukov, G. V., Arold, S. T., Imielinski, M., Theurillat, J.-P., … Chin, L. (2012). A Landscape of Driver Mutations in Melanoma. Cell, 150(2), 251–263. https://doi.org/10.1016/j.cell.2012.06.024
- Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R. M., Ozenberger, B. A., Ellrott, K. and Kling, T. (2013). The cancer genome atlas pan-cancer analysis project. Nature Genetics. https://doi.org/10.1038/ng.2764
- Perera, D., Poulos, R. C., Shah, A., Beck, D., Pimanda, J. E., & Wong, J. W. H. (2016). Differential DNA repair underlies mutation hotspots at active promoters in cancer genomes. Nature. https://doi.org/10.1038/nature17437
- Lauss, M., Donia, M., Harbst, K., Andersen, R., Mitra, S., Rosengren, F., … Jönsson, G. (2017). Mutational and putative neoantigen load predict clinical benefit of adoptive T cell therapy in melanoma. Nature Communications. https://doi.org/10.1038/s41467-017-01460-0
- Collin, F., Ning, Y., Phillips, T., McCarthy, E., Scott, A., Ellison, C., … Levy, S. (2018). Detection of early stage pancreatic cancer using 5-hydroxymethylcytosine signatures in circulating cell free DNA. BioRxiv, 422675. https://doi.org/10.1101/422675
- Dagogo-Jack, I., & Shaw, A. T. (2018). Tumour heterogeneity and resistance to cancer therapies. Nature Reviews Clinical Oncology. https://doi.org/10.1038/nrclinonc.2017.166
- Spiradenoma: Background, Pathophysiology, Epidemiology. (n.d.). Retrieved July 28, 2017, from http://emedicine.medscape.com/article/1062079-overview
- White, M. A., Ané, C., Dewey, C. N., Larget, B. R., & Payseur, B. A. (2009). Fine-scale phylogenetic discordance across the house mouse genome. PLoS Genetics, 5(11), e1000729. https://doi.org/10.1371/journal.pgen.1000729
- Wang, S. J., Tu, H. P., Ko, A. M., Chiang, S. L., Chiou, S. J., Lee, S. S., … Ko, Y. C. (2011). Lymphocyte alpha-kinase is a gout-susceptible gene involved in monosodium urate monohydrate-induced inflammatory responses. J Mol Med (Berl), 89(12), 1241–1251. https://doi.org/10.1007/s00109-011-0796-5
- Shihab, H. A., Rogers, M. F., Gough, J., Mort, M., Cooper, D. N., Day, I. N. M., … Campbell, C. (2015). An integrative approach to predicting the functional effects of non-coding and coding sequence variation. Bioinformatics. https://doi.org/10.1093/bioinformatics/btv009
- Gerstein, M. B., Kundaje, A., Hariharan, M., Landt, S. G., Yan, K. K., Cheng, C., … Dobrovic, A. (2015). Architecture of the human regulatory network derived from ENCODE data. Nature, 7(1), 91–100. https://doi.org/10.1038/nature11245

- Nik-Zainal, S., Davies, H., Staaf, J., Ramakrishna, M., Glodzik, D., Zou, X., … Stratton, M. R. (2016). Landscape of somatic mutations in 560 breast cancer whole-genome sequences. Nature, 534(7605), 47–54. https://doi.org/10.1038/nature17676

- Miller, A., Asmann, Y., Cattaneo, L., Braggio, E., Keats, J., Auclair, D., … Stewart, A. K. (2017). High somatic mutation and neoantigen burden are correlated with decreased progression-free survival in multiple myeloma. Blood Cancer Journal. https://doi.org/10.1038/bcj.2017.94

- Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Aparicio, S. A. J. R., Behjati, S., Biankin, A. V., … Stratton, M. R. (2013). Signatures of mutational processes in human cancer. Nature, 500(7463), 415–421. https://doi.org/10.1038/nature12477

- Favero, F., Joshi, T., Marquard, A. M., Birkbak, N. J., Krzystanek, M., Li, Q., … Eklund, A. C. (2015). Sequenza: Allele-specific copy number and mutation profiles from tumor sequencing data. Annals of Oncology. https://doi.org/10.1093/annonc/mdu479

- Chang, K., Creighton, C. J., Davis, C., Donehower, L., Drummond, J., Wheeler, D., … Stuart, J. M. (2013). The Cancer Genome Atlas Pan-Cancer analysis project. Nature Genetics. https://doi.org/10.1038/ng.2764

- Moss, J., Magenheim, J., Neiman, D., Zemmour, H., Loyfer, N., Korach, A., … Dor, Y. (2018). Comprehensive human cell-type methylation atlas reveals origins of circulating cell-free DNA in health and disease. Nature Communications, 9(1), 5068. https://doi.org/10.1038/s41467-018-07466-6

- Snyder, A., Makarov, V., Merghoub, T., Yuan, J., Zaretsky, J. M., Desrichard, A., … Chan, T. A. (2014). Genetic Basis for Clinical Response to CTLA-4 Blockade in Melanoma. New England Journal of Medicine. https://doi.org/10.1056/NEJMoa1406498

- Tomek, I. (1976). EXPERIMENT WITH THE EDITED NEAREST-NEIGHBOR RULE. IEEE Transactions on Systems, Man and Cybernetics. https://doi.org/10.1109/TSMC.1976.4309523

- Fischer, A., Illingworth, C. J. R., Campbell, P. J., & Mustonen, V. (2013). EMu: Probabilistic inference of mutational processes and their localization in the cancer genome. Genome Biology. https://doi.org/10.1186/gb-2013-14-4-r39

- Robles-Espinoza, C. D., Harland, M., Ramsay, A. J., Aoude, L. G., Quesada, V., Ding, Z., … Adams, D. J. (2014). POT1 loss-of-function variants predispose to familial melanoma. Nature Genetics. https://doi.org/10.1038/ng.2947

- He, B., Chen, C., Teng, L., & Tan, K. (n.d.). IM-PET: Integrated Methods for Predicting Enhancer Targets. Retrieved from http://tanlab4generegulation.org/IM-PET.html

- Scacheri, C. A., & Scacheri, P. C. (2015). Mutations in the noncoding genome. Current Opinion in Pediatrics. https://doi.org/10.1097/MOP.0000000000000283

- Gehring, J. S., Fischer, B., Lawrence, M., & Huber, W. (2015). SomaticSignatures: Inferring mutational signatures from single-nucleotide variants. Bioinformatics. https://doi.org/10.1093/bioinformatics/btv408

- Grau, J., Grosse, I., & Keilwagen, J. (n.d.). PRROC: computing and visualizing precision-recall and receiver operating characteristic curves in R. Retrieved from https://cran.r-project.org/web/packages/PRROC/vignettes/PRROC.pdf
- Rivera, M. N., Woo, J. K., Wells, J., Driscoll, D. R., Brannigan, B. W., Han, M., … Haber, D. A. (2007). An X chromosome gene, WTX, is commonly inactivated in wilms tumor. Science. https://doi.org/10.1126/science.1137509
- Marino-Enriquez, A., & Fletcher, C. D. M. (2014). Shouldn't we care about the biology of benign tumours? Nature Reviews Cancer. https://doi.org/10.1038/nrc3845
- Lawrence, M. S., Stojanov, P., Polak, P., Kryukov, G. V., Cibulskis, K., Sivachenko, A., … Getz, G. (2013). Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature. https://doi.org/10.1038/nature12213
- Ritchie, G. R. S., Dunham, I., Zeggini, E., & Flicek, P. (2014). Functional annotation of noncoding sequence variants. Nature Methods, 11(3), 294–296. https://doi.org/10.1038/nmeth.2832
- Atkin, W. S., & Saunders, B. P. (2002). Surveillance guidelines after removal of colorectal adenomatous polyps. Gut. https://doi.org/10.1136/gut.51.suppl$_5$.$v6$
- Tsao, H., Bevona, C., Goggins, W., & Quinn, T. (2003). The transformation rate of moles (melanocytic nevi) into cutaneous melanoma: A population-based estimate. Archives of Dermatology. https://doi.org/10.1001/archderm.139.3.282
- Pathology Outlines - Skin tumor Nonmelanocytic. (n.d.). Retrieved July 28, 2017, from http://www.pathologyoutlines.com/skintumornonmelanocytic.html
- Ideker, T., & Krogan, N. J. (2012). Differential network biology. Molecular Systems Biology. https://doi.org/10.1038/msb.2011.99
- Rosenthal, R., McGranahan, N., Herrero, J., Taylor, B. S., & Swanton, C. (2016). deconstructSigs: Delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution. Genome Biology. https://doi.org/10.1186/s13059-016-0893-4
- Puente, X. S., Beà, S., Valdés-Mas, R., Villamor, N., Gutiérrez-Abril, J., Martín-Subero, J. I., … Campo, E. (2015). Non-coding recurrent mutations in chronic lymphocytic leukaemia. Nature, 526(7574), 519–524. https://doi.org/10.1038/nature14666
- Krijgsman, O., Carvalho, B., Meijer, G. A., Steenbergen, R. D. M., & Ylstra, B. (2014). Focal chromosomal copy number aberrations in cancer—Needles in a genome haystack. Biochimica et Biophysica Acta (BBA) - Molecular Cell Research, 1843(11), 2698–2704. https://doi.org/10.1016/j.bbamcr.2014.08.001
- Dakubo, G. D., Jakupciak, J. P., Birch-Machin, M. A., & Parr, R. L. (2007). Clinical implications and utility of field cancerization. Cancer Cell International. https://doi.org/10.1186/1475-2867-7-2

- Jaiswal, G., Jaiswal, S., Kumar, R., Sharma, A. (2013). Field cancerization: concept and clinical implications in head and neck squamous cell carcinoma. Journal of Experimental Therapeutics Oncology.

- Lu, Q., Hu, Y., Sun, J., Cheng, Y., Cheung, K. H., & Zhao, H. (2015). A statistical framework to predict functional non-coding regions in the human genome through integrated analysis of annotation data. Sci Rep, 5, 10576. https://doi.org/10.1038/srep10576

- Slaghter, D. P., Southwick, H., & Smejkal, W. (1953). Field canerisation in oral stratified squamous epithelium. Cancer. https://doi.org/10.1002/1097-0142

- Breiman, L. (2001). Randomforest2001. Machine Learning. https://doi.org/10.1017/CBO9781107415324.004

- Nik-Zainal, S., Van Loo, P., Wedge, D. C., Alexandrov, L. B., Greenman, C. D., Lau, K. W., … Campbell, P. J. (2012). The Life History of 21 Breast Cancers. Cell. https://doi.org/10.1016/j.cell.2012.04.023

- Khurana, E., Fu, Y., Colonna, V., Mu, X. J., Kang, H. M., Lappalainen, T., … Gerstein, M. (2013). Integrative Annotation of Variants from 1092 Humans: Application to Cancer Genomics. Science, 342(6154), 1235587–1235587. https://doi.org/10.1126/science.1235587

- Brodsky, R. A., & Jones, R. J. (2004). Riddle: What do aplastic anemia, acute promyelocytic leukemia, and chronic myeloid leukemia have in common? Leukemia. https://doi.org/10.1038/sj.leu.2403487

- Chawla, N. V, Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research, 16, 321–357.

- biobambam. [https://github.com/gt1/biobambam]. (n.d.). Retrieved from https://github.com/gt1/biobambam

- Shen, L., Kondo, Y., Rosner, G. L., Xiao, L., Hernandez, N. S., Vilaythong, J., … Issa, J. P. J. (2005). MGMT promoter methylation and field defect in sporadic colorectal cancer. Journal of the National Cancer Institute. https://doi.org/10.1093/jnci/dji275

- Knudson, A. G. (1971). Mutation and Cancer: Statistical Study of Retinoblastoma. Proceedings of the National Academy of Sciences. https://doi.org/10.1073/pnas.68.4.820

- Fehr, A., Kovacs, A., Loning, T., Frierson Jr., H., van den Oord, J., & Stenman, G. (2011). The MYB-NFIB gene fusion-a novel genetic link between adenoid cystic carcinoma and dermal cylindroma. J Pathol, 224(3), 322–327. https://doi.org/10.1002/path.2909

- Miki, Y., Swensen, J., Shattuck-Eidens, D., Futreal, P. A., Harshman, K., Tavtigian, S., … Skolnick, M. H. (1994). A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1. Science. https://doi.org/10.1126/science.7545954

- He, C., Holme, J., & Anthony, J. (2014). SNP genotyping: The KASP assay. Methods in Molecular Biology. https://doi.org/10.1007/978-1-4939-0446-47

- McFarland, C., Yaglom, J. A., Wojtkowiak, J. W., Scott, J., Morse, D. L., Sherman, M. Y., & Mirny, L. (2017). The damaging effect of passenger mutations on cancer progression. Cancer Research. https://doi.org/10.15672/HJMS.2017.417

- Harrow, J., Frankish, A., Gonzalez, J. M., Tapanari, E., Diekhans, M., Kokocinski, F., … Hubbard, T. J. (2012). GENCODE: The reference human genome annotation for the ENCODE project. Genome Research. https://doi.org/10.1101/gr.135350.111

- Dees, N. D., Zhang, Q., Kandoth, C., Wendl, M. C., Schierding, W., Koboldt, D. C., … Ding, L. (2012). MuSiC: Identifying mutational significance in cancer genomes. Genome Research. https://doi.org/10.1101/gr.134635.111

- Milholland, B., Dong, X., Zhang, L., Hao, X., Suh, Y., & Vijg

- Degner, J. F., Pai, A. A., Pique-Regi, R., Veyrieras, J.-B., Gaffney, D. J., Pickrell, J. K., … Pritchard, J. K. (2012). DNase I sensitivity QTLs are a major determinant of human expression variation. Nature, 482(7385), 390–394. https://doi.org/10.1038/nature10808

- Ng, P. C., & Henikoff, S. (2003). SIFT: Predicting amino acid changes that affect protein function. Nucleic Acids Research. https://doi.org/10.1093/nar/gkg509

- Konnick, E. Q., & Pritchard, C. C. (2016). Germline, hematopoietic, mosaic, and somatic variation: Interplay between inherited and acquired genetic alterations in disease assessment. Genome Medicine. https://doi.org/10.1186/s13073-016-0350-8 , J. (2017). Differences between germline and somatic mutation rates in humans and mice. Nature Communications. https://doi.org/10.1038/ncomms15183

- Muse, S. V., & Gaut, B. S. (1997). Comparing patterns of nucleotide substitution rates among chloroplast loci using the relative ratio test. Genetics.

- Rabbie, R., Rashid, M., Arance, A. M., Sánchez, M., Tell-Marti, G., Potrony, M., … Adams, D. J. (2017). Genomic analysis and clinical management of adolescent cutaneous melanoma. Pigment Cell and Melanoma Research. https://doi.org/10.1111/pcmr.12574

- Peto J, Collins N, Barfoot R, Seal S, Warren W, Rahman N, Easton DF, Evans C, Deacon J, S. M. (1999). Prevalence of BRCA1 and BRCA2 gene mutations in patients with early-onset breast cancer. J Natl Cancer Inst, 91(11), 943–949.

- Kanchi, K. L., Johnson, K. J., Lu, C., McLellan, M. D., Leiserson, M. D. M., Wendl, M. C., … Ding, L. (2014). Integrated analysis of germline and somatic variants in ovarian cancer. Nature Communications. https://doi.org/10.1038/ncomms4156

- Tischler, G., & Leonard, S. (2014). biobambam: tools for read pair collation based algorithms on BAM files. Source Code for Biology and Medicine, 9(1), 13.
https://doi.org/10.1186/1751-0473-9-13

- J van der Horst, M. P., Marusic, Z., Hornick, J. L., Luzar, B., & Brenn, T. (2015). Morphologically low-grade spiradenocarcinoma: a clinicopathologic study of 19 cases with emphasis on outcome and MYB expression. Modern Pathology, 28, 944–953. https://doi.org/10.1038/modpathol.2015.48

- Cooper, D. N. (2002). Human gene mutation in pathology and evolution. Journal of Inherited Metabolic Disease. https://doi.org/10.1023/A:1015621710660

- Katainen, R., Dave, K., Pitkänen, E., Palin, K., Kivioja, T., Välimäki, N., … Aaltonen, L. A. (2015). CTCF/cohesin-binding sites are frequently mutated in cancer. Nature Genetics. https://doi.org/10.1038/ng.3335

- Lopes-Ramos, C. M., Barros, B. P., Koyama, F. C., Carpinetti, P. A., Pezuk, J., Doimo, N. T. S., … Parmigiani, R. B. (2017). E2F1 somatic mutation within miRNA target site impairs gene regulation in colorectal cancer. PLoS ONE.
https://doi.org/10.1371/journal.pone.0181153

- Bailey, S. D., Desai, K., Kron, K. J., Mazrooei, P., Sinnott-Armstrong, N. A., Treloar, A. E., … Lupien, M. (2016). Noncoding somatic and inherited single-nucleotide variants converge to promote ESR1 expression in breast cancer. Nature Genetics, 48(10). https://doi.org/10.1038/ng.3650

- Esteller, M. (2007). Cancer epigenomics: DNA methylomes and histone-modification maps. Nature Reviews Genetics. https://doi.org/10.1038/nrg2005

- Cooper, D. N. (2002). Human gene mutation in pathology and evolution. Journal of Inherited Metabolic Disease. https://doi.org/10.1023/A:1015621710660

- Bhattacharya, A., & Cui, Y. (2016). SomamiR 2.0: A database of cancer somatic mutations altering microRNA-ceRNA interactions. Nucleic Acids Research. https://doi.org/10.1093/nar/gkv1220

- Hudson, T. J., Anderson, W., Aretz, A., Barker, A. D., Bell, C., Bernabé, R. R., … Wainwright, B. J. (2010). International network of cancer genome projects. Nature. https://doi.org/10.1038/nature08987

- Leiserson, M. D. M., Vandin, F., Wu, H.-T., Dobson, J. R., Eldridge, J. V, Thomas, J. L., … Raphael, B. J. (2014). Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. Nature Genetics, 47(2), 106–114. https://doi.org/10.1038/ng.3168

- Kircher, M., Witten, D. M., Jain, P., O, B. J., Cooper, G. M., & Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. Nature Publishing Group, 46. https://doi.org/10.1038/ng.2892

- Stephens, P. J., Tarpey, P. S., Davies, H., Van Loo, P., Greenman, C., Wedge, D. C., … Stratton, M. R. (2012). The landscape of cancer genes and mutational processes in breast cancer. Nature, 486(7403), 400–404. https://doi.org/10.1038/nature11017

- Ikeda, Y., Kiyotani, K., Yew, P. Y., Kato, T., Tamura, K., Yap, K. L., … Grogan, R. H. (2016). Germline PARP4 mutations in patients with primary thyroid and breast cancers. Endocrine-Related Cancer, 23(3), 171–179. https://doi.org/10.1530/ERC-15-0359

- Khurana, E., Fu, Y., Colonna, V., Mu, X. J., Kang, H. M., Lappalainen, T., … Gerstein, M. (2013). Integrative Annotation of Variants from 1092 Humans: Application to Cancer Genomics. Science, 342(6154), 1235587–1235587. https://doi.org/10.1126/science.1235587

- Shihab, H. A., Gough, J., Cooper, D. N., Day, I. N. M., & Gaunt, T. R. (2013). Predicting the functional consequences of cancer-associated amino acid substitutions. Bioinformatics, 29(12), 1504–1510. https://doi.org/10.1093/bioinformatics/btt182

- Ryan, N. M., Morris, S. W., Porteous, D. J., Taylor, M. S., Evans, K. L., Manolio, T., … Lunter, G. (2014). SuRFing the genomics wave: an R package for prioritising SNPs by functionality. Genome Medicine, 6(10), 79. https://doi.org/10.1186/s13073-014-0079-1

- Reva, B., Antipin, Y., & Sander, C. (2011). Predicting the functional impact of protein mutations: Application to cancer genomics. Nucleic Acids Research. https://doi.org/10.1093/nar/gkr407

- Perou, C. M., Jeffrey, S. S., van de Rijn, M., Rees, C. A., Eisen, M. B., Ross, D. T., … Botstein, D. (1999). Distinctive gene expression patterns in human mammary epithelial cells and breast cancers. Proceedings of the National Academy of Sciences. https://doi.org/10.2515/therapie:2008034

- Weston, J., Leslie, C., Le, E., Zhou, D., Elisseeff, A., & Noble, W. S. (2005). Semi-supervised protein classification using cluster kernels. Bioinformatics. https://doi.org/10.1093/bioinformatics/bti497

- Zhou, J., & Troyanskaya, O. G. (2015). Predicting effects of noncoding variants with deep learning-based sequence model. Nature Methods. https://doi.org/10.1038/nmeth.3547

- Quang, D., Chen, Y., & Xie, X. (2015). DANN: A deep learning approach for annotating the pathogenicity of genetic variants. Bioinformatics. https://doi.org/10.1093/bioinformatics/btu703

- Ramezankhani, A., Pournik, O., Shahrabi, J., Azizi, F., Hadaegh, F., & Khalili, D. (2016). The impact of oversampling with SMOTE on the performance of 3 classifiers in prediction of type 2 diabetes. Medical Decision Making. https://doi.org/10.1177/0272989X14560647
- Stephens, P. J., Tarpey, P. S., Davies, H., Van Loo, P., Greenman, C., Wedge, D. C., … Stratton, M. R. (2012). The landscape of cancer genes and mutational processes in breast cancer. Nature, 486(7403), 400–404. https://doi.org/10.1038/nature11
- Howlader, N., Noone, A., Krapcho, M., Miller, D., Bishop, K., Altekruse, S., … Cronin, K. (2013). SEER Cancer Statistics Review 1975-2013. National Cancer Institute. Neier, M., Pappo, A., & Navid, F. (2012). Management of melanomas in children and young adults. Journal of Pediatric Hematology/Oncology. https://doi.org/10.1097/MPH.0b013e31824e3852
- Fitzpatrick, T. B. (1988). The Validity and Practicality of Sun-Reactive Skin Types I Through VI. Archives of Dermatology. https://doi.org/10.1001/archderm.1988.016
- Cannon-Albright, L. A., Goldgar, D. E., Meyer, L. J., Lewis, C. M., Anderson, D. E., Fountain, J. W., … Skolnick, M. H. (1992). Assignment of a locus for familial melanoma, MLM, to chromosome 9p13-p22. Science. https://doi.org/10.1126/science.143
- Cordoro, K. M., Gupta, D., Frieden, I. J., McCalmont, T., & Kashani-Sabet, M. (2013). Pediatric melanoma: Results of a large cohort study and proposal for modified ABCD detection criteria for children. Journal of the American Academy of Dermatology. https://doi.org/10.1016/j.jaad.2012.12.953
- Anderson, W. F., Pfeiffer, R. M., Tucker, M. A., & Rosenberg, P. S. (2009). Divergent cancer pathways for early-onset and late-onset cutaneous malignant melanoma. Cancer. https://doi.org/10.1002/cncr.24481
- Reed, D., Kudchadkar, R., Zager, J. S., Sondak, V. K., & Messina, J. L. (2013). Controversies in the evaluation and management of atypical melanocytic proliferations in children, adolescents, and young adults. JNCCN Journal of the National Comprehensive Cancer Network. https://doi.org/10.6004/jnccn.2013.0087
- Bett, B. J. (2006). Large or multiple congenital melanocytic nevi: Occurrence of neurocutaneous melanocytosis in 1008 persons. Journal of the American Academy of Dermatology. https://doi.org/10.1016/j.jaad.2005.10.040
- Austin, M. T., Xing, Y., Hayes-Jordan, A. A., Lally, K. P., & Cormier, J. N. (2013). Melanoma incidence rises for children and adolescents: An epidemiologic review of pediatric melanoma in the United States. Journal of Pediatric Surgery. https://doi.org/10.1016/j.jpedsurg.2013.06.002
- Lorimer, P. D., White, R. L., Walsh, K., Han, Y., Kirks, R. C., Symanowski, J., … Hill, J. S. (2016). Pediatric and Adolescent Melanoma: A National Cancer Data Base Update. Annals of Surgical Oncology. https://doi.org/10.1245/s10434-016-5349-2

- Pinto, E. M., Chen, X., Easton, J., Finkelstein, D., Liu, Z., Pounds, S., … Zambetti, G. P. (2015). Genomic landscape of paediatric adrenocortical tumours. Nature Communications, 6, 6302. https://doi.org/10.1038/ncomms7302
- Van Allen, E. M., Miao, D., Schilling, B., Shukla, S. A., Blank, C., Zimmer, L., … Garraway, L. A. (2015). Genomic correlates of response to CTLA-4 blockade in metastatic melanoma. Science. https://doi.org/10.1126/science.aad0095
- Shakhova, O., Zingg, D., Schaefer, S. M., Hari, L., Civenni, G., Blunschi, J., … Sommer, L. (2012). Sox10 promotes the formation and maintenance of giant congenital naevi and melanoma. Nature Cell Biology. https://doi.org/10.1038/ncb2535
- Valverde, P., Healy, E., Jackson, I., Rees, J. L., & Thody, A. J. (1995). Variants of the melanocyte–stimulating hormone receptor gene are associated with red hair and fair skin in humans. Nature Genetics. https://doi.org/10.1038/ng1195-328
- Lu, C., Zhang, J., Nagahawatte, P., Easton, J., Lee, S., Liu, Z., … Bahrami, A. (2015). The genomic landscape of childhood and adolescent melanoma. Journal of Investigative Dermatology. https://doi.org/10.1038/jid.2014.425
- Granter, S. R., Seeger, K., Calonje, E., Busam, K., & McKee, P. H. (2000). Malignant eccrine spiradenoma (spiradenocarcinoma). A clinicopathologic study of 12 cases. American Journal of Dermatopathology. https://doi.org/10.1097/00000372-200004000-00002
- Young, A. L., Kellermayer, R., Szigeti, R., Tészás, A., Azmi, S., & Celebi, J. T. (2006). CYLD mutations underlie Brooke-Spiegler, familial cylindromatosis, and multiple familial trichoepithelioma syndromes. Clinical Genetics. https://doi.org/10.1111/j.1399-0004.2006.00667.x
- Persson, M., Andren, Y., Mark, J., Horlings, H. M., Persson, F., & Stenman, G. (2009). Recurrent fusion of MYB and NFIB transcription factor genes in carcinomas of the breast and head and neck. Proceedings of the National Academy of Sciences. https://doi.org/10.1073/pnas.0909114106
- Fehr, A., Kovács, A., Löning, T., Frierson, H. F., Van Den Oord, J. J., & Stenman, G. (2011). The MYB-NFIB gene fusion-a novel genetic link between adenoid cystic carcinoma and dermal cylindroma. Journal of Pathology. https://doi.org/10.1002/path.2909
- Bignell, G. R., Warren, W., Seal, S., Takahashi, M., Rapley, E., Barfoot, R., … Stratton, M. R. (2000). Identification of the familial cylindromatosis tumour-suppressor gene. Nature Genetics. https://doi.org/10.1038/76006
- Kazakov, D. V., Zelger, B., Rütten, A., Vazmitel, M., Spagnolo, D. V., Kacerovska, D., … Michal, M. (2009). Morphologic diversity of malignant neoplasms arising in preexisting spiradenoma, cylindroma, and spiradenocylindroma based on the study of 24 cases, sporadic or occurring in the setting of brooke-spiegler syndrome. American Journal of Surgical Pathology. https://doi.org/10.1097/PAS.0b013e31819667

- Dai, B., Kong, Y. Y., Cai, X., Shen, X. X., & Kong, J. C. (2014). Spiradenocarcinoma, cylindrocarcinoma and spiradenocylindrocarcinoma: A clinicopathological study of nine cases. Histopathology. https://doi.org/10.1111/his.12448
- Van Der Horst, M. P. J., Marusic, Z., Hornick, J. L., Luzar, B., & Brenn, T. (2015). Morphologically low-grade spiradenocarcinoma: A clinicopathologic study of 19 cases with emphasis on outcome and MYB expression. Modern Pathology. https://doi.org/10.1038/modpathol.2015.48
- Singh, D. D., Naujoks, C., Depprich, R., Schulte, K. W., Jankowiak, F., Kübler, N. R., & Handschel, J. (2013). Cylindroma of head and neck: Review of the literature and report of two rare cases. Journal of Cranio-Maxillofacial Surgery. https://doi.org/10.1016/j.jcms.2012.11.016
- Gundem, G., Perez-Llamas, C., Jene-Sanz, A., Kedzierska, A., Islam, A., Deu-Pons, J., … Lopez-Bigas, N. (2010). IntOGen: integration and data mining of multidimensional oncogenomic data. Nature Methods. https://doi.org/10.1038/nmeth0210-92
- Liao, H. F., Lee, H. H., Chang, Y. S., Lin, C. L., Liu, T. Y., Chen, Y. C., … Chang, J. G. (2016). Down-regulated and Commonly mutated ALPK1 in Lung and Colorectal Cancers. Scientific Reports. https://doi.org/10.1038/srep27350
- Canisius, S., Martens, J. W. M., & Wessels, L. F. A. (2016). A novel independence test for somatic alterations in cancer shows that biology drives mutual exclusivity but chance explains most co-occurrence. Genome Biology. https://doi.org/10.1186/s13059-016-1114-x
- Alameda, J. P., Moreno-Maldonado, R., Navarro, M., Bravo, A., Ramírez, A., Page, A., … Casanova, M. L. (2010). An inactivating CYLD mutation promotes skin tumor progression by conferring enhanced proliferative, survival and angiogenic properties to epidermal cancer cells. Oncogene. https://doi.org/10.1038/onc.2010.378
- Sun, S. C. (2010). CYLD: A tumor suppressor deubiquitinase regulating NF-B activation and diverse biological processes. Cell Death and Differentiation. https://doi.org/10.1038/cdd.2009.43
- Mather, C. A., Mooney, S. D., Salipante, S. J., Scroggins, S., Wu, D., Pritchard, C. C., & Shirts, B. H. (2016). CADD score has limited clinical validity for the identification of pathogenic variants in noncoding regions in a hereditary cancer panel. Genetics in Medicine. https://doi.org/10.1038/gim.2016.44
- Marusyk, A., Almendro, V., & Polyak, K. (2012). Intra-tumour heterogeneity: a looking glass for cancer? Nature Reviews Cancer, 12(5), 323–334. https://doi.org/10.1038/nrc3261
- Fredriksson, N. J., Ny, L., Nilsson, J. A., & Larsson, E. (2014). Systematic analysis of noncoding somatic mutations and gene expression alterations across 14 tumor types. Nature Genetics, 46(12), 1258–1263. https://doi.org/10.1038/ng.3141

- Favero, F., Joshi, T., Marquard, A. M., Birkbak, N. J., Krzystanek, M., Li, Q., … Eklund, A. C. (2015). Sequenza: Allele-specific copy number and mutation profiles from tumor sequencing data. Annals of Oncology. https://doi.org/10.1093/annonc/mdu47
- Lek, M., Karczewski, K. J., Minikel, E. V, Samocha, K. E., Banks, E., Fennell, T., … Exome Aggregation Consortium. (2016). Analysis of protein-coding genetic variation in 60,706 humans. Nature. https://doi.org/10.1038/nature19057
- Mahoney, P. A., Weber, U., Onofrechuk, P., Biessmann, H., Bryant, P. J., & Goodman, C. S. (1991). The fat tumor suppressor gene in Drosophila encodes a novel member of the cadherin gene superfamily. Cell. https://doi.org/10.1016/0092-8674(91)90359-7
- Guillamot, M., Cimmino, L., & Aifantis, I. (2016). The Impact of DNA Methylation in Hematopoietic Malignancies. Trends in Cancer. https://doi.org/10.1016/j.trecan.2015.12.006
- Kakimi, K., Karasaki, T., Matsushita, H., & Sugie, T. (2017). Advances in personalized cancer immunotherapy. Breast Cancer. https://doi.org/10.1007/s12282-016-0688-1
- Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics, 25. https://doi.org/10.1093/bioinformatics/btp324
- Strietz, J., Stepputtis, S. S., Preca, B.-T., Vannier, C., Kim, M. M., Castro, D. J., … Maurer, J. (2016). ERN1 and ALPK1 inhibit differentiation of bi-potential tumor-initiating cells in human breast cancer. Oncotarget. https://doi.org/10.18632/oncotarget.13086
- Rajan, N., Andersson, M. K., Sinclair, N., Fehr, A., Hodgson, K., Lord, C. J., … Stenman, G. (2016). Overexpression of MYB drives proliferation of CYLD-defective cylindroma cells. Journal of Pathology. https://doi.org/10.1002/path.4717
- Tandy-Connor, S., Guiltinan, J., Krempely, K., LaDuca, H., Reineke, P., Gutierrez, S., … Tippin Davis, B. (2018). False-positive results released by direct-to-consumer genetic tests highlight the importance of clinical confirmation testing for appropriate patient care. GENETICS in MEDICINE. https://doi.org/10.1038/gim.2018.38
- Khurana, E., Fu, Y., Chakravarty, D., Demichelis, F., Rubin, M. A., & Gerstein, M. (2016). Role of non-coding sequence variants in cancer. Nat Rev Genet, 17(2). https://doi.org/10.1038/nrg.2015.17
- Tsoucas, D., & Yuan, G. C. (2017). Recent progress in single-cell cancer genomics. Current Opinion in Genetics and Development. https://doi.org/10.1016/j.gde.2017.01.002

Propositions

Making sense of cancer mutations

Looking into the wilderness beyond genes

Mamunur Rashid

1. Consensus approaches using multiple mutation detection tools out-perform individual tools alone, even if they have been highly optimised for a task (Chapter 2)
2. Orthogonal validation of somatic mutations detected using next generation sequencing is a vital prerequisite for depositing them in large scale databases. (Chapter 2&3)
3. In the absence of validated, gold standard data sets for developing prediction algorithms, astute choices of alternative approaches can create robust approximations (Chapter 6).
4. Difficult classification problems should not be solved by over-simplifying them via over- or under-sampling (Chapter 6).
5. Bioinformatics software developed in academia should be recognised as an integral part of research infrastructure development and funding strategies should more emphasis on ensuring documentation, distribution, support, and usability of these tools.
6. Cancer research has been blessed with a significantly larger share of fund compared to many other life-altering diseases. Healthcare research funding should be better distributed to include more non-cancer diseases such as malaria or mental health.
7. The contemporary framework of political correctness closes the door to legitimate criticisms and dialogue leading to more isolated and polarized societies.
8. In the era of information overload, scientists have a greater responsibility to effectively communicate their work beyond like-minded peers. In the absence of established scientific facts, gaps are more likely to be filled by social media-driven pseudo-science.
9. Overselling the promises of Artificial Intelligence (AI) driven systems in the healthcare industry in recent years has become the single biggest hurdle in its fruition.
10. Continuous search for the best experience (e.g. activity, food) might narrow the travel experience altogether rather than enhancing it.

These propositions are regarded as opposable and defendable, and have been approved as such by the supervisor Prof. Dr. M. J. T. Reinders.