Development of a reproducible transcriptomics variant calling workflow and its application to colorectal cancer.

A thesis presented in partial fulfilment of the requirements for the degree of

Master of Science in Genetics

At Massey University, Albany, New Zealand

Tom Manning

2020

Abstract

Colorectal cancer (CRC) is one of the most common cancers worldwide. It has some of the highest rates in New Zealand, exacerbated by short-comings in available diagnostic tools and survival discrepancies between Maori and non-Maori demographics. In this project, a bioinformatics workflow was developed to make “high confidence” single nucleotide polymorphism (SNP) variant calls from transcriptomics/RNA-seq data. While calling variants from whole genome and exome sequencing is common, standard workflows for calling variants from RNA-seq data do not exist.

Here, we aimed to use two common RNA-seq pre-processing methods which we then complemented with an ensemble of variant calling tools, improving confidence in any variants called. We then applied this pipeline to two independent CRC datasets with the hope that those variant calls could improve our understanding of the disease, one of the most significant aggregators of cancer-related mortality.

Variant calls were made including those with clinical implications, such as the same *KRAS* gene variant being called between both geographically distinct populations. Multiple “novel” variants, or those lacking clinically significant annotations, were also obtained for known oncogenic targets (e.g. *MAPK1* and *AKT1*).

RNA-seq variant calling remains problematic. The results of this study have provided us with some direction and considerations for future work, such as including normal samples to better distinguish between germline and somatic variants, permit the use of more somatic variant calling tools, etc. Future work is also needed to understand how or if those novel variant calls could improve our understanding of CRC.

Acknowledgements

My personal thanks go to Dr Sebastian Schmeier of Massey University, my supervisor for this project. Sebastian has provided numerous resources for help aspiring bioinformaticians, including myself, but also helped produce this piece of scientific literature. Sebastian has since left Massey University.

I would also like to thank all the lecturers, tutors, demonstrators, and peers that all contributed in various ways throughout my time at Massey University. Of those lecturers, I would like to specifically mention:

Dr Gabrielle Schmidt-Adam, who has since left Massey but made some of my early undergraduate courses some of the more enjoyable;

Dr Heather Hendrickson, whose relentless enthusiasm for bacteriophages is so infectious I feel the need to mention it here, and;

Dr Evelyn Sattlegger, who has also indirectly supported me through my post-graduate studies.

I would also like to thank Dr Rachel Purcell for providing me with the data used in this study, Sebastian's PhD student Arielle Sulit who helped me with some troubleshooting early in the pipeline's development, and the general bioinformatic community whose inquisitiveness often answered questions before I'd thought to ask them.

Jiayan Lin is also worthy of my thanks. As one of Massey's team of proof-readers, she significantly improved my writing and worked around my needs and better accommodated my productivity during the trying time that is 2020.

Speaking of 2020, I must also acknowledge that during the final few months of preparing this thesis occurred under the shadow of the covid-19 global pandemic. Despite this, I have been able to continue working on my thesis in relevant comfort. I've been fortunate and privileged to live in New Zealand during this time, when so many other places are suffering.

Thank you.

Table of Contents. Page No.

List of Tables and Figures. Page No.

1.0 – Introduction.

Colorectal cancer (CRC) is one of the most common cancers worldwide. It is the second most common in females (behind breast cancer) and third in men (behind lung and prostate cancer) [1-4]. Setting aside the sex-specific cancers, CRC is the second most common cancer in the world. Over 140,000 new cases were estimated to have occurred in 2018 alone [5]. CRC is also the cause of the second-most cancer-related deaths in the world, more than 8% of the total [6-8], primarily due to metastasis of the disease [9-11]. The US National Cancer Institute estimated that a total of over 1.5 million people (akin to the population of New Zealand's largest city) were diagnosed with the disease, and over 500,000 died from CRC in 2016 [12]. In 2006, the New Zealand Ministry of Health (MIH)'s "Colorectal Cancer Screening Advisory Group" found short-comings regarding the diagnostic tools for CRC available in New Zealand. They also found that Maori people have disproportionately worse post-diagnosis survival rates than non-Maori [13]. More recently, the MIH's "Cancer: New Registrations and deaths 2013" revealed that 3075 people had been diagnosed with CRC while 1252 people had died from the disease.

1.1 – CRC between demographics.

The prognosis for CRC is generally favourable, especially in "developed" countries like North America, Western Europe, Australia, New Zealand, and Japan [1, 14]. This may not be the case in developing countries where CRC prevalence may be increasing, possibly as these populations adopt a more "Western" lifestyle [15, 16]. Regardless of prognosis, CRC has higher rates in developed regions, with some of the highest being in Australia and New Zealand [16-18]. For comparison, the age-standardized rate (ASR) for these regions is as high as 44.8 (men) and 32.3 (women) per 100,000 person-years. Conversely the lowest rates are found in Western Africa (ASR of 4.5 and 3.8 for men and women, respectively [17]). As CRC is often observed in older demographics and our populations continue to age, it is thought that as many as 2.2 million new cases will develop over the next two decades [9].

Even in countries where CRC prognosis is favourable, both socioeconomic backgrounds [19] and poor patient compliance with screening recommendations (e.g. invasive colonoscopies) can lead to less favourable outcomes [6, 20-22]. Effective screening has been shown to be particularly important with regards to CRC prognosis. Stage I survival rates can be as high as 91% but can drop to 11% by stage IV [6, 22-24]. Effective screening likely also contributed to a reduction in CRC-related mortality over the last decade, reportedly more than 50% [25]. This is despite the availability of the faecal occult blood test (FOBT), an un-invasive, economical, and generally effective screening method. Poor adoption of the FOBT screen may be because of its reduced accuracy when compared to colonoscopies [6, 22].

1.2 – Genes with known CRC implications.

As with most cancers, between 70-90% of CRC cases are due to sporadic, somatic mutations of the genome [5, 8, 16, 22, 26, 27]. More than 80% of CRC cases reportedly involve inactivation of the *adenomatous polyposis coil* (*APC*) tumour suppressor gene. This gene is a major component of the β-catenin destruction complex [8] and it inhibits the canonical/β-catenin dependant *wingless/integration-1* (*wnt*)-signalling pathway [28-30]. As with other cancers [31], upregulation of this pathway results in more proliferative, "stem cell-like" cells, which are able to perpetuate their lineage [8, 32, 33]. Around 80% of somatic *APC* gene mutations are found within codons 1285-1580, known as the *APC* "mutation cluster region", which constitutes only 8% of its genetic code [8]. The frequency of this gene's mutation in CRC means that patients without variants for *APC* may suffer worse prognoses, as a *wnt*-signalling independent process likely drives the disease, thus requiring alternative therapeutic approaches [8].

Most of the remaining CRC cases belong to hereditary forms of the disease, such as familial adenomatous polyposis (FAP). FAP is characterized by the formation of numerous adenomatous polyps in the colonic epithelium. FAP cases, 70-90% of which involve germline variants, typically have an earlier age of onset than in sporadic cases. Some novel variants involved in FAP have been identified as recently as 2018 [34]. Both classical FAP and attenuated FAP (a milder form of the disease) are autosomal dominant diseases involving *APC*, 95% of which are caused by truncating nonsense or frameshift mutations in the gene [35].

An autosomal recessive form of the disease also exists. This disease involves the DNA damage repair gene *mutY DNA glycosylase* (*MUTYH*). This disease is sometimes referred to as *MUTYH*-associated polyposis or MAP [34]. There is also a non-polyposis autosomal dominant disorder, Lynch syndrome, which gives people a predisposition towards developing CRC and other cancers. Rather than *APC*, Lynch syndrome involves germline defects in one of the following four mismatch-repair (MMR) genes: *mutL homolog 1* (*MLH1*), *mutS homologs 1 and 6*, and *post-meiotic segregation increased 1 homolog 2* [36, 37]. For heterozygous carriers of these defects, subsequent somatic mutation or epigenetic changes affecting the remaining functional allele then leads to cancer formation, also known as carcinogenesis, oncogenesis, or tumourigenesis. Interestingly, MMR deficiency usually occurs in only a small fraction of advanced CRC cases [38].

1.3 – The genetic diversity or "heterogeneity" of CRC.

CRC is a complex heterogenous genetic disease [39] that develops as the colonic epithelium acquires adenomas (such as the benign polyps described above). These adenomas then transition into malignant carcinomas in a process known as the "adenoma-carcinoma sequence" [8]. The accumulation of somatic mutations or "variants" in the tumour (point mutations, gene fusions, etc.) drive both this adenoma-carcinoma transition and a tumour's on-going evolution [40]. Even the immune system, which would normally destroy cells with pathogenic variants, can be utilized by a tumour as it develops. For example, tumour-associated macrophages can promote hallmarks of cancer, like cell proliferation (the generation of new cells), metastasis (the

development of secondary malignancies from a primary tumour), and angiogenesis (the formation of new blood vessels that "feed" a growing tumour) [23].

A CRC tumour's position can profoundly influence a tumour's heterogeneity. Whether it originated on the right of the colon (which includes the caecum, ascending colon, hepatic flexure and the right side of the transverse colon) or left (left half of the transverse colon, splenic flexure, descending colon and sigmoid) can result in distinct molecular profiles. These different molecular profiles can also require different clinical approaches.

Tumours on the right side of the colon, for example, often have much worse prognoses, and can include gene-inactivating hypermethylated CpG islands. These islands are a biomarker that are indicative of massive gene inactivation, also referred to as the CpG island methylator phenotype or CIMP-high. Right-side tumours are also often enriched for hypermutable microsatellite instabilities (MSI), a result of silencing the *MLH1* MMR gene. Meanwhile, tumours on the left are more likely to have chromosomal instability, involving more chromosomal duplications or deletions, but such tumours are also less likely to be affected by hypermutation [41].

This heterogeneity gives rises to as many as three to six different CRC subtypes or endotypes [42-44]. The CRC Subtyping Consortium recently unified previously independent classifications into a single cohesive system with demonstrated prognostic value. This system details four biologically distinct consensus molecular subtypes or CMS: CMS1 (MSI immune), CMS2 (classical/canonical CRC), CMS3 (metabolically dysregulated CRC), and CMS4 (CRC with epithelial-mesenchymal transitions) [41, 45-48].

1.4 – The significance of the microbiome pertaining to CRC.

CRC's sporadic nature suggests that environmental factors contribute greatly to its development. Age [9, 22], smoking [9, 16, 49], alcohol consumption [9, 22, 49], obesity [5, 7], low physical activity [5, 9], chronic inflammation [7, 22] and diet [16, 50] are all factors likely to increase CRC risk [5, 7, 9, 16, 22, 49, 51]. These factors can also have a dramatic effect on one's microbiome. For example, one study has shown that more "Westernized" diets, which include more saturated fats, animal proteins, dietary additives, and less fibre, have seen microbiomes with less α-diversity (intra-microbial richness) and more β-diversity (inter-microbial richness [10]). A growing body of evidence has suggested that such "gut dysbiosis", the mal-adaptation of microbial communities within the colon, may contribute to the initiation and development of CRC [7, 10, 22, 24, 26, 41].

One specific example of gut dysbiosis is the increased occupation of oral microbes in the gut. This includes species like *Fusobacterium nucleatum* which has already been associated with CRC [22, 52]. Antibiotic-treated mice, implanted with adenocarcinomas and then intra-gastrically administered with *Escherichia coli* to simulate gut dysbiosis, have developed bigger tumours and more hepatic metastases [10]. Both *F. nucleatum* and enterotoxic *Bacteroides fragilis* (ETBF) are bacterial species are thought to play a role in the adenoma-carcinoma sequence [6, 14, 16, 21, 22, 25, 51, 53, 54]. For example, *F. nucleatum* encodes for the

fusobacterium adhesion protein A, a virulence factor that can activate *wnt*-signalling. This is akin to the inactivation of the *APC* gene [54-56], and ETBF has been found to function similarly [4].

The human microbiome can also host bacteria which are beneficial for patients with CRC. Members of the *Lactobacilli* genus have been suggested to improve the results of chemotherapy, and work synergistically with prebiotic short-chain fatty acids (SCFAs) that have been administered to treat cancer [5-7, 16, 57]. *Clostridium butyricum* has been associated with the apoptosis, or programmed cell death, of CRC cells [51]. The commensal form of *B. fragilis* has also been observed in mouse models to discourage the carcinogenesis [39].

In 2014, a CRC classifier based on microbial "operational taxonomic units" to classify bacterial groups, was found to be more accurate compared to the FOBT. Combining both this classifier and FOBT resulted in both higher specificity and increased sensitivity [6, 22]. A study on 2015 found it possible to diagnose CRC using microbial structures found in faecal matter [6, 16]. Correlations between cancer-related mutations (specifically loss-of-function/LOF mutations) and defined sets of microbial communities have been used to predict a tumour's mutational profile [58]. For the first time in 2017, individual bacterial species were associated with three CRC CMS groups, with associations for the fourth CMS likely not made due to the sample size (n=34) of the study [59]. These associations were achieved in part via a bioinformatic method known as transcriptomics or RNA sequencing (RNA-seq), from which this study's data was obtained.

1.5 – The versatility and limitations of NGS and RNA-seq data.

Next generation sequencing (NGS) technologies have greatly contributed to the development of personalized medicine. This means that clinical therapies can be adjusted based on an individual's genetic make-up [60]. RNA-sequencing (RNA-seq) is a technique where the RNA transcripts of a biological sample (the "transcriptome") are sequenced. This allows us to study any transcript (such as premature and mature messenger, micro, ribosomal, and non-coding RNAs, etc.) by adjusting the strategy used to isolate specific transcripts. For example, thymidine homopolymers can be used as probes that hybridize to the poly-adenine tail of a messenger RNA (mRNA).

RNA-seq is often used to compliment other techniques [61, 62], such as validating the findings of whole-genome or whole-exome sequencing studies (WGS and WES, respectively [63]), but is primarily used for studying gene expression. While publicly available RNA-seq datasets are not necessarily paired with WGS or WES data, RNA-seq can be used directly to study biological processes beyond gene expression, such as transcript splicing [64, 65]. The versatility and relative cost of RNA-seq, compared to other methods, means this technology is one of the most popular high-throughput technologies used. In fact, it is becoming increasingly common to use RNA-seq in medical settings to understand both rare Mendelian diseases and cancers [61, 66-68].

As mentioned above, changes in gene expression [69] is one of the most common uses for RNA-seq. One such study described how the *protein arginine methyltransferase 1* (*PRMT1*) gene

could contribute to CRC malignancy. For example, reduced expression or downregulation of *PRMT1* following propionate treatment (an SCFA produced via dietary fibre fermentation [70]) correlated with induced apoptosis in a HCT116 CRC cell line [71]. Incidentally, other SCFAs (such as the fermentation products of *C. butyricum)* have been shown to have similar anti-cancer effects [7, 70, 71].

In another study, increased expression of *transforming growth factor beta receptor II* was associated with right-sided CRC tumours. Meanwhile, the left side was associated with overexpression for ligands of the *epidermal growth factor receptor* (*EGFR*)'s protein, and the *vascular endothelial growth factor 1* (*VEGF*) gene [41]. A 2016 study by Kim *et al.* discovered that overexpression of *aldehyde dehydrogenase 1 family member A1,* and i*nsulin-like growth factor receptor-binding protein 1,* had significant and time-dependant suppressive effects on a colorectal carcinoma cell line, for both cell proliferation and tissue invasiveness [72]. Induced tissue inflammation studies have shown gene expression profiles akin to CMS4, which comprises nearly a quarter of all CRC cases. These expression profiles included both β-catenin hyperactivity and down-regulation of the *MYC* proto-oncogene's targets [73].

RNA-seq can also be used in other ways, such as to identify splice isoforms [74, 75]. This is the process of "splicing" together a gene's exons, which form a final mature mRNA which results in functionally different proteins [76]. Such transcripts cannot be studied easily, if at all, using other non-RNA NGS techniques. It is thought that approximately 75% of all human genes produce multiple splice isoforms. This can make accurate analysis of data involving splice sites difficult [77].

Aberrant spliced isoforms are also common and often overexpressed in cancer, and these novel splicing events may serve as diagnostic biomarkers. Such biomarkers could indicate chromosomal rearrangements, such as the "Philadelphia chromosome". This chromosome is an abnormally short chromosome 22, caused by a translocation with chromosome 9, and causes chronic myeloid leukaemia. While not pertaining to human disease, a recent 2016 study by Jakhesara *et al.* discovered 14 transcripts that may serve as horn cancer biomarkers for the Indian cattle species *Bos indicus* [78].

While all NGS methods are reasonably accurate, there is still a risk of sequencing biases or artefacts resulting from errors being reported for given nucleotide positions [79]. For example, the Illumina platform, one of the most popular second-generation NGS technologies, uses a method known as "sequence-by-synthesis". This method possesses both a low error rate (<1%) and reasonable costs (possibly being the technology that results in the USD1000 genome [80]). Despite its low error rates, Illumina can still struggle when sequencing low complexity regions or LCRs, such as tandem repeats (the repetition of the same short nucleotide sequence, such as "GATA"). Illumina struggles with such sequences due to the relatively short length of its' sequencing reads, hundreds of base pairs or bp [81], which could consist entirely of a tandem repeat. Illumina is not alone in this regard, as Ion Torrent sequencing also struggles with LCRs [82]. While other third-generation technologies like PacBio produce significantly longer read lengths (thousands of bases or kb), they also possess both significantly higher error rates (between 5-20%), and increased costs [68, 83].

The Guanine-Cytosine content (GC%) of a sequence can also influence biases and artefacts. This can be especially problematic for RNA-seq data, where GC% can also affect read amplification during sequencing [84, 85]. Another RNA-specific issue is the use of reverse transcriptase enzymes, which synthesises complimentary DNA (cDNA) for an RNA sequence prior to sequencing, which are thought to have increased error rates compared to DNA polymerases [86].

The differences between technologies may also require different computational or "bioinformatic" pipelines for downstream analysis, which may or may not be validated [82]. While these technologies continue to improve and reduce these artefacts [87], they have not yet been completely resolved. As so much "big data" is being generated from these techniques, robust and rigorous standards are required to better process this data, and not burden current knowledge with false-positive reporting [60, 79, 88].

1.6 – RNA-seq in the study of coding and non-coding regions.

Genetic variation, or the differences between individuals regarding a specific genomic nucleotide sequence, can exist anywhere in the genome. However, some of the more pathological changes often occur in protein-coding regions. RNA-seq can be particularly useful when studying such pathological variants, such as if a variant perturbs splice-sites with putative regulatory functions [63]. Malignant hyperthermia susceptibility, for example, is a pharmacogenetic disorder which appears to only be caused by missense mutations in coding exons [89]. However, protein-coding genes constitute less than 2% of the human genome [4], and 80% of the entire genome is thought to possess some biochemical or regulatory function. Variants in these non-protein coding regulatory regions, like gene enhancers or promotors [27], are thought to be responsible for as much as 30% of the known disease-causing variants. This is because non-coding variants can influence a gene's expression and/or the processing of its' protein product [90]. As a result, many non-coding RNAs, like micro RNAs (miRNA, which mostly downregulate genes [63]) and long non-coding RNAs (lncRNA), are often mutated and dysregulated in cancers [27]. Fortunately, these variants can also be studied using RNA-seq data [61].

Such studies have shown that lncRNAs can serve as diagnostic markers [91], and have been shown to regulate oncogenic processes like increased cell "stemness", oxidative stress response and cell death, and inflammation [4, 50]. Perturbations to miRNAs and lncRNAs (defined as being below 30 or over 200 nucleotides long, respectively [92]) have been linked to specific CRC-related processes. These include *wnt*-signalling described previously, the *mammalian target of rapamycin* (*mTOR*) pathway which regulates the cell cycle [3, 5, 50]), and specific genes like *MYC* and *apoptosis-inducing factor* [2].

One example of a lncRNA being involved in disease pathogenesis is the *B. fragilis-associated lncRNA1* (*BFAL1*). *BFAL1* has been found highly expressed in CRC tissues, upregulated in ETBF-treated cells, promote tumour growth via the *mTOR* pathway, and act as a "sponge" for the microRNAs *miR-155-5p* and *miR200a-3p.* Both *miR-155-5p* and *miR200a-3p* would otherwise regulate the *Ras homolog enriched in brain* gene, also involved in the *mTOR* pathway [4]. RNA-seq can also be used to investigate allele imbalance and more complex diseases

governed by mechanisms like allele-specific expression, imprinting, or X chromosome-inactivation [90, 93]. Other documented uses of RNA-seq include studies on gene fusions [94], inferring unannotated gene function [95], cancer drug resistance [56], and variant calling [96-102] for which there is currently no standard methodology.

1.7 – The impact and distribution of genetic variants in CRC.

Some common variants found in innate immune receptors are considered significant risk factors both within and between diseases. For example, there are genetic variants which affect both inflammatory bowel disease and CRC [103]. Many genes are frequently oncogenic targets, like the *tumour suppressor p53 (TP53), Kirsten rat sarcoma homolog (KRAS), serine/threonine-protein kinase B-raf (BRAF), neuroblastoma RAS viral oncogene homolog (NRAS),* and *EGFR* [19, 49, 67, 104]. Cancerous variants often play a role in various signalling pathways [19] like those above, and others like the *mitogen-activated protein kinase* (*MAPK*) pathway that regulates cell proliferation. Mutated genes involved in these pathways can cause them to be constitutively activated, aggravating oncogenesis and tumour progression. Reportedly, more than 40% of CRC patients possess mutations in the above *MAPK* pathway [23].

Genetic variants can worsen CRC cases in a multitude of other ways. Polymorphisms in the interleukin 17 family members *IL17A*, *IL17E*, and *IL23*, all involved in the adaptive immune system, have been associated with both increased CRC risk and poor prognoses [14]. Another example involves *activating transcription factor 6 (ATF6),* a gene involved in regulating the unfolded protein response which, if prolonged, typically drives pro-apoptotic pathways. Mice with a more active variant of *ATF6* have demonstrated caecum microbiome dysbiosis, increased epithelial cell proliferation, and loss of the mucus barrier that lubricates and protects the epithelium. These events all occurred prior to spontaneous tumour formation within the mice [105]. As recently as 2019, another study on mice has shown that deletion of the *mutated in colorectal cancer* gene, close in proximity to *APC* gene on chromosome 5, may promote CRC by failing to repair DNA damage caused by inflammation [73].

As mentioned, a tumour's position can also have specific variant profiles. Tumours on the right side have been associated with mutations in mismatch repair genes, both *KRAS* and *BRAF*, the tumour suppressive *microRNA-31*, and mutations involving the *RAS* and *mTOR* pathways. Tumours originating on the left side meanwhile have been associated with chromosome instability and mutations in genes *TP53*, *NRAS*, *APC*, *small/mothers against decapentaplegic family member 4* (*SMAD4)*, *KRAS* without *BRAF*, and the microRNAs *-146a, -147b, and -1288* [41]. Despite all this information, more work is still required to understand the pathogenesis of diseases like CRC to improve both detection and treatment [106-108].

1.8 – Identifying genetic variants, or "variant calling" with RNA-seq data.

Variant calling is the process by which we aim to identify genetic variants, many of which could be implicated with disease risk. Reference-based variant calling is one of the more common

bioinformatic methods for identifying genetic variation. Essentially this process involves: processing NGS reads , i.e. removing low quality bases and artificially ligated adapters; aligning or "mapping" the reads to complimentary sequences within a reference genome; identifying or "calling" nucleotide variation between the reference and the sample, and; applying variant filters, which help distinguish true positives and negatives from erroneous sequencing artefacts [86].

Reference-base variant calling has some benefits over reference-free methods, such as requiring fewer computational resources and having higher sensitivity [109]. The kinds of variants that can be identified using bioinformatic techniques are diverse, from simple single nucleotide variants or polymorphisms (SNVs/SNPs), to increasingly complex variants more likely to have functional roles in disease [63]. Examples of more complex variants include multiple nucleotide variants/polymorphisms (MNVs/MVPs), structural variants (SVs), insertions and deletions (collectively known as "indels"), and chromosomal rearrangements [19, 87].

With regards to RNA-seq variant calling, there has been some success identifying frequently mutated target genes in CRC [110]. In cases where a variant increases gene expression, RNA-seq may be better than other NGS technologies in calling these variants, as increased expression can improve RNA-seq coverage [86, 99]. This also means the opposite is true for variants which reduce expression. Such variants would require deeper sequencing, which refers to how many reads overlap at a given position, to distinguish low frequency variants from erroneous sequencing artefacts [19]. However, deeper sequencing also increases sequencing costs [86, 100].

In some cases, more accurate reference-based variant calling can be achieved by using the most recent release of the human genome. This is currently GRCh38, which has various improvements over the previous release, GRCh37. For example, the GRCh37 release lacks sequences surrounding chromosome centromeres, known as peri-centromeric sequences. These centromeric sequences were omitted as they consist of "alpha satellites": dinucleotide tandem repeats, approximately 171bp in length, which can be difficult to sequence [111]. Exclusion of these peri-centromeric sequences means that genes within these regions, even those with known disease-causing variants, cannot be called using a GRCh37-based variant calling method [112].

Another consideration for variant calling pertains to all NGS methods that relying on the polymer chain reaction, or PCR, which is likely to produce optical duplicates. These are sequencing reads produced by the amplification of the same polynucleotide during sequencing [113]. Without identifying duplicates prior to variant calling, the depth for these sequences could be artificially inflated. This could lead to erroneous variant calls if the original sequence contains a sequencing artefacts.

The length of sequencing reads can also affect variant calls, as demonstrated by a recent study in plant genetics where longer reads (as few as 25 bp more) significantly reduced the false positive rate [66]. Longer read lengths also mean fewer ambiguously multi-mapping reads, i.e. reads that map to multiple genomic locations purely by chance, but this again increases sequencing costs. The samples themselves, be they fresh, formalin-fixed, paraffin-embedded, etc. can also influence error rates during sequencing [86].

Which variant calling tool to use is another issue, regardless of the NGS technology is used, as many variant callers are developed independently. This may mean different variant calls could be made from the same data [60, 114]. Different variant callers may also represent the same variant in different ways [60, 115]. Consider this example provided by Tan *et al.* [116], which describes a 14-bp reference sequence (REF) of "GGGCACACACAGGG" and a 2-bp deletion in the 12-bp alternative sequence (ALT) "GGGCACACAGGG". This seemingly simple variant could be represented in a myriad of ways. For example, the second pair deleted at the sixth position could be represented as ALT = C. Alternatively, the first pair deleted after the third position could be represented as ALT = GCA. Incidentally, deletion of that first pair could also be represented as ALT = G. While none of these representations are incorrect, this lack of a standardised format means that both calling variants and then annotating them with different variant databases, which themselves may be formatted differently, could lead to confounding results [117].

One of the most difficult challenges to overcome for RNA-seq variant calling concerns eukaryotic splice sites. If an RNA-seq read contains a splice junction, i.e. it consists of two or more exons that have been "spliced" together, then mapping tools that are "splice unaware" may try to map this read as if it was a sequence of DNA. This would result in either one exon being mapped correctly while the other or others are falsey identified as variant sequences, or (more likely) the tool will fail to align the read entirely [86].

This has necessitated the development of "splice-aware' sequence aligners, such as the STAR tool, as well as pre-processing RNA-seq data for specific applications like variant calling [64, 86, 101, 118-120]. Pre-processing RNA-seq data is one way to improve a variant calling pipeline's sensitivity (the ability to make variant calls) and specificity (being able to distinguish them from false positives) [86].

If we consider the above issue regarding splice junctions, one method to improve RNA-seq variant calling is to split such reads into their exon segments. This makes the reads more "DNA-like" and avoids some of the issues when making calls around splice junctions. This "DNA-like" approach is currently being used by the Broad Institute as part of their "best practises" for RNA-seq variant calling. This methodology, currently still in development, uses the popular Genome Analysis Toolkit (GATK) [40, 86, 120-124] to make variant calls.

However, the current GATK RNA-seq variant calling workflow makes several compromises to consider. For example, the GATK DNA-seq workflow uses the "Variant Quality Score Recalibrator". This is a tool that uses "truth sets" available for DNA-seq to improve downstream analysis. Such truth sets are not currently available for RNA-seq data, and so the GATK team has instead relied on a series of "hard filters" for making RNA-seq variant calls. These filters aim to exclude likely false positives based on the annotations provided by variant calling tools. Unfortunately, this means some true positives will be lost while some false positives are retained, simply because their annotated scores are sufficient for the filter being used [125].

While the GATK RNA-seq best practises are certainly an improvement over making raw variant calls, they are still a far cry from being of a "gold standard". Often in bioinformatics, these "gold standard" set of tools are difficult to ascertain. This is especially true for RNA-seq variant calling

[60, 66] partly because, like the aforementioned truth sets, there are no validated RNA-seq data sets available like Illumina's platinum genomes [126]. Validating RNA-seq variant calling pipelines is therefore reliant on synthetic datasets (which can never truly simulate real data), comparing performance to existing pipelines [60], or via orthogonal methods like Sanger sequencing [60, 86]. This is also the case for other NGS methods, although it appears that orthogonal methods are becoming increasingly less useful as these technologies improve [127-129].

It is also possible to never truly achieve one gold standard for all instances, given the breadth and depth of bioinformatic subjects worth studying. For example, low-frequency variants may require modern variant callers which can assess variant allele frequencies (VAF), which refers to how often a variant is observed in a sample at a given position. This is unlike "traditional" variant callers that rely on probable genotypes, such as whether a sample is homozygous or heterozygous at a given position [86].

Low frequency variants also require significantly more depth to distinguish them from sequencing artefacts [68]. Often traditional tools are not designed for such depths and so may arbitrarily remove some sequences, a process known as "down sampling". This may omit low frequency variants entirely if they are among those sequences that were down sampled. Traditional variant callers also struggle with identifying more complex variants, for example they may confuse the individual nucleotides of an MNP (e.g. GCA) as separate SNPs (G, C, and A).

Reliance on any one tool or methodology could have significant clinical implications. If we consider the above confusion surrounding MNPs and SNPs, a therapy that targets a SNP variant could be administered to a patient who in fact carries an MNP [67]. Variant calling pipelines may therefore benefit by incorporating multiple variant callers, a methodology known as consensus or ensemble variant calling. The Consensus Variant Calling System (CoVaCS) is one example which uses three variant callers (VarScan, HaplotypeCaller, and Freebayes [130]). This example demonstrated more specificity and slightly more sensitivity than when using these tools individually [131]. This study also demonstrated how ensemble methods, like CoVaCS and the Consensus Genotyper for Exome Sequencing [132], are most effective when one tool compliments another's weakness, e.g. if one tool better calls SNP variants and another better calls indels [131]. This ensemble method of variant calling has also been demonstrably used when using RNA-seq in a recent study for the black poplar, *Populus nigra* [133].

As with most cancer studies [27, 54, 55, 64, 86], often normal-tumour paired tissues samples are compared when using variant calling to better understand cancer pathogenesis [19, 134]. This allows researchers to better distinguish germline and somatic variants [86, 135]. Any biological sample could become "contaminated" with cells from different tissue, for example if normal tissue is sampled near an invasive tumour or if the tissue has already been infiltrated during metastasis [136, 137]. Sampling both normal and cancerous tissue, therefore, helps avoid this concern.

Being aware of these processes during experimental design helps reduce the false discovery rate [138], as germline variants are both frequently observed between individuals and are often clinically insignificant. This is especially true for individuals from more genetically diverse ancestries, such as those from African or recently admixed populations [88]. For example, the 1000 Genomes Project (1KGP), originally based on GRCh37, found that on average one variant is expected for every eight bases within a person's exome [124]. An individual's genome may differ from GRCh37 by as many as 5 million sites, mostly from SNPs and approximately 2100-2500 being SVs [87]. Most individuals are thought to possess over 100 LOF genes, over 10,000 genes with peptide-altering sequences, and around half a billion variants that overlap known regulatory regions [87, 139-141]. Additionally, current efforts to update the 1KGP with the more accurate GRCh38 reference may find these initial estimates to be conservative [142].

1.9 – *Our study's aims and objectives.*

Here, we present a bioinformatic workflow which has been developed upon existing RNA-seq variant calling methodologies. Our study aims to develop a workflow to reliably call variants from existing samples for which clinical diagnoses may be too late. As a result, we favoured a pipeline that would minimise the number of false positive variant calls over speed of analysis. We then aimed to use the workflow to make high confidence RNA-seq variant calls for two sets of RNA-seq data which included both primary CRC tumours and their subsequent liver metastases (an organ not commonly affected by primary tumours [143]). Any variants called as part of this pipeline could further our understanding of CRC. To achieve this aim we will:

1) Map RNA-seq reads to the most recent release of the human genome, GRCh38;
2) Pre-process mapped RNA-seq reads via two different methods to try minimising bias/artefacts;
3) Use an ensemble of variant calling tools to call variants;
4) Select variants based on "intersecting" calls from all our variant calling methods (a method being defined as a combination of a pre-processing method from 2) and a variant caller from 3) above;
5) Annotate high confidence calls with known variant database annotations to support a variant's status (such as germline, somatic, pathogenic, true positive, etc.), to compensate for our lack of normal samples for one data set;
6) Perform gene set enrichment analysis and prediction of deleteriousness to provide focus on variants, which may include unannotated or "novel" variants, that are more likely to be oncogenic, and;
7) Use gene expression data to understand possible mechanisms by which these variants could exacerbate carcinogenesis.

While the significance of this specific work is aimed toward developing a better understanding of CRC pathogenesis, the wider significance of this work will be to establish a reproducible and automated computational methodology that can be use for other cancerous samples and similarly improve how these diseases can be understood.

2.0 - Materials and Methods.

Our study intended to make high confidence variant calls from the RNA-seq data from a cohort of patients from New Zealand (NZ) and South Korean (SK). Both data sets contained sequencing data for colorectal primary tumours (the original location where the cancer developed) and their subsequent secondary liver metastases, while the SK data also contains a set of normal samples.

We opted to develop a reference-based variant calling pipeline using the most recent release of the human genome, GRCh38. This means that the RNA-seq data is aligned to the reference so that discrepancies between the reference and the data can be "called out" in the output files produced after sequencing alignment. As this study used RNA-seq data, the alignment files then need to be pre-processed to improve variant calling sensitivity and specificity. Pre-processing occurred before making variant calls from an ensemble of calling tools.

Finally, variant calls were filtered to further improve confidence of these variants being true positives. The remaining "high confidence" variants were analysed using various techniques to try and determine any pathological impacts they may have pertaining to CRC pathogenesis. These various steps are detailed in the following section.

2.1 – The "NZ" and "SK" RNA-seq data sets.

The primary data set was a collection of deeply sequenced RNA-seq data (Illumina-based, paired-ended, 150-bp RNA-seq reads) derived from 13 paired biological samples (primary-metastasis). These biological samples were obtained as part of the Purcell *et al.* study [59] and are referred to as the "New Zealand/NZ" data. This data was adapter-cleaned using "fastq-mcf" [144], and "SolexaQA++" was used to dynamically trim sequencing bases of low quality [145].

The NZ data set was not available at the beginning of the study. This necessitated the development of the pipeline using a second publicly available data set (Illumina-based, paired-ended, 100bp RNA-seq reads) accessed via the "National Center for Biotechnology Information" (NCBI, accession PRJNA218851, ID 218851 [72, 146]) while the NZ data completed sequencing. The SK data was derived from 56 samples (18 normal-primary-metastasis triplets) and was referred to as the "South Korean/SK" data. Normal, colorectal and liver metastatic samples are referred to as NRM, CRC, and LM respectively.

Unlike the SK data, the NZ data lacked NRM samples that would normally be sequenced in variant calling studies along with cancerous samples. The sequencing of NRM samples would have helped distinguish between germline and somatic variants. NRM samples were not obtained for the NZ data set, as the original study did not pertain to making germline/somatic variant calls.

The aim of the pipeline, therefore, was to call both somatic and germline variants using a combination of different RNA-seq pre-processing methods, an ensemble of variant callers, and then using database annotations to help distinguish between germline and somatic variants.

2.2 – Computational workflow management using Snakemake.

The current workflow used "Snakemake" [147], a scalable workflow engine based on the Python programming language. This tool allowed us and will allow future users to automate and reproduce the workflow for multiple samples. Incorporating Snakemake into the workflow required writing a "Snakefile" (Appendix 1) that contained a series of rules with shell commands that connect tasks to each other. In these commands, various inputs and outputs (e.g. "sample.file") were substituted with generic "wildcard" terms (e.g. "[SAMPLE].file") which Snakemake could then use to automate the various rule's inputs and outputs. Snakemake also allowed us to run rules in parallel, given we had access to multiple CPU cores, by first producing an ordered graph of rules that could be executed automatically as input became available.

We also created "tool environment" YAML-files for each of the rules to make the workflow more easily reproducible on other systems. These YAML files allowed Snakemake to download and install the same tool versions used in the workflow before execution. This meant that a "hard-coded" and controlled bioinformatic environment can be recreated for each run automatically without having to install tools manually. The "GATK Best Practises for variant calling on RNAseq, in full detail" [97, 148-150] was used as the basis upon which the pipeline was developed. However, some changes were made to this methodology, discussed below, to increase confidence in the variants called.

2.3 – Choice of reference genome and preparation for read alignment.

We chose the "GRCh38_no_alt_plus_hs38d1_analysis_set" reference genome [151], referred to as "our reference". GRCh38 has been described as a more accurate representation of the human genome with improved alignment and fewer false positives in other studies [152]. In comparison to GRCh37, GRCh38 has over 80 million fewer unannotated ("N") nucleotides, reduced sequencing gaps (including centromeric regions), decreased GC% for 17 of the 24 autosomes, and a significantly increased (over 20 million nucleotides) exome [153]. GRCh38 also includes the rCRS mitochondrial sequence [142, 154, 155], while the GRCh37 mitochondrial sequence includes a 2 bp insertion which may complicate alignment [156]. The "GRCh38_no_alt_plus_hs38d1_analysis_set" reference also contains sequences for the Epstein-Barr virus and some human decoy sequences. These act as "sinks" that reduce false positives as described in our introduction.

Before executing the workflow, a "sequence dictionary" was generated using Picard's [157] "CreateSequenceDictionary". This output was then updated with the Single Nucleotide Polymorphism Database (DBSNP [158], specifically its "00-All" release) with Picard's "UpdateVcfSequenceDistionary". This avoided downstream incompatibility between the original DBSNP file header and some downstream tools. The resulting file was then "block gzipped" using Tabix [159]. A "genome fasta file index" (different from the genome's index detailed below) was also required for the GATK [160] tool "SplitNCigarReads" before the execution of the workflow, which was achieved using Samtools' "faidx" function [161].

2.4 – Genome indexing and sequence read alignment.

We elected to use the Spliced Transcript Alignment to a Reference (STAR) alignment tool. This is a fast, efficient, and splice-aware RNA-seq mapping tool with comparable performance to other aligners such as GSNAP, GSTRUCT, and MapSplice [162]. STAR also has a higher tolerance for accepting mismatched and "soft-clipped" reads by default. This means that bases at the beginning and end of a read weren't included in alignment as they are prone to sequencing artefacts. However, these bases can still provide useful information for a read's sequence when calling variants. This allowed STAR to align more reads than other aligners [163]. STAR is also recommended for various GATK best practise workflows [121], including its RNA-seq best practises workflow [164].

Before aligning sequencing reads to the reference, the pipeline first generated a "genome index" via STAR's "genomeGenerate" function. This genome index is required by STAR and speeds up sequencing alignment. To produce a genome index, both the reference and its annotations ("GRCh38_full_analysis_set.refseq_annotation", gtf file format [165]) were required as inputs. While indexing was time consuming and genome indexes were available for download, STAR users are also recommended to generate their own indexes using up-to-date assemblies and annotations [166].

After indexing, STAR's "multi-sample 2-pass mapping" protocol was used to align and map the sequencing reads to the reference. This ran STAR twice for each sample, where the first run produced a list of splice junctions for a sample that then improves alignment during the second pass. While the current pipeline generates output files for the first pass, providing use with files for extra statistics, this was not required.

We did not use the basic 2-pass option recommended by the GATK workflow and instead compiled a list of splice junctions found for all samples using a custom "AWK" script. Compiling this list meant that all samples would be made aware of more splice junctions than if they had only considered those found during an individual sample's first pass. This can improve alignments and allow more sequences to be mapped.

We also did not include GATK's recommended "FastqToBam" or "MergeBamAlignment" steps. These steps would have converted information contained within a samples' unaligned RNA-seq FASTQ reads into a BAM file (the binary equivalent to the "Sequence Alignment/Map file). This would have then been merged with the samples' STAR-produced BAM file. Instead the STAR option "—outSAMunmapped Within" was used which both retained unmapped RNA-seq data and bypassed incompatibility between both the "FastqToBam" and "MergeBamAlignment" tools and the SK data.

The output BAM option "sorted-by-coordinate was required for downstream tool compatibility. The pipeline then added read group annotations as per the GATK best practises workflow, achieved via Picard's "AddOrReplaceReadGroups" tool.

2.5 – Pre-processing prior to RNA-seq variant calling.

Before making RNA-seq variant calls, the output BAM files were pre-processed to increase variant calling sensitivity and specificity. GATK's currently in development best practises workflow for calling variants in RNA-seq relies on hard filters with known issues, especially in the absence of truth sets for human RNA-seq data sets. To address this, a second pre-processing step was included in the pipeline using the Opossum tool (OP, [167]). While OP was primarily designed to improve RNA-seq variant calling with the Platypus variant caller (PT, [109]), OP is also compatible with other tools like GATK's "HaplotypeCaller" (HC) included in our ensemble variant calling methodology. Variants that could be called from both types of pre-processed BAM files are less likely to be false-positives, as this would require these independently developed tools and methods to give rise to the same artefact. While OP was a single tool and required only one non-default setting (command line option "--SoftClipsExist True"), the GATK best practises involved several tools discussed below.

First, Picard's "MarkDuplicates" tool was used to mark duplicate reads, broadly defined as "reads with identical sequences or starting position with regards to the reference sequence" [85]. These "markdup" BAM files then required indexing which we achieved via SAMBAMBA's "index" function [168]. GATK version 3, rather than version 4 (referred to as GATK3/GATK4 respectively [121]) of the tool "SplitNCigarReads" was then used in this pipeline. This tool both splits RNA-seq reads into exon segments to avoid errors surrounding splice junctions and removed or "clipped" sequences that overhung into intronic segments. GATK3's version of this tool was used as GATK4's version at the time of the pipeline's development had not been validated [169]. SplitNCigarReads also reassigned the BAM file's STAR mapping quality scores into GATK-equivalent values for downstream compatibility. The optional "AnalyzeCovariates" step was omitted before performing "base quality score recalibration (BQSR)" with GATK4's "BaseRecalibrator". BaseRecalibrator required both DBSNP as input and for the DBSNP to be indexed specifically with Tabix. The recalibration table output by BQSR was then used as input for GATK4's "ApplyBQSR" tool, completing GATK RNA-seq pre-processing.

2.6 – Ensemble Variant Calling.

We used an ensemble variant calling strategy in conjunction with the two pre-processing methods to reduce false positives and improve confidence in the variants called. For both the GATK and OP pre-processed BAM files, we used GATK4's HC, GATK4's Mutect2 (MT2 [170]), and Freebayes (FB [130]). For the OP-only BAM files we also used the PT variant caller, which we found was incompatible with the GATK RNA-seq pre-processing method.

For both FB and PT tools, we used their default settings as no publications have provided settings specific for RNA-seq variant calling. However, for HC we included the following command line options as recommended by GATK's RNA-seq best practises: include the DBSNP as an input, "dont-use-soft-clipped-bases", and "stand-call-conf 20.0". For MT2 we also included "dont-use-soft-clipped-bases" and used the genome aggregation database (GNOMAD, specifically the "af-only-gnomad.hg38" release [171]) as the input for MT2's "germline-

resource" option. Other options for variant filtration were included for both HC and MT2, described below.

Finally, each of our seven different variant calling "methods" (defined as a combination of either GATK or OP RNA-seq pre-processing and one variant calling tool) produced a single output file in the "Variant Calling Format/VCF" format [172]. Each method's output will be referred to as GATKFB, GATKHC, GATKMT2, OPFB, OPHC, OPMT2, and OPPT.

2.7 – Variant Filtering.

Variant filtration is a method of improving confidence in the variants called and removing potential false positives [173]. Specific filters for RNA-seq data, the HC variant caller, and SNP variants were provided by the GATK best practises and included in the pipeline. These filters were: a "SNP cluster", described as three or more SNPs that occurred within a sequence window of 35bp; a Fischer-Strand (FS) bias score of below 30.0, which filters out sequencing artefacts that occurred predominantly on one of the sequenced strands, and; a Quality-by-Depth (QD) value of above 2.0, which is a normalized quality score based on the sequencing depth supporting a variant. GATK4's "VariantFiltration" tool was used to apply these filters specifically to the HC VCF files. For the MT2 VCF files we used the MT2-specific filtering tool, GATK4's "FilterMutectCalls".

Variants that passed these filters were annotated with a 'PASS' in their respective VCF files, which allowed us to remove all other variants without this annotation using SnpSift's "Filter" function [174]. These filters needed to be applied before merging a sample's VCF files as, after merging, some variant calling tool-specific annotation, such as HC's QD values, etc. were lost. The OPPT output file also required sorting via Picard's "SortVCF" to be compatible with GATK3's "CombineVariants" tool, which we used to merge the VCF files produced by each of our variant calling methods.

Within a sample's "merged" VCF file (which was then "block gzipped" compressed and indexed via Tabix), variants that were called by all seven methods possessed the unique annotation "set = Intersection". This annotation was then used to filter for our "highest confidence" variant calls further in the pipeline. To first filter for SNP variants, the focus of this study, GATK4's "SelectVariants" tool was used with the command line option "select-type-to-include SNP" on the merged VCF files.

The variants that remained were then annotated with SnpEff to predict their functional impact [174, 175], with frame-shift mutations being an example of variants with a putative "high" impact. The SnpSift "Annotate" function was used to add variant annotations from known variant databases: the "Clinical Variants" database, specifically the available "clinvar" and "clinvar_papu" releases (referred to collectively as "ClinVar" [176]); the Exome Aggregation Consortium, specifically the "ExAC.r0.3.1.sites.vep" release (referred to as ExAC [177]); the Catalogue of Somatic Mutations In Cancer, specifically the "CosmicCodingMuts" release (referred to as COSMIC [178]), and; the above mentioned releases for DBSNP and GNOMAD.

These annotations were added to help distinguish between germline and somatic variations, e.g. variants with COSMIC annotations were more likely to be somatic, etc.

"SnpSift Filter" was then used remove all variants which lacked the "set = Intersection" annotation described above leaving only the highest confidence calls. Given our interest in clinically significant variants, "SnpSift Filter" and the command line option "(ANN[0].IMPACT has 'HIGH')" was used to isolate variants with predicted "high" impact annotation as provided by SnpEff. The final VCF files therefore contained only SNP variants called in high confidence with putative high impact SnpEff annotations.

A custom script was then used to compile each variant's QUAL score to produce a tab-delimited text file for further analysis using Microsoft Excel. Finally, statistical outputs for the unprocessed, GATK-processed, and OP-processed BAM files were produced using the Samtools' "stats" and "flagstat" functions for comparison.

2.8 – Gene set enrichment analysis with "Enrichr".

Gene set enrichment analysis can be defined as a "computational method that determines whether *a priori* defined set of genes shows statistically significant and concordant differences between two biological states [179]". Enrichr [180, 181], is a web-based tool that allowed us to perform such an analysis. This was achieved by uploading a gene set that is then compared to various databases of existing biological libraries. Enrichr then reported back, giving users functional categories in which genes from the input set appear statistically more often than by chance. Each of these results were ranked based on a p-value, adjusted p-value, a z-score, and a "combined" score.

Our gene sets consisted of only variant-affected genes that were supported by more than 2 samples and had an average QUAL score above that of the median QUAL across all variants for that data set. For the SK data the following gene sets were produced: "All" (all variant-affected genes across all samples), "CRC" (genes affected by variants in CRC samples), "LM" (as previous for LM samples), and "cancerous" (genes whose variants were not observed in NRM samples but may have been shared across CRC and LM samples). As the NZ data lacked NRM samples, only two gene sets were produced for "CRC" and "LM" as described above, with the LM gene set effectively also serving the same purpose as the "All" gene set.

As well as noting the functional categories that were enriched as part of our analysis, we also made a tally or count each time a gene was observed for an Enrichr functional category. The tally/count was made provided that entry: could be associated with cancer, such as when an Enrichr entry considered cancerous cell lines, signalling pathways relating to cancerous hallmarks (e.g. proliferation, angiogenesis, etc.), proto/oncogenes (e.g. *TP53*, *KRAS*, etc.); pertained to human entries, e.g. provided human gene names and descriptions, etc. and; possessed the lowest p-value (<0.01) within that database. For brevity, we only considered genes among the highest three tallies for each gene set, our logic being that this would allow us to

focus on genes observed frequently in cancerous-related functional categories. These genes and their variants are therefore more likely to be involved in oncogenesis.

2.9 – Deleterious SNP predictions with "PredictSNP2".

PredictSNP2 is a web-based tool that allowed us to upload a variant's details in a simplified format, allowing deleteriousness to be predicted by a collection of tools [182]. These tools included: the Combined Annotation Dependent Depletion (CADD), Deleterious Annotation of genetic variants using Neural Networks (DANN), and Functional Analysis through Hidden Markov Models (FATHMM), among others.

PredictSNP2 was specifically used to assess the deleterious nature of the "novel" variant calls, defined as "variant calls made in high confidence that lacked clinically significant or explicit annotations provided by known databases". Each of the tools that encompass PredictSNP2 provided an expected accuracy for their deleterious predictions, with an overall scored based on the combined output of these tools. PredictSNP2 was included in our analysis given its ease of use, accurate prediction of known pathogenic variants described in our results, as well as providing support for the SnpEff high impact annotation where appropriate.

2.10 – Gene expression analysis with Subread's "featureCounts" and DESeq2.

To further investigate those novel variants described above, gene expression was analysed using a combination of bioinformatic tools. First, gene expression was quantified using the unprocessed BAM files after Star2Pass alignment as input for the "featureCounts" software [183], found within the Subread 2.0.0 package [184]. This tool outputs the number of reads assigned to features or "meta-features" such as exons or genes and some statistical information from the overall summation of the results.

Differential gene expression was then achieved using DESeq2 [185] and reported back as the log2 fold change (log2FC) between two compared sample sets. Differential expression was also accompanied by an adjusted p-value to indicate statistical significance in the log2FC, for which we only considered adjusted p-values below 0.05 as being significant. The computational methodology for reproducibility is available at the following link [186]. We then compared the average of a gene's expression across all samples for a given data set against that of the "variant's expression", i.e. we compared the TPM values for the samples from which a variant was obtained against that of the average for that gene's expression across all samples of the relevant data set.

3.0 – Results.

We will first describe the number of reads obtained for each data set after sequencing, providing some preliminary statistics that were obtained from the data. Next we show results of comparing the performance between the GATK and Opossum (referred to as OP) pre-processing methods and our ensemble of variant callers. Our focus then begins to shift towards specific variant calls made during our analysis and how they may or may not influence CRC pathogenesis. Table 1 provides some general statistics obtained from both data sets.

Table 1

Data set	No. of patients	No. of samples	Primary (CRC) samples	Metastatic (LM) samples	Normal (NRM) samples	Reads mapped (million)	Average reads mapped per sample (million)
NZ	12*	24*	Yes	Yes	No	120.2*-288.7	187.0*
SK	18	56	Yes	Yes	Yes	43.7-128.3	90.2

Table 1 – Summary of the NZ and SK data sets. * refers to the exclusion of a problematic sample (CRC335, 50.5 million mapped reads) and its paired metastasis (LM335, 213.5 million reads mapped), which would have lowered the average reads mapped per sample to 182.8 million.

3.1 – Statistical and GATK/OP performance comparisons.

Section 3.1 primarily compares the statistical outputs between the two pre-processing methods in order to compare the performances of both the GATK and OP pre-processing methods, to inform future studies. This section will also compare the outputs for our various variant calling methods that made up our ensemble of variant callers, to justify our use of the "set = Intersection" filter to obtain our "high confidence" variants, described below.

3.1.1 – Preliminary statistics obtained for both data sets.

Here we shall describe some statistics obtained from the NZ and SK data sets, to inform us as to any discrepancies between their results. For the SK data (56 samples, 18 normal-tumour-metastases triplets), over $1.3*10^9$ reads were obtained that were uniquely mapped to the reference. This corresponded to between 68.0-83.6% of a samples' total reads (average $2.41*10^7$ reads per sample, supplementary material 1).

Also, we found the average runtime for this data set to be 46 hours, 43 minutes, and 5 seconds (46h:43m:05s, supplementary material 1). Runtime was roughly linear with the number of uniquely mapped reads and no notable differences had been observed between sample types (CRC, NRM, or LM, Figure 1a). After pre-processing, variant calling, and variant filtration, over $3.70*10^7$ variants (including non-SNP variants) were called for the SK data (supplementary material 2).

Our NZ data (26 samples, 13 tumour-metastatic pairs) included one sample (CRC335) with far fewer reads than all other samples. This sample was therefore excluded from further analysis along with its paired LM sample (LM335). Despite the removal of these samples, approximately $1.15*10^9$ reads were uniquely mapped to the reference genome which corresponded to

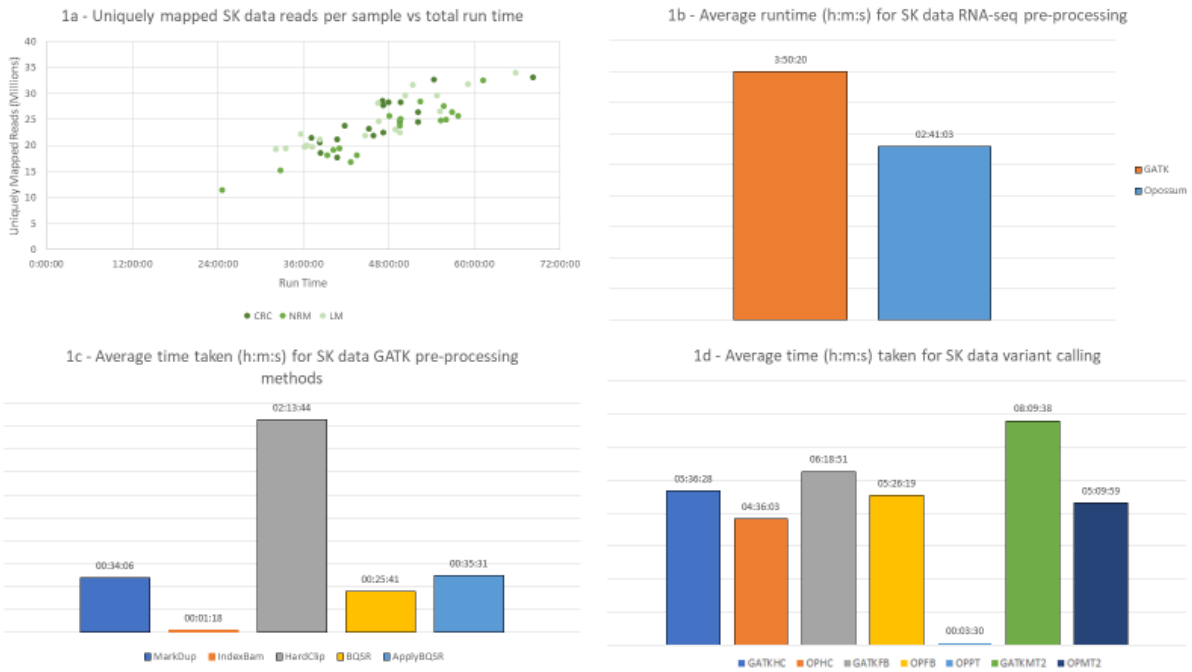

Figure 1 – Runtime data for the SK data set. 1a – Variant calling runtime appeared roughly linear with number of uniquely mapped RNA-seq reads. 1b – On average, GATK RNA-seq pre-processing took longer than OP pre-processing. 1c –The "SplitNCigar" or "HardClip" tool contributed most notably to GATK pre-processing runtime. 1d – Generally OP variant calling "methods" (defined as a combination of one pre-processing method and one variant caller) were quicker to process than their GATK contemporaries, with OPPT being substantially quicker than all other methods.

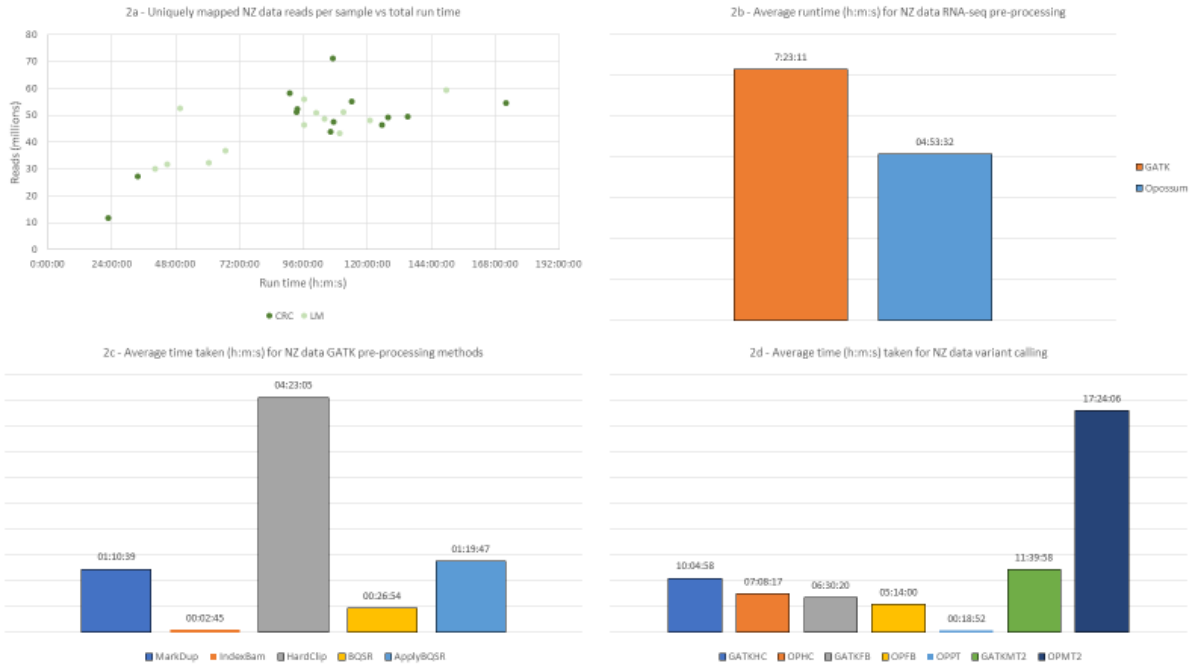

Figure 2 – Runtime data for the NZ data set. 1a – Variant calling runtime again appears roughly linear with number of uniquely mapped RNA-seq reads, with the reduced linearity likely because of the reduced number of NZ samples compared to SK. 1b – GATK RNA-seq pre-processing again took longer than OP pre-processing on average. 1c –The "SplitNCigarReads/HardClip" tool again contributed most notably to GATK pre-processing runtime. 1d – As with the SK variant calls, generally the same trends were seen for GATK vs. OP variant calling methods with one exception, as OPMT2 took substantially longer than all other methods. OPPT remained the fastest combination in comparison.

between 70.8-87-8% of a samples' total reads. An average of $4.79*10^7$ reads were obtained per sample with an average runtime of 97h:47m:48s (supplementary material 3), and over $2.79*10^7$ SNP and non-SNP variants were called across these samples (supplementary material 4). Runtime again appeared roughly linear to the number of uniquely mapped reads (Figure 2a).

3.1.2 – More differences were noted between GATK and OP pre-processed files than between CRC, NRM, and LM sample types.

To inform future studies on RNA-seq variant calling, we compared the performance of the two RNA-seq pre-processing methods. For the SK data, on average OP pre-processing took 2h:41m:03s while GATK pre-processing took 3h:50m:20s (Figure 1b). Similar was seen for the NZ data (GATK's average of 7h:23m:11s vs. OP's average of 4h:53m:32s, Figure 2b). For both data sets the step "SplitNCigarReads" (referred to as "HardClip" in Figures 1c and 2c) contributed the most to the GATK pre-processing runtimes. Variant calling for GATK BAM files also took longer generally for both data sets than for the OP BAM files with one exception described below. For both data sets, the variant calling "method" (defined as a combination of one pre-processing method and one variant calling tool) of OP pre-processing and the Platypus or "PT" (referred to as OPPT) was notably quicker in calling variants than all other methods. This method average runtime was 00h:03m:30s for the SK data and 00h:18m:52s for the NZ data, while all other methods took at least four hours to call variants (Figures 1d and 2d). The one exception described above was the variant calling method which combined OP with Mutect2 (referred to as OPMT2) which took over five hours longer than its contemporary method (GATKMT2) for the NZ data (17h:24m:06s vs.11h:39m:58s, Figure 2d).

We then investigated the statistical outputs provided by the Samtools "stats" and "flagtstat" options. When detailing a Samtools output, the tool that provided the output (i.e. stats or flagstat) will be used to prefix that output (e.g. "flagstat: properly paired", "stats: reads properly paired", etc.). For the SK data, the unprocessed BAM file outputs for "flagstat: properly paired", "flagstat: both in a pair mapped", "stats: reads mapped", "stats: reads mapped and paired", and "stats: reads properly paired" were of equal value. This value differed for GATK BAM files but was shared across the same outputs described above (supplementary material 5). Similar was observed for the NZ data: "flagstat: properly paired", "flagstat: both in a pair mapped", "stats: reads mapped and paired", and "stats: reads properly paired" were equal values within a sample, however the output for "stats: reads mapped" differed numerically (supplementary material 6).

Regardless of data set, both the unprocessed and OP BAM file outputs for "Stats: reads duplicated/Flagstat: duplicates" was 0, while the GATK BAM files provided a different value for each sample. Similarly, the OP BAM files also reported 0 values for: "Stats: non-primary alignments/Flagstat: secondary", "Stats: reads paired/Flagstat: paired in sequencing", "Stats: reads properly paired/Flagstat: properly paired", "Stats: reads mapped and paired/Flagstat: both in a pair mapped". This meant that some more meaningful comparisons between GATK and OP BAM files required using different outputs between pre-processing methods, e.g. GATK's "stats: reads mapped and paired" output against the OP output for "stats: reads mapped".

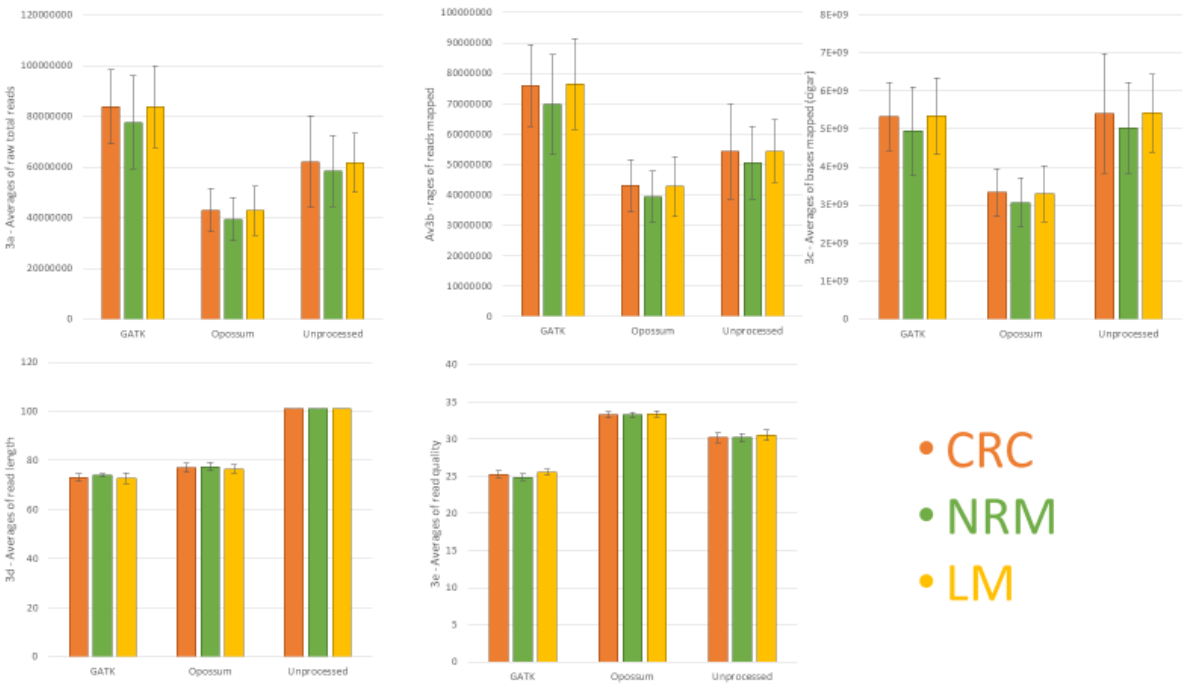

Figure 3 – Comparison of statistical outputs for the SK data's CRC, NRM, and LM BAM files. For all comparisons (3a – Averages of raw total reads, 3b – Averages of reads mapped, 3c – Averages of bases mapped (cigar), 3d – Averages of read length, and 3e – Averages of read quality), little if any differences appeared to exist between different sample types. Error bars display one standard deviation where available.

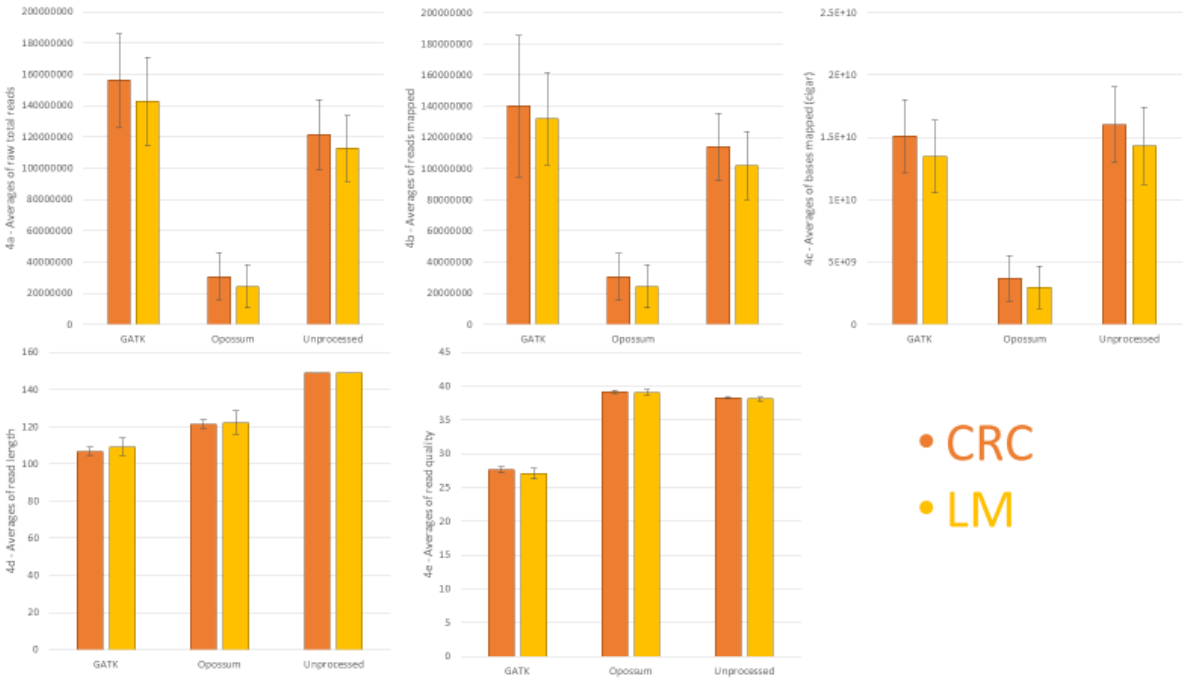

Figure 4 – Comparison of statistical outputs for the NZ data's CRC and LM BAM files. Like the SK data above, little if any differences were observed between different the two sample types for all comparisons (4a – Averages of raw total reads, 4b – Averages of reads mapped, 4c – Averages of bases mapped (cigar), 4d – Averages of read length, and 4e – Averages of read quality). Error bars display one standard deviation where available.

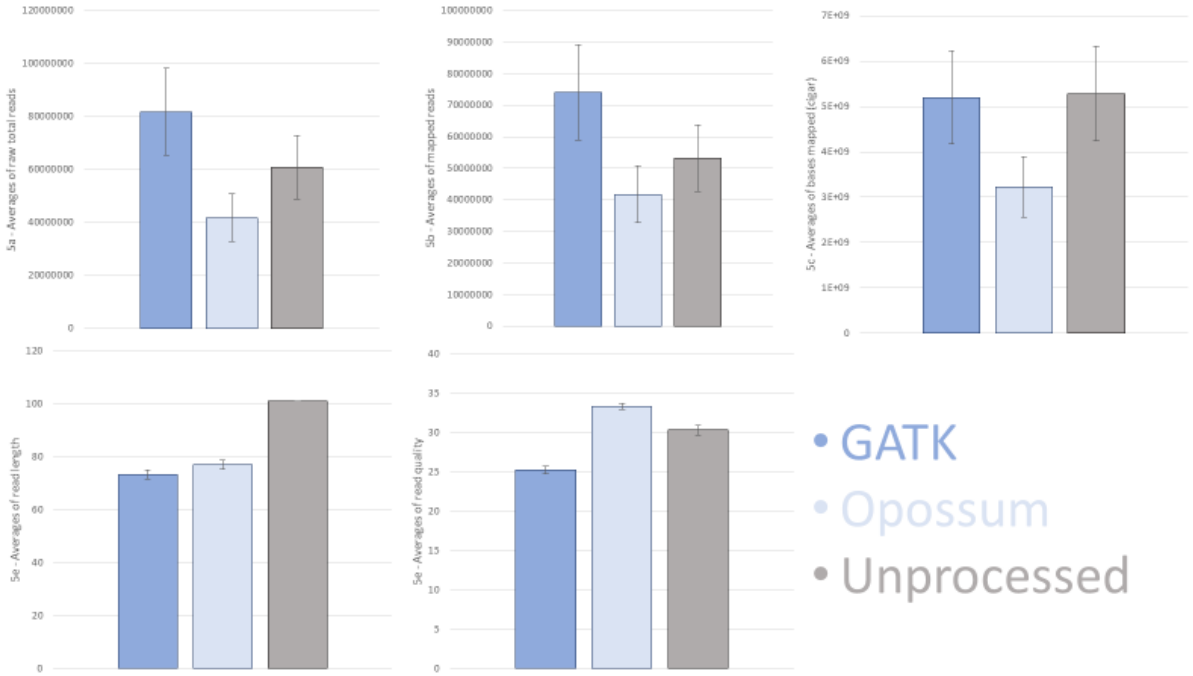

Figure 5 – Comparison of statistical outputs for the SK data's unprocessed, GATK pre-processed, and OP pre-processed BAM files. 5a – OP produced lower average of raw total reads compared to both GATK and unprocessed with GATK producing more. 5b – Similar was seen for averages of mapped reads as for raw total reads, likely because of GATK splitting reads. 5c – OP provided the lowest average for bases mapped (cigar). 5d – Both GATK and OP averages of read length were reduced compared to unprocessed data. 5e – OP produced a higher average of read quality, with GATK producing less than when unprocessed. Error bars display one standard deviation where available.

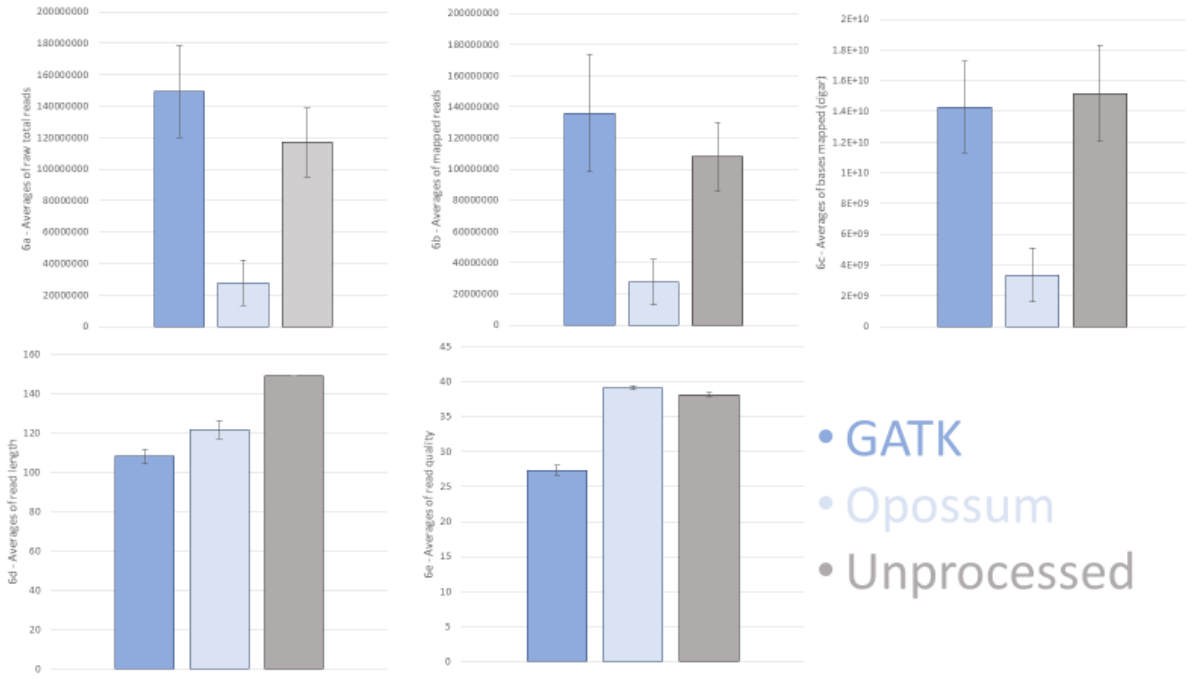

Figure 6 – Comparison of statistical outputs for the NZ data's unprocessed, GATK pre-processed, and OP pre-processed BAM files. Previous trends seen for the SK data's averages for raw total reads (6a), mapped reads (6b), and bases mapped (cigar, 6c) were more notable for the NZ data, possibly due to increased sequencing depth. Averages of read length (6d) and read quality (6e) were like what had been seen previously.

Regardless of pre-processing, no major differences appeared to exist between sample types (CRC, LM, or NRM, Figures 3a-e, 4a-e). Greater differences were observed when comparing averages between the unprocessed and pre-processed files. For the SK data, GATK consistently produced a higher average for "raw total reads", "reads mapped", and "bases mapped (cigar)" than OP for both data sets (Figures 5a-c, 6a-c). GATK's averages for "raw total reads" and "reads mapped" was also higher, though less notably, in comparison to the unprocessed averages (Figures 5a-b, 6a-b). GATK also output slightly lower averages for "bases mapped (cigar)" than the unprocessed BAM files (Figures 5c and 6c). Both GATK and OP had reduced averages for "read length" compared to the unprocessed files, with OP files showing a slightly higher average than GATK (Figures 5d and 6d). OP also displayed notably higher averages of "read quality" than GATK. For the NZ data's unprocessed and OP BAM files, the averages were more similar than for the SK data. GATK also produced slightly lower averages for "read quality" than the unprocessed files (Figures 5e and 6e).

The starkest difference between pre-processing methods were the values for "maximum read length". GATK allowed for much longer maximum lengths (SK data average $9.83*10^5$, NZ data average $1.57*10^6$) while OP appeared to have a hard threshold at 201 for the SK data, roughly double the SK data's RNA-seq 101bp read length (supplementary material 5). Like the OP BAM files, the unprocessed files also had a hard threshold for the length of the reads at 101bp. A similar but not identical observation was made for the NZ data, whose sequencing read length was 151bp. While the unprocessed BAM files reflected this (maximum read length 150bp), the OP files maximum read length ranged from 297-299bp (supplementary material 6).

3.1.3 – GATK pre-processing appears less specific/more sensitive than OP pre-processing.

To complement our above results assessing differences between pre-processing methods, we then compared the variant calling performance after pre-processing to inform future studies. We first wanted to justify our use of using an intersection of seven different methods to increase our confidence in the variants called. One way we could demonstrate this is how our intersection increased variant calling specificity which likely also limited the number of false positive calls in our results.

A custom script was used to count how many variants were called by the 127 different combinations of methods employed in our methodology. This allowed us to then count the total number of variants called for each sample, which methods contributed to those calls, and how many variants were annotated by the DBSNP. As an example, when investigating the number of variants called by the GATKFB method, we searched each of the different combinations or "sets" which included this tool ("set = gatkfb", "set = gatkfb | gatkhc", "set = Intersection", etc. supplementary materials 7-11).

Tables 2 and 3 show the number of variant calls, including non-SNP variants, made by a single tool in comparison to the intersection of all seven tools for the NZ and SK data sets, respectively. However, it must be noted that while the Freebayes (FB), HaplotypeCaller (HC), and Mutect2 (MT2) tools all had access to both pre-processed BAM files, PT only had access to one kind of

Table 2

SK dataset	CRC	NRM	LM	Total
FB	11977523	11488035	11484389	34949947
HC	2671629	2603655	2540106	7815390
MT2	1143594	1126101	1035265	3304960
PT	985104	968532	937659	2891295
Intersection	42909	42643	36423	121975

Table 3

NZ dataset	CRC	LM	Total
FB	12522716	10695267	23217983
HC	7150515	5689275	12839790
MT2	5387798	4528917	9916715
PT	2361478	1795798	4157276
Intersection	76632	60004	136636

Tables 2 and 3 – Comparison of variant calls made by individual tools compared against the intersection of seven methods for the SK and NZ datasets, respectively. Here we see that the intersection of all seven methods reduced the number of variants (which likely included false positives) in comparison to the individual variant callers.

pre-processed file (OP). With this in mind, we see that our intersection of seven callers did increase specificity and limited the number of calls in comparison to the calls made by the FB, HC, and MT2 tools. FB provided vastly more variant calls than all other tools.

Satisfied that our intersection of variant calling methods had increased specificity, we then considered how many of these variants would be likely true positives. As RNA-seq data lacks any truth sets, we decided to use existing annotations provided by the DBSNP (which can annotate both known SNP and non-SNP variants) as support for a variant being more likely a true positive call.

It must be noted that the sporadic nature of somatic variants means that they are less likely to be annotated, and so unannotated variants are not definitively false positives. This is especially true for MT2, which is more specifically a somatic variant caller and so is more likely to make calls that would not be annotated. We considered making a similar comparison using the Catalogue of Somatic Mutations in Cancer (COSMIC) annotations but the preliminary results for this approach (not included in this study) resulted in far fewer variant annotations, and showed little if any differences in the trends seen when focusing on the DBSNP annotations. For example, the variant calling method with the most DBSNP annotations was the same when looking for the most COSMIC annotations, etc.

Figure 7 shows the percentage of annotated variants (including non-SNP variants) for each of our methods against that of the intersection. Each data set (SK and NZ) were separated into sample subtypes (e.g. CRC, NRM, and LM where appropriate). Figure 7 shows that the intersection retained a reasonable percentage of annotated variants (over 60% and 40% of the known SK and NZ variants, respectively) compared to our other methods. A higher percentage of variants were annotated by both the HC and PT tools, but they also lacked the specificity of the intersection annotation. This is perhaps best demonstrated by the FB tools' lower percentage for the SK data, which was sequenced at a shallower depth compared to the NZ data, whereas our intersection's percentage increased as depth decreased. As expected, the MT2 tool provided the fewest number of annotated variants.

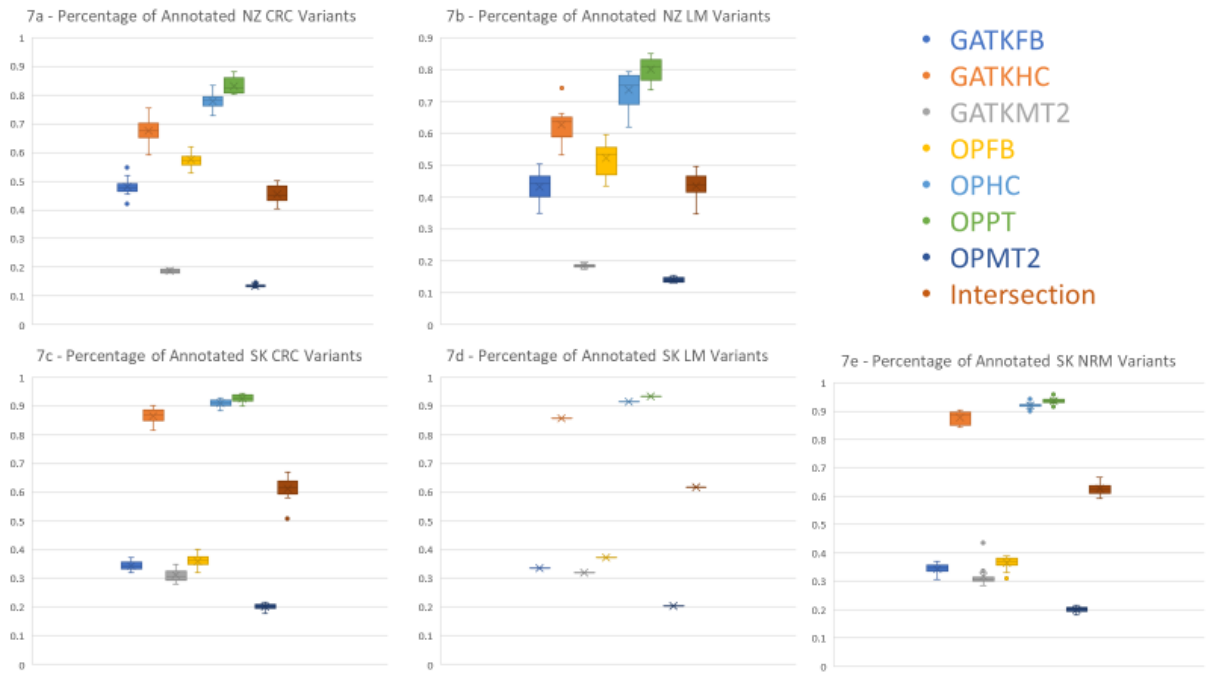

Figure 7 – Percentage of DBSNP annotated variants called by our ensemble of variant calling methods. For all comparisons, GATKHC, OPHC, and OPPT had a higher percentage of variant calls in comparison to the intersection. For the NZ samples specifically (Figs 7a, 7b), The percentage of annotated GATKFB variants was comparable to the intersection, with the OPFB method having a higher percentage of annotated variants. This changed for the SK samples (Figures 7c-e), with the lower sequencing depth seeming to influence the number of variants that could be annotated by FB. MT2 consistently annotated the fewest variants for all samples. As MT2 is primarily a somatic variant caller, it is more likely to make novel variant calls that have not been annotated by any database. While the intersection did not annotate as high a percentage of variants as some individual tools, it possessed the increased specificity discussed above and would include somatic variant calls thanks to the inclusion of MT2.

After this comparison, we then compared SNP variant calls specifically and found that, for the SK data, FB again contributed most calls to the total, approximately $2.24*10^7$. HC called an order of magnitude fewer SNP variants than FB ($4.39*10^6$), while MT2 contributed the fewest ($2.83*10^6$). Interestingly for the SK data, PT contributed the third highest number of variants ($2.89*10^6$) despite its runtime and calling variants from only OP BAM files, while all other tools had access to both pre-processed BAM files (Figure 8a).

We then considered how many of the SK SNP variants had been called in concordance between both pre-processing methods, rather than being exclusive to one method. For example, variants with both “gatk” and “op” annotations (“set = gatkfb | gatkhc | opfb”, “set = Intersection”, etc.) were considered “concordant” variant calls between pre-processing methods. Conversely, those with annotations like “set = gatkfb | gatkc”, “set = opfb | ophc”, are examples of “GATK only” or “OP only” exclusive variants, respectively.

This was followed by investigating how many SNP variants had been called by individual tools. If we used FB as our example again, a call was considered in concordance if both of that tool’s methods contributed to the call (e.g. any set annotation with both “gatkfb” and “opfb”). Discordant calls, meanwhile, possessed only one of these methods. This comparison could not be made for the PT tool which could only be used with OP pre-processed BAM files.

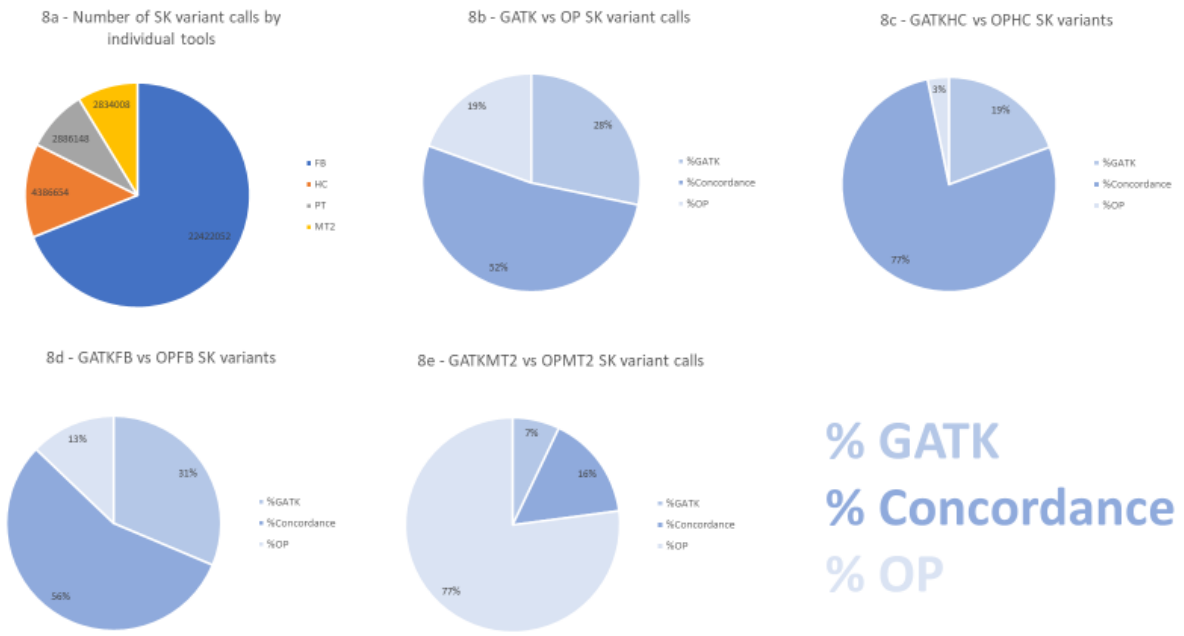

Figure 8 – SK variant calls and GATK/OP concordance. 8a – FB outperformed all other tools regarding the number of variant calls made. PT performed well despite having access to only the SK OP BAM files. 8b – All variant calls made between GATK and OP BAM files. Most calls (52%) were made in concordance between BAM files. GATK again seems more sensitive (28%) and OP more specific (19%). 8c - HC variant calls showed the highest consensus between pre-processing (77%), likely because of the HC-specific filters available to both pre-processing methods. GATK remained more sensitive than OP (19% vs. 3% respectively). 8d - FB variant calls were like the results seen for all variants: 56% in concordance, 31% GATK only and 13% OP only. 8e – Despite being part of the GATK set of tools, an overwhelming number of MT2 variant calls came from the OPMT2.

When considering all SK SNP variants, most (52%) were concordant between pre-processing methods, with 28% being GATK only and 19% OP only (Figure 8b). When considering variants that were called by HC, a much higher number were called in concordance than any other method (77%), leaving 19% as GATKHC only and 3% OPHC only (Figure 8c). The FB tool's results were akin to comparing all SNP variant calls: 56% were concordant, 31% were GATKFB only, and 13% OPFB only (Figure 8d). Despite OP being designed primarily for PT, and MT2 is a component of the Genome Analysis Toolkit, most MT2 SNP variant calls where OPMT2 only (72%), leaving 16% in concordance and 7% GATKMT2 only (Figure 8e).

For the NZ data, FB again contributed the highest number of SNP variant calls ($1.57*10^7$), which was followed by MT2 at $8.74*10^6$. HC contributed the third highest number of variants ($8.22*10^6$), and PT called the least at $4.16*10^6$ variants, roughly half that of MT2 and HC (Figure 9a). Less concordance was seen for the NZ data compared to the SK data. 40% of all variants were concordant between pre-processing methods, with both GATK-only and OP only variant calls making up 30% of the total each (Figure 9b). HC again provided the highest concordance between methods, but the NZ result was much lower than the SK result at 57%, leaving 37% as GATKHC only and 6% for OPHC only (Figure 9c). For FB, the OPFB only variants were also just 6%, with concordant calls making up 49% of the total leaving 45% to GATKFB only (Figure 9d). Little changed for the NZ MT2 variant calls compared to the SK

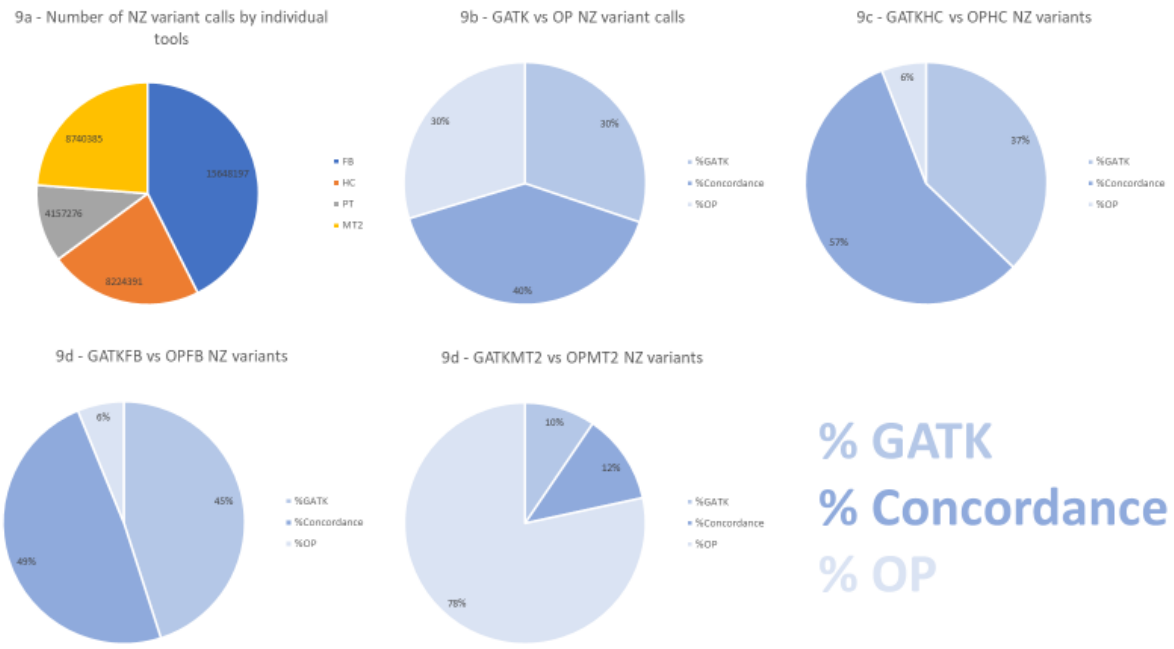

Figure 9 – NZ variant calls and GATK/OP concordance. 8a - FB again contributed the most variants calls to the NZ total. PT performance dropped to roughly half the values seen for HC and MT2. 8b – Most variants were still made in concordance between pre-processing methods, but not to the same extent seen previously. 8c – HC-specific variant calls maintained higher concordance between pre-processing methods, but was again lower compared to previous results. 8d – OP retained its specificity for FB-specific variant calls, while GATK only and consensus variant calls were now more comparable than for the SK data. 8e – The majority of MT2 variant calls still came from OP-MT2.

data with most calls being OPMT2 only (78%), leaving 12% in concordance and 10% for GATKMT2 only (Figure 9e).

3.1.4 – The ClinVar, COSMIC and DBSNP data bases helped discern between variant types and true positives.

Despite using an ensemble of seven different methods which must intersect to be considered variants of "high confidence", our results still returned a high number of SNP variant calls. We then applied various additional filters to try and limit this to a more reasonable number of variants for analysis. To achieve this, we selected for variants which: possessed high impact annotations provided by SnpEff, had been called in high confidence for at least two samples, and possessed an average QUAL value (an annotation provided by HC, retained after merging VCF files, and is indicative of variant quality) that was above that of the median QUAL score for all variants within a data set that were supported by multiple samples.

For the SK data, this returned 333 intersecting SNP variant calls that were also annotated with a "high impact" annotation via SnpEff, 167 of which were supported by more than one sample. 84 variants remained after filtering based upon the SK data's median QUAL value (QUAL = 586.9333), 73 of which were found in CRC samples, 54 in NRM, 78 in LM, and 30 were shared between both CRC and LM but not NRM (supplementary material 12). For the NZ data set's

variants, 342 were intersecting SNPs with “high impact” annotations. 107 of these variants were supported by more than one sample. 54 of these variants possessed an average QUAL value above that of the median QUAL scores for the NZ data set (QUAL = 412.7679), of which 47 of these variants were obtained in CRC samples and 53 from LM samples (supplementary material 13).

Like when we used the DBSNP to support a variant call as a likely true positive, we could also use this and other databases, such as the Clinical Variants (ClinVar) and Catalogue of Somatic Mutations in Cancer (COSMIC), to help us distinguish between germline and somatic variant calls and assess our variant calling methodology. When describing these database annotations below, DBSNP IDs were prefixed with “rs”, COSMIC IDs prefixed with “COSM”, while ClinVar IDs were a purely numerical value.

As well as explicitly describing known variants as germline/somatic or benign/pathogenic based on their annotations, the absence of any annotations and the samples from which they were called could also be useful. For example, variants called only in the SK data’s NRM samples were more likely to be benign germline variants, whereas those only called in cancerous samples were more likely to be somatic and pathogenic.

Of the SK data’s 54 NRM variants, which were more likely germline and benign rather than somatic and pathogenic, 11 had ClinVar annotations that were all were described as germline (three “benign”, four “likely benign”, three of “uncertain significance”, and one “pathogenic”). Seven of the remaining 43 variants possessed COSMIC annotations (three with multiple IDs), five of which had been excluded from the COSMIC website as part of the COSMIC team’s efforts to reduce noise from hypermutated samples [187]. For the remaining two, one had been observed as somatic but also neutral with regards to pathogenicity, while the other had been observed both somatic and pathogenic.

This left 36 variants, of which 30 possessed DBSNP IDs, leaving the remaining six unannotated. One DBSNP ID (rs1021487491) had been merged with another (rs477171), the latter of which also possessed a ClinVar ID (129251) suggesting this variant was germline. While none of the remaining DBSNP annotations provided any clinical information, they did support these variant calls as being true positives.

3.2 – Notable variants called during analysis.

Section 3.2 shifts our focus towards specific variants that were called as part of this analysis, and how they may pertain to CRC pathogenesis. We first noted frequent variant calls made to understand if they are likely pathogenic or benign. We then considered the possibility that, for the SK data, variants may have been called in “cancerous” (CRC, LM) samples with high confidence, but less stringently (in that they possessed a “set” annotation other than the Intersection) in its paired normal sample. This could then be investigated by searching for a “cancerous” variant’s position within the NRM VCF file before applying the intersection filter.

We then used gene set enrichment analysis via Enrichr to assess if our gene sets likely pertain to cancer, as well as observing which genes were frequently observed in Enrichr's oncogenic functional categories. We then used PredictSNP2 to make deleterious predictions about some of our more "novel" variant calls which lacked more definitive clinical annotations (i.e. ClinVar/COSMIC annotations explicitly stating if they were somatic and/or pathogenic) before more closely studying the individual VCF entries for those novel variants. Finally, we used differential gene expression to further investigate these more novel variants to interpret possible mechanisms by which they could exacerbate CRC pathogenesis.

3.2.1 – Some of the most frequent variant calls were germline and benign across both data sets.

We first considered genes that were frequently called for as variants in our analysis. These variants could either be common germline benign SNPs with no involvement in CRC, or somatic variants with pathogenic implications so profound that they are common to CRC cases.

Gene variants called in high frequency for the SK data set were for the *HECT, UBA and WWE domain containing E3 ubiquitin protein ligase 1 (HUWE1*, 51 samples), *glucose-6-phosphate dehydrogenase* (*G6PD*, 50 samples), and *KIAA1161* (also known as *myogenesis regulating glycosidase* or *MYORG*, 11 samples) genes. The *HUWE1* and *G6PD* gene variant's ClinVar annotations (129251 and 470162 respectively) were both described as benign. The *KIAA1161* gene variant also seems unlikely to be pathogenic given how frequently it was observed in our SK samples (19.6%) and yet possessed only a DBSNP annotation. A somatic, pathogenic variant this frequent would be unlikely to lack a clinically significant annotation.

The *TP53* gene was affected by the most diverse variants that were called in high confidence (five different for the SK data and two for the NZ data). All these SNP variants occurred between positions 7,673,793-7,674,948, and for the SK data were also absent from any SK NRM samples. One *KRAS* gene variant (position 25,245,347) was observed in high confidence for five cancerous samples for both data sets and was also absent from SK NRM samples. This result suggests that this is likely a somatic variant, which was also called less stringently in other cancerous samples. These variants for the *TP53* and *KRAS* genes will be described in more detail in sections 3.2.2 and 3.2.3 below.

The second-most frequent variant that was called exclusively in SK cancerous samples and pertained to the *fatty acid synthase* gene (four samples, position 82,084,074). However, this variant was both annotated with a germline and benign ClinVar annotation (462162) and was found less stringently called in six NRM samples before filtering. All other variants were supported by one CRC and one LM sample. Two of these variants were both annotated with benign ClinVar IDs and called less stringently in their respective NRM sample. Of the 13 variants with COSMIC IDs (excluding the *TP53*/*KRAS* variants above), six were called in NRM samples before filtering, five of which had also been excluded from the COSMIC website as described previously.

Of the seven variants that remained, five possessed DBSNP IDs (including one for *kinesin-like protein 9* or *KIF9*). However, only one variant for *aldehyde dehydrogenase 1 family member A2* (*ALDH1A2)* had not been called less stringently in any NRM samples and so was putatively somatic. The remaining two unannotated variants were called in the *calpin 2 (CAPN2)* and *scm polycomb group protein homolog 1 (SCMH1)* genes. While the *CAPN2* gene variant was called less stringently in the respective patient's NRM sample, the *SCMH1* gene variant was not and so was also putatively somatic. Two additional gene variants were exclusive to LM samples: one for *itchy E3 ubiquitin protein ligase* or *ITCH* (position 34412548), and one for *cytoplasmic FMR1 interacting protein 1* (position 22914854). That said, the *ITCH* variant's ClinVar ID (538759) suggested it was both germline and benign, and both variants had been called less stringently in their respective NRM VCF files.

As the NZ data lacked NRM samples to help identify putative germline/benign variants, we became reliant on database annotations in this regard. Of the 54 NZ variants, 37 possessed DBSNP annotations, 15 possessed COSMIC annotations, and nine with ClinVar annotations. Four of the most frequently affected gene for the NZ data set (With the fifth being *TP53*) included the same *HUWE1* (24 samples) and *G6PD* (23 samples) gene variants described above. Other frequent variants for the NZ data set included ones for the *cytochrome p450 family 3 subfamily A member 5* (*CYP3A5*, 20 samples), and *DExD-box helicase 52* (*DDX52*, 9 samples) genes, which were also annotated by the DBSNP. Like the *KIA1161* gene variant above, the frequency of both the *CYP3A5* and *DDX52* gene variants and their lack of clinically significant annotations could suggest they are not somatic and pathogenic.

Of the seven with ClinVar annotations (excluding *HUWE1* and *G6PD* above), two were described as benign, one likely benign, one of uncertain significance, one likely pathogenic, and two pathogenic. Of the 15 COSMIC annotations, two had been excluded from the COSMIC website, one of which also possessed a benign ClinVar annotation. For the remaining 13 COSMIC IDs, all but one (COSM4640837) were described as having been observed somatic while all were predicted to be pathogenic. None of the DBSNP IDs readily helped identify any of the NZ data's variants as being germline, somatic, benign, or pathogenic. More specific details for some of these variants are available in Table 4.

3.2.2 – Various oncogenic processes were enriched following gene set enrichment analysis.

In order to better understand what oncogenic processes were driving CRC and LM in our data sets, and focus on variant-affected genes that more likely to contribute to oncogenesis, we used the web-based gene set enrichment analysis tool Enrichr [180, 181].

Before compiling our sets of genes required for Enrichr analysis, genes affected by variants annotated with benign ClinVar annotations and those with COSMIC IDs that had been excluded from the COSMIC website were removed. For the SK data set, this resulted in the following gene sets: "All" (60 genes), "CRC" (54 genes), and "LM" (57 genes, supplementary material 14). The NZ data's lack of normal samples meant only two gene sets were produced for

Table 4

SK Data								
Gene	Variant	Total samples	CRC samples	LM samples	NRM samples	Known IDs	Germline/Somatic	Benign/Pathogenic
HUWE1	**X:53534126,T>C**	**51**	**18**	**17**	**16**	**rs1021487491/rs477171;129251**	**Germline**	**Benign**
G6PD	**X:154532439,A>G**	**50**	**16**	**17**	**17**	**rs2230037;470162**	**Germline**	**Benign**
KIAA1161	**9:34372875,G>C**	**11**	**4**	**4**	**3**	**rs4879782**	**Unknown**	**Unknown**
KRAS	**12:25245347,C>T**	**5**	**2**	**3**	**0**	**rs112445441;COSM1140132;12580**	**Somatic**	**Pathogenic**
FASN	17:82084074,G>A	4**	1	3	0**	rs45557233;462063	**Germline**	**Benign**
TP53*	17:7674230,C>T	2	1	1	0	rs28934575;COSM121035;12365	**Somatic**	**Pathogenic**
KIF9	3:48288340,C>A	2**	1	1	0**	rs76198671	**Unknown**	**Unknown**
ALDH1A2	15:58138445,A>G	2	1	1	0	rs1230944831	**Unknown**	**Unknown**
CAPN2	1:223755650.G>T	2**	1	1	0**	.	**Unknown**	**Unknown**
SCMH1	1:41113329,G>A	2	1	1	0	.	**Unknown**	**Unknown**
ITCH	20:34412548,T>C	2	0	2	0**	rs3761146;538759	**Germline**	**Benign**
CYFIP1	15:22914854,C>T	2	0	2	0**	rs147012760	**Unknown**	**Unknown**
NZ Data								
Gene	Variant	Total samples	CRC samples	LM samples	NRM samples	Known IDs	Germline/Somatic	Benign/Pathogenic
HUWE1	**X:53534126,T>C**	**24**	**12**	**12**	**-**	**rs1021487491/rs477171;129251**	**Germline**	**Benign**
G6PD	**X:154532439,A>G**	**23**	**12**	**11**	**-**	**rs2230037;470162**	**Germline**	**Benign**
CYP3A5	7:99672916,T>C	20	11	9	-	rs776746	**Unknown**	**Unknown**
DDX52	17:37628628,T>C	9	4	5	-	rs7224513	**Unknown**	**Unknown**
TP53*	17:7675088,C>T	6	2	4	-	rs28934578;COSM10648;12374	**Germline**	**Pathogenic**
KRAS	**12:25245347,C>T**	**5**	**3**	**2**	**-**	**rs112445441;COSM1140132;12580**	**Somatic**	**Pathogenic**
KIAA1161	**9:34372875,G>C**	**4**	**2**	**2**	**-**	**rs4879782**	**Unknown**	**Unknown**
ABHD12	20:25300697,G>A	2	1	1	-	rs1046073;337987	**Germline**	**Benign**
AMFR	16:56326967,C>T	2	1	1	-	rs759617526;COSM5513199	**Somatic**	**Pathogenic**
CARM1	19:10920494,C>T	2	1	1	-	COSM7270215	**Somatic**	**Pathogenic**
DSP	6:7584384,C>T	2	0	2	-	rs2076300;COSM150043;44947	**Germline**	**Benign**
KDM6A	X:45089908,T>C	2	1	1	-	rs142238688;471949	**Germline**	**Benign**

Table 4 – Germline/somatic and benign/pathogenic variant calls supported by known databases. Variants with coloured backgrounds were the same across both data sets. TP53* acknowledges that multiple TP53 variants were called for both data sets with only one given as an example for this table. ** refers to only high confidence calls, with lower confidence calls having been made for the respective samples highlighted.

CRC and LM samples, consisting of 42 and 46 genes respectively (supplementary material 15, Table 4). One other gene set for the SK data was proposed ("cancerous" variants shared between CRC and LM samples) but as it contained so few genes (17) we found it could not provide a meaningful gene set for enrichment analysis.

Table 5 details the functional categories that were observed as part of our Enrichr analysis. These functional categories possessed one of the lowest, if not the lowest, p-value for that functional category. We can see from this table that various cancers that were explicitly stated by Enrichr, with others having clear connotations with either the hallmarks of cancer (e.g. Panther 2016: Apoptosis signalling pathway Homo sapiens) or CRC specifically (e.g. Jensen TISSUES: Intestine).

For brevity, we then considered the three most observed genes for each Enrichr functional category. Genes involved in multiple oncogenic processes were more likely to play an important role in CRC pathogenesis. Details for how these observations were recorded are described in section 2.8.

For all gene sets, *TP53* was consistently among the most observed variant-affected genes, while the *MAPK1* gene was among the most observed for three gene sets (SK all, SK CRC, and SK LM). The *tyrosine-protein kinase Met* (*MET)* and *SWI/SNF related matrix associated actin*

Table 5						
Data set	No of genes	Functional category	Notable entry	p-value	Genes involved	Including
SK	60	BioPlanet 2019	Pathways in cancer	5.47E-06	8	MAPK1, TP53, MET, JAK1
		WikiPathways 2019 human	TCA Cycle Nutrient Utilization and Invasiveness of Ovarian Cancer	8.80E-04	2	MAPK1, JAK1
		KEGG 2019 Human	Pathways in cancer	2.67E-05	9	MAPK1, TP53, MET, JAK1
		BioCarta 2016	Stat3 Signalling Pathway Homo sapiens	2.45E-04	2	MAPK1, JAK1
		Reactome 2016	TNFR1-induced NFkappaB signalling pathway Homo sapiens	6.35E-05	3	OTUD7B
		NCI-Nature 2016	IFN-gamma pathway Homo sapiens	2.34E-04	3	MAPK1, JAK1
		Panther 2016	Apoptosis signalling pathway Homo sapiens	1.37E-05	5	MAPK1, TP53
		GO Biological Process 2018	Negative regulation of apoptotic process	1.33E-05	9	TP53, MET, JAK1
		GO Cellular Component 2018	Focal adhesion	1.06E-05	8	CAPN2, MAPK1, JAK1, TP53, MET
		Jensen TISSUES	Intestine	3.58E-09	38	CAPN2, MAPK1, TP5, MET
		ClinVar 2019	Heptocelular carcinoma	2.13E-06	3	MET, TP53
		DisGeNET	Mammary neoplasms	1.02E-09	26	CAPN2, MAPK1, JAK1, TP53, MET
		OMIM Expanded	Hepatocellular carcinoma	1.24E-04	4	TP53, MET, JAK1
		Rare diseases GeneRIF ARCHS4 Predictions	Malignant cylindroma	7.95E-09	9	CAPN2
		Rare Diseases GeneRIF Gene Lists	Malignant mesothelioma	2.10E-06	6	MAPK1, TP53, MET
		Rare Diseases AutoRIF Gene Lists	Primary effusion lymphoma	3.94E-05	5	MUM1, MAPK1, TP53, JAK1
NZ	46	BioPlanet 2019	Wnt signaling pathway	4.13E-07	7	AKT1, TP53
		WikiPathways 2019 human	Oncostatin M Signaling Pathway	1.50E-05	4	AKT1, KRAS, TP53
		KEGG 2019 Human	Central carbon metabolism in cancer	1.50E-05	4	AKT1, KRAS, TP53
		BioCarta 2016	Apoptotic Signaling in Response to DNA Damage Homo sapiens	4.07E-06	3	AKT1, TP53
		Reactome 2016	Deactivation of the beta-catenin transactivating complex Homo sapiens	2.57E-06	4	XPO1, AKT1
		NCI-Nature 2016	Glucocorticoid receptor regulatory network Homo sapiens	1.23E-06	5	AKT1, TP53
		Panther 2016	p53 pathway feedback loops 2 Homo sapiens	1.51E-04	3	AKT1, KRAS, TP53
		GO Cellular Component 2018	Chromatin	5.69E-05	6	TP53, MET, JAK1
		Human Phenotype Ontology	Transitional cell carcinoma of the bladder	1.70E-05	3	AKT1, KRAS, TP53
		Jensen TISSUES	Intestine	1.11E-08	31	XPO1, AKT1, SHISA5, KRAS, TP53
		Jensen DISEASES	Thyroid cancer	9.07E-06	3	AKT1, KRAS, TP53
		ClinVar 2019	Neoplasm of the breast	3.20E-05	3	AKT1, KRAS, TP53
		DisGeNET	Granulosa cell tumour	1.36E-07	5	AKT1, KRAS, TP53
		OMIM Disease	Breast cancer	1.88E-03	2	AKT1, TP53
		Rare Diseeases GeneRIF ARCHS4 Predictions	Rectosigmoid neoplasm	6.23E-06	6	XPO1, AKT1
		Rare Diseases GeneRIF Gene Lists	Pancreatic cancer	6.65E-08	7	AKT1, KRAS, TP53

Table 5 – Notable enriched functional categories obtained from Enrichr gene set enrichment analysis. The 60 and 46 genes included for the SK and NZ gene sets respectively excluded variants annotated as benign/likely benign from their ClinVar annotations of COSMIC IDs that had been excluded from the respective website. The gene names specified in the right-most column are those that become more relevant in later sections of our analysis.

dependant regulator of chromatin subfamily A member 4 (SMARCA4) genes were one of the most observed for two different gene sets each. *MET* being frequently observed for the SK all and SK LM gene sets, while *SMARCA4* was frequently observed for both NZ gene sets. This left the *Janus kinase 1 (JAK1)*, *histone deacetylase 1 (HDAC1)*, and the *RAC-alpha serine/threonine protein kinase 1* (*AKT1)* genes, which all were among the most observed affected genes for one gene set each (SK CRC, NZ CRC, and NZ LM respectively, Table 6).

The *TP53* gene was found to possess multiple variants in our results. For the SK data, all but one of the five *TP53* variants possessed both DBSNP and ClinVar IDs, while all possessed COSMIC annotations. For those five variants with ClinVar annotations (12365, 127814, 127819, 182935, and 376617), all but one SNP was described as either pathogenic or likely pathogenic, containing a mix of germline and somatic submissions. The variant that was not explicitly stated as either pathogenic or likely pathogenic (182935) possessed five submissions described as germline (three likely pathogenic, one pathogenic, and one unknown), and five somatic (all likely pathogenic).

The SK *TP53* gene variant with only COSMIC annotations (e.g. COSM10756) was described as observed both somatic and pathogenic. For the NZ data, one *TP53* gene variant possessed DBSNP, COSMIC and ClinVar IDs (ID 12374) which had been called in six samples and described as both germline and pathogenic. The remaining *TP53* variant was annotated with only COSMIC IDs (e.g. COSM5755152) was both observed somatic and pathogenic.

Table 6							
Gene	SK All	SK CRC	SK LM	NZ CRC	NZ LM	Annotations	Significance
TP53	19	20	22	18	23	DBSNP, COSMIC, ClinVar	Mostly somatic*, pathogenic
MAPK1	22	17	18	-	-	-	-
MET	21	-	21	-	-	DBSNP, COSMIC, ClinVar	Both germline/somatic, pathogenic/uncertain significance
SMARCA4	-	-	-	21	1	DBSNP, ClinVar	Germline, uncertain signifiance
JAK1	-	16	-	-	-	DBSNP	-
AKT1	-	-	-	-	27	-	-
HDAC1	-	-	-	-	-	DBSNP	-

Table 6 – Enrichr counts between the SK all, SK CRC, SK LM, NZ CRC, and NZ LM gene sets. Gene names were counted if observed within the Enrichr functional category that pertain to human genes, cancer or related processes, and with a p-value below 0.01. *TP53* was the most frequently observed gene across all gene sets, followed by *MAPK1*. TP53 also possessed the most variants for those genes and so possessed multiple annotations (mostly somatic, with one germline variant) with explicitly pathogenic annotations. While *MET* and *SMARCA4* are both genes with proto-onco implications, both variants appear to be under investigation given their uncertain significance annotation. No other frequently observed variants possessed explicitly pathogenic annotations.

All the remaining genes with high Enrichr counts were affected by only one high confidence variant. The *MET* gene variant possessed DBSNP, COSMIC, and ClinVar annotations. The COSMIC annotation (COSM3995297) described the variant as having been observed both somatic and pathogenic, while its ClinVar ID (411875) described it as both germline and of uncertain significance. The *SMARCA4* gene variant's ClinVar annotation (486461) was also observed as germline and of uncertain significance. Of the remaining four gene variants without COSMIC/ClinVar annotations, both *JAK1* and *HDAC1* possessed only DBSNP annotations (rs758343641 and rs779483240 respectively), leaving the *MAPK1* and *AKT1* gene variants completely unannotated. Both the *MAPK1* and *AKT1* gene variants were seemingly "novel" given their lack of explicitly pathogenic annotations, i.e. ClinVar/COSMIC IDs that explicitly stated some degree of pathogenicity.

3.2.3 – Some variant calls were "unanimously" predicted deleterious by PredictSNP2.

After finding some putatively "novel" variants described above, i.e. those affecting known oncogenic targets and which lacked explicitly pathogenic annotations provided by COSMIC and ClinVar, we then used the absence of these annotations as an opportunity to find more novel variants which may play a role in CRC pathogenesis.

For variants without the above annotations, we used PredictSNP2 [182], a web-based tool that predicts the deleteriousness of genetic variants, to compliment the functional impacts already predicted by SnpEff. As this tool uses an ensemble of tools to predict deleteriousness, variants that were predicted as deleterious by all of PredictSNP2's tools were referred to as "unanimously" deleterious.

Initially we assessed the accuracy of PredictSNP2 by using two known pathogenic variants for the *KRAS* and *TP53* genes. Our logic was that if PredictSNP2 predicted these known pathogenic variants as deleterious, this would support the accuracy of the tool. Table 7 shows that both the *TP53* and *KRAS* gene variants were predicted as deleterious with the same expected accuracy of 87%. Of the five individual tools comprising PredictSNP2, *TP53* gene variant was

Table 7									
Gene	Variant	Region	Function	PredictSNP2	CADD	DANN	FATHMM	FunSeq2	GWAVA
MAPK1	22:21769268,T>C	exonic	nonsynonymous	87%	53%	67%	83%	62%	51%
KRAS	12:25245347,C>T	exonic	nonsynonymous	87%	84%	70%	76%	62%	?
TP53	17:7674230,C>T	exonic	nonsynonymous	87%	84%	71%	83%	61%	52%
SCMH1	1:41113329,G>A	exonic	synonymous	93%	88%	95%	92%	93%	66%
ALDH1A2	15:58138445,A>G	UTR5	-	88%	80%	75%	91%	83%	84%

Table 7 – Initial assessment of PredictSNP2's deleterious predictions. Two known pathogenic gene variants for *KRAS* and *TP53* were used to assess the accuracy of the PredictSNP2 tool, which were both correctly predicted to be deleterious. When then included the *SCMH1*, *ALDH1A2*, and *MAPK1* gene variants for assessment as the former two variants were found exclusive for the SK data's cancerous samples regardless of variant filtering, the latter was called in NRM samples but also affected a known oncogenic target, and all three lacked explicitly pathogenic database annotations. While the *SCMH1* and *ALDH1A2* gene variants were predicted non-deleterious (indicating PredictSNP2's ability to make predictions that are not deleterious), the *MAPK1* variant was predicted as deleterious as both the *KRAS* and *TP53* gene variants. This table accurately reflects the output of PredictSNP2, which did not provided any functional details for the *ALDH1A2* gene variant.

unanimously predicted deleterious, while the *KRAS* gene variant was predicted deleterious by four tools with the fifth providing an ambiguous prediction.

Table 7 also provides deleterious predictions for the *MAPK1, SCMH1* and *ALDH1A2* gene variants previously described. The unannotated *MAPK1* gene variant was included in this initial analysis as it affected a known oncogenic target but lacked a clinically significant annotation. The latter two variants were included as they were also unannotated and putatively somatic genetic variants.

We found that not only was the *MAPK1* gene variant predicted as deleterious as the known pathogenic *TP53* and *KRAS* gene variants described above, it also possessed the same expected accuracy of deleteriousness (87%). With regards to the *SCMH1* and *ALDH1A2* gene variants, both returned overall non-deleterious predictions, although this result was not unanimous among PredictSNP2's individual tools.

All the novel high confidence variant calls from both the SK and NZ data sets were then retrieved and had their deleteriousness predicted by PredictSNP2. This included COSMIC variants excluded from the website and ClinVar annotations either of uncertain significance or conflicting reports of pathogenicity.

For the SK data, 57 variants were obtained that lacked explicitly pathogenic annotations. Five of these variants were found to be incompatible with PredictSNP2 (including the previously described and frequently observed *HUWE1, G6PD, and KIAA1161* gene variants) due to conflicting reference alleles between our reference and the available GRCh38 references used by PredictSNP2. For the remaining SK novel variants, 28 were overall predicted non-deleterious, one of uncertain deleteriousness, leaving 23 predicted overall as deleterious (supplementary material 16).

Of those 23, nine gene variants were predicted unanimously by PredictSNP2 as deleterious: *filamin A* (*FLNA*, ClinVar 452140), *multiple myeloma oncogene 1* (*MUM1*, also known as *interferon regulating factor 4* or *IRF4), CAPN2, MAPK1, JAK1, KIF9, MET* (ClinVar 411875), *ovarian tumour deubiquitinase 7B* (*OTUD7B),* and *receptor-type tyrosine-protein phosphatase F*

Table 8								
SK data	Region	Function	PredictSNP2	CADD	DANN	FATHMM	FunSeq2	GWAVA
MUM1	exonic	synonymous	99%	82%	99%	57%	99%	58%
CAPN2	splicing	-	89%	50%	54%	55%	74%	66%
MAPK1	exonic	nonsynonymous	87%	53%	67%	83%	62%	51%
JAK1	exonic	nonsynonymous	87%	60%	60%	62%	61%	51%
KIF9	exonic	nonsynonymous	87%	84%	71%	56%	62%	51%
OTUD7B	exonic	stopgain	81%	65%	73%	62%	65%	76%
PRPTF	exonic	nonsynonymous	87%	84%	75%	83%	62%	51%
NZ data	Region	Function	PredictSNP2	CADD	DANN	FATHMM	FunSeq2	GWAVA
PHYH	exonic	nonsynonymous	87%	77%	66%	63%	61%	52%
XPO1	exonic	nonsynonymous	87%	84%	77%	83%	62%	50%
PHF23	exonic	stopgain	81%	58%	61%	60%	65%	76%
SHISA5	exonic	stopgain	81%	55%	57%	67%	65%	65%

Table 8 – PredictSNP2 deleterious predictions for both data sets more novel variants. Variants called in high confidence for both datasets which lacked explicitly pathogenic database annotations provided by ClinVar and COSMIC were considered further with PredictSNP2. This table details only the variants with unanimous deleterious predictions for each dataset, i.e. predicted as deleterious by all tools that encompass PredictSNP2: seven for the SK data (*MUM1*, *CAPN2*, *MAPK1*, *JAK1*, *KIF9*, *OTUD7B*, and *PTPRF*) and four for the NZ data (*PHYH*, *XPO1*, *PHF23*, and *SHISA5*).

(*PTPRF*). As the *FLNA* and *MET* variants both had ClinVar IDs of uncertain significance and so are likely being investigated elsewhere, they were not considered further in this study.

For the NZ data, 37 variants lacked explicitly pathogenic annotations, of which six (including the *CYP3A5* and *DDX52* gene variants) were found to be incompatible with PredictSNP2 for the same reasons as described above. Of the remaining 31, 16 were predicted non-deleterious and three of uncertain deleteriousness. This left 12 variants overall predicted as deleterious (supplementary material 17), of which four were predicted unanimously deleterious by PredictSNP2: *phytanoyl-CoA 2-hydroxylase* (*PHYH*), *exportin 1* (*XPO1*), *plant homeodomain finger protein 23* (*PHF23*), and *shisa family member 5* (*SHISA5*). Table 8 below provides the PredictSNP2 results for each of the novel and unanimously predicted deleterious variants considered for this analysis.

3.2.4 – Most novel variants were described as either "stop gained" or "structural interaction" variants often affecting multiple transcripts.

Following the PredictSNP2 results, the VCF entries for each unanimously predicted deleterious variant were inspected more closely. The details contained within these files, when coupled with the expression data further below, would help inform us as to how these variants could advance oncogenesis, i.e. if the expression data matches the impact annotation afforded that variant. General details for these variants can be found in Table 9. Often a file possessed annotations

Table 9			
SK variant	DBSNP	High Impact	Other annotations
MUM1	rs1323312366	Stop gained:	Synonymous low impact: Tyr677Tyr;
		Gln703*	Modifier impacts with NMD
CAPN2	-	Splice Donor	LOF
MAPK1	-	Structural interaction:	Missense moderate impact: Lys340Arg/Lys296Arg;
		1TVO/1WZY/406E	Sequence feature with low impact
JAK1	rs758343641	Structural interaction:	Missense moderate impact: Asp895Glu
		3EYG/4E4L/4E4N/	
		4E5W/4EHZ/4EI4/	
		4FK6/4I5C/4IVB/	
		4IVC/4IVD/4K6Z/	
		4K77	
KIF9	rs76198671	Structural interaction:	Missense moderate impact: Arg12Leu/Arg26Leu;
		3NWN	Sequence feature with low impact
OTUD7B	rs782649998	Stop gained:	
		Arg490*	
PTPRF	rs17849101	Structural interaction:	Excluded from COSMIC webite
	COSM7244022	4N5U	Missense moderate impact: Arg635Cys;
NZ variant	DBSNP	High Impact	Other annotations
PHF23	-	Stop gained:	Modifier impacts with NMD;
		Arg58*	LOF
		Arg63*	
		Arg67*	
		Arg71*	
SHISA5	rs769953108	Stop gained:	Synonymous low impact: Cys14Cys;
		Gln91*	Modifier impacts with NMD;
		Gln163*	LOF
		Gln187*	
		Gln194*	
XPO1	-	Structural interaction:	Missense moderate impact: Arg417Cys
		3GB8	Modifier impacts with NMD;
PHYH	rs200245065	Structural interaction:	Missense moderate impact: Val231Leu/Val314Leu/Val331Leu
		2A1X	

Table 9 – VCF file entry and SnpEff impact annotation details for unanimously predicted deleterious variants. The *PTPRF* gene variant's COSMIC annotation had been excluded from the website, and one of the available DBSNP annotations had been linked with ClinVar. Discrepancies were observed between some variants that shared the same high impact annotation (e.g. "stop gained") but lacked other details such as expected "NMD" annotations.

for multiple "transcripts", which could refer to different splice isoforms, different starting positions, overlapping genes, etc. Occasionally the VCF files also provided details regarding the crystal structure of the gene's protein product.

For the SK variants (supplementary material 18), the *MUM1* gene variant had 13 transcripts of which two were annotated as a high impact "stop gained" variants. These both described the same protein change ("Gln730*"). One other transcript was annotated as a "low" impact synonymous variant, leaving the remaining ten transcripts annotated as "modifier" variants, unevenly split between an upstream gene variant, a downstream gene variant, and a non-coding

transcript exon variant. Three of these modifying variants (two downstream gene and one non-coding transcript exon) also pertained to a different gene, *AC004623.2*.

Our results for the *CAPN2* gene variant had six transcripts, three of which were annotated as high impact "splice donor and intron" variants, with the other three all being modifying upstream gene variants. The *CAPN2* variant was also annotated with an additional loss-of-function (LOF) for this gene after its list of transcript annotations ("LOF=CAPN2|ENSG00000162909|12|0.17").

The *MAPK1* gene variant had results for 12 different transcripts. Six of these detailed a high impact "structural interaction variant" for one of three different crystal structures: "1TVO:A_59-A:340", 1WZY:A_59-A340", and "406E:A_60-A340". Three other *MAPK1* gene transcripts also had three "moderate" impact missense mutation describing one of two different amino acid changes, Lys340Arg and Lys296Arg.

The *JAK1* gene variant's results had 28 transcripts with 24 providing annotations for high impact "structural interaction variants". These 24 transcripts were unevenly divided between 13 different crystal structures: 3EYG, 4E4L, 4E4N, 4E5W, 4EHZ, 4EI4, 4FK6, 4I5C, 4IVB, 4IVC, 4IVD, 4K6Z, and 4K77. The *JAK1* variant's VCF file also described a moderate impact missense variant for a "Asp895Glu" amino acid change. The remaining three transcripts were all modifying variants.

Our results for the *KIF9* gene variant had 26 transcripts with 12 detailing high impact "structural interaction variants". There were evenly split between three KIF9 "2NWN" crystal structures: "A_12-A_315", "A_12-A_64", and "A_12-A_66". Nine of the remaining transcripts also described a "moderate" impact missense variant with an "Arg12Leu" amino acid change. One of these moderate variants (ENST00000443784.5) also provided an annotation for nonsense-mediated decay (NMD). Two transcripts were of a "low" impact and provided an annotation describing a "nucleotide phosphate binding region" sequence feature, while the remaining transcripts were all modifying variants.

The remaining SK gene variants to consider pertained to the *OTUD7B* and *PTPRF* genes. The *OTUD7B* gene variant's results had two transcripts, one of which was a high impact stop gained variant. This annotation detailed an "Arg490*" change, while the remaining transcript was a modifying downstream gene variant. The results for the *PTPRF* gene variant possessed nine transcripts, two of which provided annotations for a high impact "structural interaction variant" for the same "4N5U:A_606-A_635" crystal structure. Three other transcripts described moderate impact missense variants split between two different amino acid changes, "Arg635Cys" and "Arg291Cys", leaving the remaining modifying transcripts evenly split between an upstream gene variant and an intron variant.

For the NZ data's variants (supplementary material 19), the *PHYH* gene variant's results provided only four transcript annotations which all related to *PHYH.* Only one annotation was given a high impact, which was a structural interaction for the crystal structure "2A1X:A_151-A_331". The remaining three transcripts were all missense variants with a moderate impact for three different amino acid changes: "Val331Leu", "Val231Leu", and "Val314Leu".

The *XPO1* gene variant's results provided annotations for 17 transcripts. Six of these were high impact structural interaction variants split evenly across two different changes for the 3GB8 crystal structure ("A_417-A_464", and "A_417-A_471"). Three transcripts detailed the same moderate missense variant for the amino acid change Arg417Cys. *XPO1* was also annotated with NMD for two specific transcript: a modifying 3' UTR variant, and downstream gene variants. However, NMD was not provided as an annotation at the end of the list of annotated transcripts as seen for the *PHF23* and *SHISA5* gene variants described below.

The *PHF23* gene variant results pertained to 26 transcripts, with 13 referring to other genes (six for *segment polarity protein dishevelled homolog DVL-2*, six for *gamma-aminobutyric acid receptor-associated protein*, and one for *CTD-2545G14.7*). Of the transcripts regarding *PHF23*, five were annotated as high impact "stop gained" variants with four different changes: "Arg67*" (two transcripts), "Arg63*", Arg58*", and "Arg71*". The PHF23 gene variant's results were also annotated with LOF and NMD annotations. This NMD annotation differed from what was seen for the *KIF9* gene's results in that it was it was found at the end of the list of annotations and was not associated with any one transcript.

Finally, the *SHISA5* gene variant's results provided 31 transcripts, 17 of which related to *SHISA5*, with the others all being modifying downstream gene variants for the *ataxia telangiectasia and Rad3 related-interacting protein* gene and the *three prime repair exonuclease 1* gene. A "Stop gained" high impact variant were provided for six of the *SHISA5* transcripts, which provided details on four different amino acid changes: "Gln194*", "Gln91*", "Gln163*", and "Gln187*". The *SHISA5* results also provided annotations for both LOF and NMD annotations. Unlike previous NMD annotations, the *SHISA5* gene variant's results provided an NMD annotation both at the end of its list of transcripts and for specific transcripts: a 3' UTR transcript, and two downstream gene variants, which were all modifying variants.

Interestingly, all but one of these NZ variants came from the same patient (samples CRC337/LM337). Closer inspection of the variant calls for the NZ data's different samples and patients revealed that the CRC/LM337 samples provided far more variants in comparison to other samples. With an average of 7.96 variants per sample, both the CRC337 and LM337 samples possessed 28 high confidence, high impact SNP variant calls each.

3.2.5 – More differential expression was seen between normal and primary tumour samples.

Given the nature of our data (RNA-seq) for both data sets, we investigated the expressions for those genes which had been affected by the above novel variants called in our analysis. Differential gene expression was performed using a combination of Subread's "featureCounts" for gene expression quantification and DESeq2 for calling statistically differentially expressed genes. Comparisons were made between the SK data's CRC vs. NRM, CRC vs. LM, and the NZ data's CRC vs. LM.

9651 genes were found to have significantly disrupted expressions for the SK data's CRC vs. NRM comparison (29.1%). Meanwhile, only 3316 genes (10.0%) were significantly disturbed

for the SK data's CRC vs. LM comparison and only 1373 genes (4.4%) whose expression was significantly disrupted for the NZ data's CRC vs. LM comparison (supplementary materials 20 and 21).

Tables 10 and 11 show the fold changes of expression between the SK data's NRM vs. CRC, and the NZ data's CRC vs. LM, respectively. Using an adjusted p-value threshold of 0.05, we found that none of the genes affected by the novel NZ variants were significantly different in their expression. Similar was observed for the SK CRC vs. LM comparison. We did see a significant difference (i.e. adjusted p-values below 0.05, highlighted in red in Table 10) for each of the genes affected by novel variants for the SK data (*MUM1*, referred to as *IRF4* in Table 10, *CAPN2*, *MAPK1*, *JAK1*, *KIF9*, *OTUD7B* and *PTPRF*).

Table 10

SK	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj	weight
IRF4	406.2584002	1.81088047	0.37078192	5.235522892	1.65E-07	3.09E-06	1.633191834
CAPN2	6760.106822	0.935180641	0.187040778	5.067723109	4.03E-07	8.75E-06	1.121150765
MAPK1	2350.091977	0.232021937	0.085232488	2.72693636	6.39E-03	2.10E-02	1.271268855
JAK1	3252.021424	0.244909352	0.088984193	2.758516494	5.81E-03	2.20E-02	1.091160255
KIF9	161.9234794	-0.612232848	0.16686869	-3.627968682	2.86E-04	1.79E-03	1.130335945
OTUD7B	1104.776539	0.46170555	0.098938519	4.677958263	2.90E-06	3.57E-05	1.443894561
PTPRF	14440.35052	0.870560297	0.161259654	5.4709404	4.48E-08	1.53E-06	1.091160255

Table 11

NZ	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj	weight
PHF23	817.4975312	0.303658128	0.164423648	1.836611616	6.63E-02	2.97E-01	1.554095305
SHISA5	5245.358612	0.21715024	0.197965193	1.131390155	2.58E-01	5.93E-01	1.522691836
XPO1	4935.542661	0.048776901	0.088533469	0.5529768	5.80E-01	8.80E-01	1.33810142
PHYH	845.5732747	-0.106638021	0.206265096	-0.520404864	6.03E-01	7.89E-01	1.842669537

Tables 10 and 11 – Variant expression for the SK and NZ data sets, respectively. These tables provide some of the statistics obtained during our differential expression analysis. We see that for all the novel variants obtained from the SK data, the genes that were affected by those variants had their expression disrupted for the SK data, suggesting that are targets for CRC oncogenesis. For the NZ data there were no significant changes in the log fold of expression for those genes for which novel variants had been called, with their adjusted p-values highlighted to as they were below 0.05. Adjusted p-values below 0.05 are highlighted in red. For Table 10, *IRF4* refers to *MUM1*.

As seen in the supplementary materials 20 and 21, we did also find that for the SK data's CRC vs. NRM comparison, the *SHISA5*, *XPO1*, and *PHYH* gene expressions were significantly perturbed with adjusted p-values of $1.66*10^{-6}$, $7.71*10^{-6}$, and $9.12*10^{-3}$, respectively. While the SK data did not appear affected by variants for these genes, these were all genes for which a novel variant had been called for within the NZ data. For the NZ data, both the *MUM1* (adjusted p-value $5.53*10^{-3}$, also referred to as *IRF4* in the supplementary material 21) and *KIF9* (adjusted p-value $1.80*10^{-2}$) genes, which lacked the variant calls as found in the SK data, also had significantly perturbed expressions.

As all SK data's novel variants found their expressions were significantly disrupted when comparing the SK data's CRC and NRM samples, we then compared the expression for the samples containing those variants against the averages of expression for that gene within the SK

data set. Units provided were “transcripts per million” or “TPM” (supplementary material 22, with the NZ expression values also being available in supplementary material 23).

As all seven of these novel variants were putatively germline variants, it was unlikely that they alone would notably disrupt normal gene expression. That said, it is possible that these more minor alleles could also be risk alleles. We therefore considered it possible that these gene variants could be aggravated following oncogenesis and lead to more significant differences in gene expression.

However, Figure 10 shows that most of the samples which contained their respective novel variant possessed expression values within one standard deviation of the average. One exception to this trend was the *JAK1* gene variant, whose gene expression was more than one standard deviation below that of the *JAK1* gene’s mean expression for both the NRM and CRC samples. We also saw that for the NRM sample affected by the *CAPN2* gene variant, its expression was slightly above one standard deviation of the average of *CAPN2* gene expression, although this difference was not observed for its paired CRC sample.

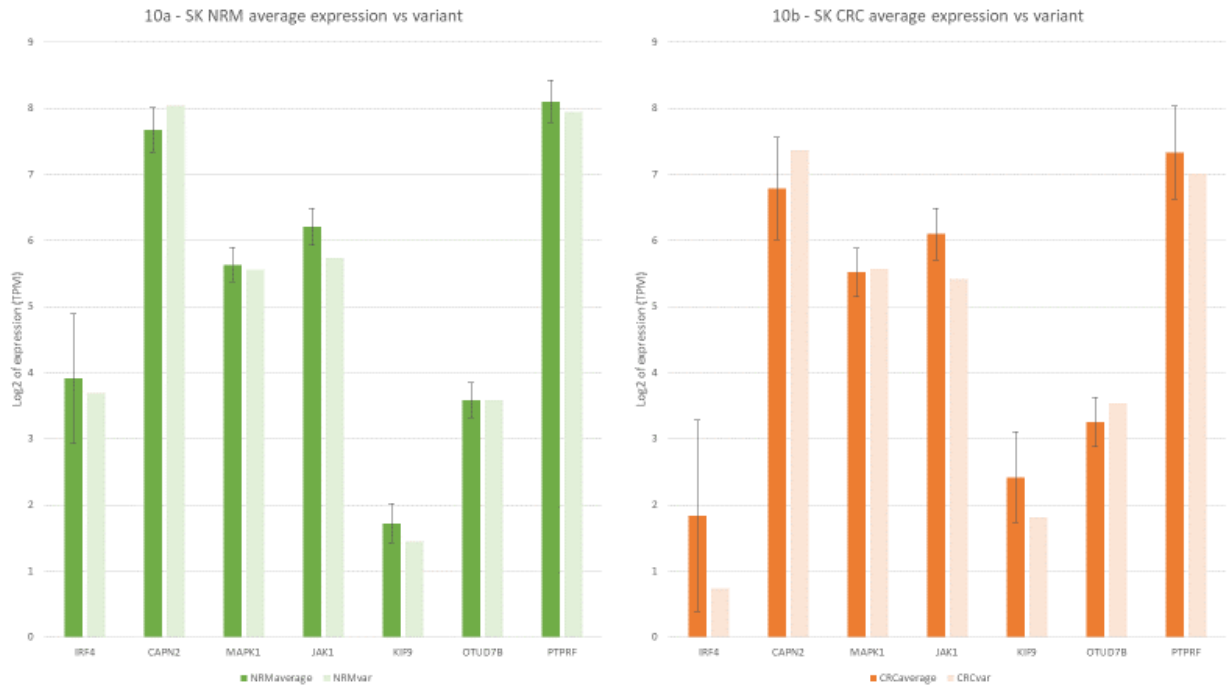

Figure 10 – Differences in expression for the SK genes affected by novel variants. 10a – average of expression vs. variant for NRM samples. We see that little difference is seen in expression between our variant samples and the average for that gene obtained from the data set. That said, *CAPN2* gene variant expression does seem slightly elevated and *JAK1* gene variant expression does seem slightly below that seen for their respective averages. 10b – Average of expression vs. variant for CRC samples. Again, we see little difference between a variant’s sample and the average obtained from the SK data’s CRC samples. CAPN2 variant expression now exists within one standard deviation of the average, whereas the *JAK1* gene variant continues to be below that seen for the average of expression. These results suggest that these genes for samples not affected by novel variants may be similarly disrupted by other means.

4.0 – Discussion.

CRC is one of the most prevalent cancers worldwide. Even in developed regions like New Zealand, CRC rates of incidence remain high. Despite this, the pathogenesis of this disease remains not well understood. Therefore, we have used two RNA-seq data sets, one from a cohort of NZ patients and the other from a publicly available South Korean data set, to try to identify in high confidence genetic variants that may further understanding of the disease.

We developed an RNA-seq variant calling pipeline with an emphasis on making high confidence variant calls, using two different RNA-seq pre-processing steps. These were the "GATK best practises" (referred to as GATK) and the Opossum pre-processing tool, referred to as OP. We also employed an ensemble of variant calling "methods", defined as a combination of one of the above pre-processing methods and one of the following variant calling tools: Freebayes (FB), HaplotypeCaller (HC), Platypus (PT), or MuTect2 (MT2). This ensemble helped increase our confidence in the variants called with those that intersected, i.e. called by all seven variant calling methods, being considered as our variants of "highest confidence".

The pipeline was also developed to make variant calls without a reliance on normal samples. Such samples, often used for variant calling studies in cancer, were not available for the NZ data set. However, normal samples were available for the SK data. Their inclusion helped us to assess if our current methodology, using known variant databases (including the DBSNP, COSMIC, and ClinVar), helped to distinguish between germline, somatic, benign, and pathogenic variants.

Unfortunately, no variants were identified as being exclusive to metastatic samples, which we had hoped would improve knowledge of the metastatic process. However, we did have success in calling known germline and somatic variants, annotated with the above data bases, for known cancer targets such as *TP53*, *KRAS*, etc. Some high confidence variant calls were also made for some more "novel" variants for both data sets, i.e. those with annotations not explicitly pathogenic, or those that lacked annotations entirely. We will now discuss in detail both the performance of our two employed pre-processing methods, our ensemble of variant callers, and some noteworthy variant calls made as part of our analysis.

When discussing performance, we will first consider the statistical outputs obtained from both data sets, then compare pre-processing methods. We then conclude discussing performance by considering the individual tools, their strengths and weaknesses, and help justify our use of the intersection of all seven methods. When considering the noteworthy variants, we will describe how we came to focus on a handful of variants from numerous calls, how these variants were annotated, how expression for those variant-affected genes may or may not have been perturbed, before considering how these variants specifically may or may not be involved in CRC pathogenesis.

4.1 – Overview of the current RNA-seq high confidence variant calling pipeline.

As variant calling from RNA-seq data alone remains a developing field where no standard exists, making high confidence variant calls required us to develop the pipeline using a strict set of

filtering criteria. The purpose of this was to have the highest confidence in the variants called as being true positives and avoid errors brought about by sequencing artefacts, known short comings in current best practises, or tool biases. Not addressing these issues could distract us from considering variants that are potentially involved in carcinogenesis. We decided to use a reference-based variant calling method, which aligns RNA-seq reads with a reference genome, so that variants between the data and the reference can be identified. This alignment produced a sequence/alignment map file (SAM) for each sample and, in our case, we used the binary equivalent (BAM) to minimise our footprint on the available computer resources.

Which reference to use depended partly on which bioinformatic tools we employed. We ultimately decided upon the most recent release of the human genome, GRCh38 reference. More specifically, we used the "GRCh38_no_alt_plus_hs38d1_analysis_set" which we refer to as the "reference" or "chosen reference". GRCh38 alone has several benefits over the older release, GRCh37, already described in our introduction, while our chosen reference also has other benefits relevant to our interests. The "GRCh38_no_alt_plus_hs38d1_analysis_set" for example includes masks for the two pseudo-autosomal regions (PAR) on chromosome Y. As these PAR regions are already represented on chromosome X, representing both would mean any reads that map to these areas would be considered ambiguous multi-mapping reads. Regardless of such reads containing true positive variants, these reads would likely be disregarded from variant calling because of their ambiguous nature.

While the chosen reference contains masks to hide duplicate regions, GRCh37 also has its own masked regions such as for the centromeric arrays for chromosomes 5, 14, 19, 21, and 22. The repetitive nature of these regions in this older reference meant that originally, they were difficult to accurately sequence resulting in them being masked. The improved sequencing technologies available when constructing the GRCh38 build, including the chosen reference, meant that GRCh38 now more accurately represents these arrays and allows for more true positive variants to be called in these regions [112, 153, 156].

Additionally, the chosen reference contained non-human sequences for the Epstein-Barr virus (EBV) which is sometimes obtained during sequencing, either because of natural infection or if EBV was used to immortalise a cell line. The chosen reference also contained "human decoy sequences", either partially assembled sequences missing from the primary GRCh38 assembly [112] or repetitive non-centromeric sequences that are, again, difficult to align reads to. Both the EBV and decoy sequences acted as "sinks" for false positive variant calls. Without these sequences available, some sequencing reads may partially and incorrectly aligned to similar sequences elsewhere in the genome, again increasing the number of ambiguous multi-mapping reads.

The "full analysis set" GRCh38 release was not chosen, as this contained alternative contigs as detailed in the reference's "README" file [151]. These alternative contigs are not currently compatible with our chosen read alignment tool, the Spliced Transcripts Alignment to a Reference (STAR [142]). This tool was chosen given its popularity with regards to mapping RNA-seq reads, its recommendations for some RNA-seq best practises with GATK, and our existing familiarity with the tool. Researchers interested in using the "full analysis set" reference

would therefore need to use an alternative alignment tool like HISAT2 [188] or BWA-MEM [189]. The benefits the "full analysis" reference may have over our chosen reference is that it provides known germline variants that exist between populations, reducing false positives.

Once the reference was decided upon and our BAM files were produced, we then required variants to be present in BAM files that had been pre-processed specifically for RNA-seq variant calling using two independent methods: the Broad Institute's "GATK best practises", and the "Opossum" (OP) pre-processing tool. The Broad Institute has begun developing a set of "best practises" for calling variants from RNA-seq data using their Genome Analysis Toolkit (GATK). However, their methodology, currently still in development, is reliant on the use of hard filters that can arbitrarily include false positives or exclude true positives. The OP tool offers an alternative method of RNA-seq pre-processing for variant calling, more specifically designed for the Platypus (PT) variant caller but is also compatible with other variant calling tools. It was therefore less likely (though still possible) that variants present in BAM files pre-processed by two different RNA-seq pre-processing methods would be false positives.

An ensemble of up to four different variant calling tools was then used to call variants between the different BAM files. Like using two pre-processing methods, if the same variant is called between two independent variant calling tools, it is less likely to be a false positive call, with more tools adding more support for that variant. Additionally, as variant calling tools use BAM files to call variants, and for each sample we had two BAM files (one GATK pre-processed and the other OP), each of our variant calling tools (except for PT explained below) could call variants from two different BAM sources.

Following the logic that a variant's presence in multiple pre-processed BAM files, and called by multiple variant callers, to be more likely true positives, we therefore considered variants that had been called by an "intersection" of all our variant calling "methods" (defined as a combination of one pre-processing method and one variant calling tool) as our variants of "highest confidence". Seven methods were used in this study: GATKFB, GATKHC, GATKMT2, OPFB, OPHC, OPMT2, and OPPT. Initially an eighth method was considered which would have combined both GATK pre-processing and the PT variant caller. However, we found that GATK pre-processing alone could not overcome the fundamental compatibility issues between PT and RNA-seq data, for which the OP tool was explicitly designed. Each of these methods produced a single "Variant Calling Format" (VCF) file for each sample. These VCF files we then merged into one single VCF per sample. This merger then annotated variants based on which combination or "set" of methods made their call, with the highest confidence variants called by all seven methods possessing the unique annotation "set = Intersection".

We decided upon SNP variants as the primary focus of this study as their simplicity make them easier variants to identify. Variants were filtered further based on possessing a "high functional impact" annotation, provided by the SnpEff tool, which would be more likely to contribute to CRC development. Variants were then annotated with multiple known variant databases: the Single Nucleotide Polymorphism Database (DBSNP, which also includes small-scale multi-base deletions or insertions, retrotransposons, and microsatellite repeat variants), the Catalogue of Somatic Mutations in Cancer (COSMIC), the Clinical Variants database (ClinVar), the Exome

Aggregation Consortium (ExAC), and the Genome Aggregation Database (GNOMAD). With regards to our study, it became apparent that the DBSNP, COSMIC, and ClinVar databases were more relevant to our needs. As a result, the ExAC and GNOMAD databases shall not be referenced again and should be excluded from future pipelines to reduced sample processing time.

Despite all our filtering efforts so far, we still managed to call a substantial number of variants from both our data sets, As a result we made further filtering efforts. We first selected for variants that were present in multiple samples (>1) for three reasons. First, these variants could be germline variants or "risk alleles" that worsen oncogenesis. Second, if these variants are frequently observed between different patients, they could serve as biomarkers or even therapeutic targets. Third, these variants could be so significant to pathogenesis of the disease that they are actively preserved through various stages.

We then decided to filter the remaining variants based upon if their average QUAL score (a HC annotation retained after VCF merger that also indicates a variant's quality) was above that of the median QUAL score for all variants within a data set (SK median QUAL = 586.9333, NZ median QUAL = 412.7679).

We appreciate that this high specificity (i.e. likelihood of true positive variant calls) likely makes high sacrifices for variant calling sensitivity (i.e. the ability to make variant calls). Future efforts can be made to better increase the pipeline's sensitivity whilst also maintaining high specificity. Future work will also benefit from the ongoing development of RNA-seq calling best practises and software when making variant calls from such data.

4.2 – Different sequencing depth had notable outcomes on our results.

Both the SK and NZ data sets produced over one billion reads each, with between 68-87.8% of those reads being mapped uniquely to the chosen reference. Our variant calling pipeline was able to call almost 3.70 million variants across the 56 SK samples, while the NZ data produced nearer 2.79 million from 26 samples (supplementary materials 2 and 4). The NZ data's samples were more deeply sequenced in comparison to the SK data. This explains why so many variants were called for the NZ data despite having far fewer samples compared to the SK data.

That said, the NZ sequencing depth did vary considerably. One NZ sample, CRC335, produced notably fewer sequencing reads (11.8 million) than other samples within the NZ data set (average 46.4 million). This discrepancy led to this sample's exclusion from further analysis, along with its paired metastases (LM335), which left us with 24 NZ samples.

A samples' runtime was roughly linear with the number of sequencing reads for that sample, regardless of data sets. Figures 1a and 2a illustrate this relationship and compares a sample's uniquely mapped reads to its runtime. Our stringent variant calling pipeline, developed to overcome known issues when calling variants with RNA-seq, resulted in a computationally expensive pipeline. Average run times for the SK and NZ data sets were approximately 46 and

98 hours respectively. The result for the NZ data demonstrates one of the more dramatic effects the increased sequencing depth, with average runtime almost doubling as a result.

Considering the average runtime for our data sets, we then broke down which components of our pipeline contributed most to these result. Figures 1b and 2b show the difference in average runtime between pre-processing methods, with OP being notably quicker for both data sets. OP pre-processing appeared to be much more efficient with an average runtime of 2h:41m:03s and 4h:53m:20s for the SK and NZ data sets, respectively.

Meanwhile, the GATK best practises had notably longer average runtimes, at 3h:50m:20s for the SK data and 7h:23m:11s for the NZ data. The NZ data's longer runtimes are again likely because of the increased depth at which it was sequenced. Considering the individual components of the GATK best practises for RNA-seq variant calling, the "SplitNCigarReads" tool (referred to as "HardClip" in Figures 1c and 2c) contributed the most to GATK's pre-processing runtime. This tool is responsible for splitting reads into exonic segments, to avoid some issues with making variant calls around splice junctions.

"SplitNCigarReads" is likely also responsible for the increased number of reads for the GATK pre-processed BAM files, described below. However, it is worth noting that we used an older version of this tool, available in version 3.8 of GATK. While a new version of this tool is available in version four of GATK, and may perform better in comparison, we could not find any documentation supporting version four of the tool as having been verified [169].

We also compared the average runtime for each of our variant calling methods (Figures 1d and 2d). Generally, all the OP pre-processed methods called variants quicker than their GATK contemporaries, with one exception. The OPMT2 method took much longer to process than GATKMT2 for the NZ data set (17h:24m:06s vs. 11h:39m:58s, Figure 2d). This was also notably longer to process than all other methods, regardless of data set. As MT2 is primarily a somatic variant caller, and the NZ data was sequenced more deeply than the SK data, this difference in depth is likely the cause of the increased runtime for this set of tools.

For both data sets, the OPPT method took substantially less time to process than all other methods, taking a matter of minutes to call variants compared to the hours taken by other tools. This quick processing speed initially made us sceptical as to how many variants could have been called in such a short period of time and will be discussed further below.

4.3 – Opossum pre-processing appears both more specific/less sensitive than GATK and less computationally expensive.

As part of our GATK vs. OP comparisons, we compared the general statistics obtained by the Samtools "stats" and "flagtstat" options for both data sets. This was to see if there were any notable differences between the results obtained for each sample type, i.e. primary colorectal cancer tumours (CRC), liver metastases (LM) or normal samples (NRM) where appropriate, between the pre-processed and unprocessed BAM files.

At times, comparing GATK and OP BAM files could be difficult due to different methodology for these approaches to RNA-seq pre-processing. For example, one aspect of RNA-seq pre-processing is the marking of "duplicates", i.e. where the same RNA-seq fragment is duplicated multiple times during sequencing amplification. As such, a duplicate value of 0 for the unprocessed files is expected. However, for the OP pre-processed files a value of 0 was also reported, meaning that duplicate reads are discarded from OP's outputs. This differs from the GATK approach, whose output retains marked duplicates for further processing downstream. OP also reported 0 values for the Samtools stats/flagstat outputs for: "non-primary alignments/secondary reads", "reads paired/paired in sequencing", "reads properly paired/properly paired", and "reads mapped and paired/both in a pair mapped". This meant a more meaningful comparison between OP and GATK statistics could require different outputs, i.e. OP's "raw total sequences" against GATK's "reads mapped/mapped and paired/properly paired".

For the SK data's CRC, LM, and NRM samples (Figures 3a-e) and for the NZ data's CRC and LM samples (Figures 4a-e), no notable differences were observed between the averages of a samples' "raw total reads", "reads mapped", "bases mapped (cigar)", "read length", or "read quality". Contrasting this, comparisons between the unprocessed, GATK pre-processed, and OP pre-processed BAM files using the same statistical outputs above did reveal some differences. For both data sets, the averages for the OP BAM files "raw total reads", "reads mapped", and "bases mapped (cigar)" were consistently less than the average values for both the unprocessed and GATK BAM files (Figures 5a-c, 6a-c). Meanwhile, for both data sets the averages of "read length" were higher for unprocessed BAM files (Figures 5d and 6d), and the averages of "read quality" were higher for the OP BAM files (Figures 5e and 6e).

As mentioned above, the GATK pre-processed files had more reads than both the OP pre-processed and unprocessed BAM files. This was again likely because of the GATK method of splitting reads into exons. This increased number of reads also likely contributed to the GATK variant calling runtimes being longer than OP. Understandably, more reads would require more time to process, and suggests the GATK method of pre-processing is more sensitive while OP is more specific. This sensitivity vs. specificity is also reflected in the average quality of a pre-processed read, with GATK often being a lower quality than both OP and the unprocessed BAM files.

Another notable difference between GATK and OP BAM files were the average and maximum read lengths. The read lengths during sequencing were 100bp for the SK data, and 150bp for the NZ data. This was reflected in the unprocessed files, i.e. an average/maximum read length of 101bp/101bp for the SK data, and 149bp/150bp for the NZ data. However, while for both the GATK and OP pre-processed files average read length never exceeded the actual length, the maximum read lengths differed significantly between pre-processing methods. For OP, maximum read length was always roughly double that of the actual read length (e.g. 201bp for the SK data, and approximately 303bp for the NZ data), however the GATK maximum read lengths were well into the millions of bp. This is again likely a result of GATK splitting reads

into exons which may then mapped to distant parts of the genome, and further demonstrates GATK's sensitivity vs. OP's specificity in pre-processing.

4.4 – Retaining likely true positives using intersections vs. individual variant calling tools.

Before discussing some of the more notable variants called in our analysis, we wished to justify our use of the "set = Intersection" filter for these variant calls. We used this filter as it was intuitive that more variant calling methods making the same variant call would support this call being a true positive. However, given the sheer number of variants called, the intersection filter was also a way of drastically limiting the number of variants we would consider, as well as avoiding distraction from false positive calls.

As RNA-seq data lacks any "truth sets" akin to those available for methods based on DNA-seq data, we had to develop an alternative form of testing. We decided to use annotations provided by the DBSNP as support for variant calls being true positives. It was unlikely that a false positive variant, called as part of our analysis, would also be annotated with a known DBSNP ID. That said, it must be noted that many true positives likely exist that are yet to be known by any database, and so the absence of any annotations is not indicative of a variant being a false positive. This is especially true for somatic mutations given their sporadic nature, and so is also true for variants called by MT2, which is more specifically a somatic variant caller.

We also considered using the COSMIC database in much the same way to provide support for any true positive somatic variants calls. However, preliminary tests found that COSMIC annotated variants showed the same trends as was seen when annotating with the DBSNP (i.e. PT methods had the higher percentage of annotated variants, MT2 methods the fewest, etc. Tables 2, 3, and Figure 7). Additionally, far fewer variants were annotated when using the COSMIC database, and we also found some COSMIC annotations had been excluded from the COSMIC website. It was revealed that these excluded annotations came from hypermutated cancerous samples (over 15,000 mutations). Their exclusion was part of the COSMIC team's efforts to reduce noise in this database, given the difficulty distinguishing between variants that have arisen from hypermutation, germline variation, or technical artefacts [187].

Tables 2 and 3, respective to the SK and NZ data sets, show the number of variants called by our individual variant calling tools and our intersection filter. We see that this filter lowered the number of variants called by at least an order of magnitude than when using an individual tool. The FB tool also demonstrated a sensitivity that is somewhat notorious in the variant calling community. That said, it must be noted that the FB tool lacked any kind of variant filtering that is available for these other tools. HC benefited from the RNA-seq and SNP variant calling best practises developed by the Broad Institute/ Meanwhile, MT2 has its own filtering tool included in the GATK tool set ("FilterMutectCalls"), and the OP pre-processing method was designed specifically for the PT variant caller. Any variants called by FB alone, therefore, may suffer from the tool's high sensitivity and likely includes many false positives. This result helps justify the use of an ensemble of variant callers, which can complement each tools' strengths and weaknesses, such as improving specificity for overly sensitive tools.

While these results demonstrated that the intersection filter increased specificity, Figures 7a-e demonstrate how this filter also retains a reasonable number of annotated variants, i.e. putative true positives, when compared to our seven different variant calling methods. It is also noteworthy that the intersection retained similar percentages of annotated variants regardless of data set and sequencing depth. Meanwhile, both FB methods underperformed for the SK data when compared to the NZ data. While both HC methods and the OPPT method had a higher percentage of annotated variants compared to the intersection, they also lack the specificity discussed above. As expected, both MT2 methods had the lowest percentage of annotated variants, likely because of the sporadic nature of somatic mutation. This could also partly explain why the intersection did not perform as well as the HC and PT methods, as the intersection includes calls made by MT2, and so also included novel somatic variants that, despite being true positives, my lack any known variant annotations.

As mentioned above, OPPT's quick runtime initially seemed suspicious and made us sceptical as to how many variant calls it could have made. However, Tables 2 and 3 show that, despite its runtime, a reasonable number of variants, including non-SNP variants, were called using that tool. Also, for the SK data, PT contributed the third highest number of SNP variants (approximately $2.88*10^6$) toward the total, below the number of calls made by HC (approximately $4.39*10^6$) but above that of MT2 (approximately $2.83*10^6$). FB again demonstrated its high sensitivity here, calling an order of magnitude more SNP variants than these other tools (approximately $2.24*10^7$, Figure 8a). This result for PT was surprising, not only considering its runtime, but also because PT only had access to the OP BAM files. All other variant calling methods had access to both GATK and OP BAM files.

While not as impressive, for the NZ data OPPT still made approximately $4.16*10^6$ SNP variant calls. This was roughly half of the HC and MT2 results (approximately $8.22*10^6$ and $8.74*10^6$ SNP variants respectively, Figure 9a). FB again called notably more SNP variants than the other tools (approximately $1.56*10^7$). OPPT calling roughly half the variants of the HC and MT2 tools is at least comparable in terms of performance, if we again consider OPPT had access to only half of the BAM files. This does call into question why OPPT's performance differed so greatly between the SK and NZ data. It is possible that the OPPT method retains the same level of sensitivity and specificity regardless of sequencing depth, meanwhile the HC and MT2 tools' sensitivity increases as depth also increases. It is worth mentioning that the original OP publication also found their combination of OP and PT was faster (7 hours) than when combining HC with either GATK and OP pre-processing methods (14h:45m and 15h:35m, respectively [167]).

To assess OP and GATK pre-processing method performance, the number of "concordant" variant calls was compared between these methods. This was done for all variants called within a data set, and between individual variant callers within a data set. Considering the sheer number of calls made, for brevity we first considered a variant call as "concordant" if that call was made by any combination of GATK and OP pre-processing methods (e.g. possessed set annotations like the intersection", "gatkfb-ophc", etc.). Calls were therefore considered "discordant" if they had been made following only one pre-processing method (e.g. "set = gatkfb-gatkhc" would be

a discordant “GATK only” call). When considering pre-processing concordance of an individual variant caller, variants were considered concordant if both of that tool’s methods made the same call, with “set = gatkfb-opfb” being an example of an FB concordant call. Discordant variant calls therefore were made following only one respective pre-processing method.

For the SK data, a general concordance was observed with 52% of the variants being called by any combination of GATK and OP methods. More variants were called by only GATK methods, 28%, with 19% being OP only (Figure 8b). The HC variant caller showed the highest concordance of any tool (77%, Figure 8c). This result was not unexpected, given that the GATK best practise filters for RNA-seq and SNP variant calling with HC could be applied to both of our pre-processed BAM files. The results seen for the FB tool were like when comparing all variant calls. This again demonstrates the sensitivity of the tool and how many variants it contributed to the total (Figure 8d). 56% of the variants called by FB were in concordance, with 31% being “GATK only” and 19% being “OP only”.

Juxtaposed with the general concordance seen with these other tools, we see that the number of variants called by the MT2 variant caller were overwhelmingly “OP only” at 77%. This left just 16% called in concordance, with 7% being GATK only (Figure 8e). This is a cause for concern and could suggest there is a fundamental incompatibility between the OP pre-processing tool and the somatic MT2 variant caller. However, unlike the incompatibility seen between the GATK pre-processing method and the PT variant caller, no obvious errors were obtained as a result of using the OPMT2 variant calling method.

Our comparisons for the NZ data where subtly different than those seen for the SK data. Considering all variants, more calls were still concordant, although this was less than half of all variant calls (40%, Figure 9b). The HC method still saw the highest amount of concordance but to a lesser extent than what was observed for the SK data, with 57% in concordance, 37% GATKHC only, and just 6% OPHC only (Figure 9c). FB’s result also changed considerably, with 49% in concordance, 45% GATKFB only, and only 6% OPFB only (Figure 9d) . This difference in results could be a result of the OPPT method’s poorer performance for the NZ data compared to the SK data. This would mean that less concordance could be obtained between methods. With the main difference between the SK and NZ data being sequencing depth, this could suggest that the OPPT method is more effective at shallower depths than the other tools.

While HC and FB’s concordance results for the NZ data differed from the SK data, for the MT2 tool little changed. The number of OPMT2 only variant calls was still notably higher (78%) than either those called in concordance (12%), or those called by GATKMT2 only (10%, Figure 9e). This again raises some concerns regarding the OPMT2 method as, for the NZ data, the OPMT2 method took significantly longer to call variants than any other method as mentioned previously. This means that if these variant calls are problematic, the final output of the pipeline is also delayed possibly as a function of sequencing depth. Further work would then need to investigate this variant calling method thoroughly before including it in any future pipelines.

4.5 – Genes frequently affected by variations called during analysis were often benign germline variants or known oncogenic targets.

On concluding our GATK vs. OP pre-processing comparisons and assessing our ensemble of variant calling tools, we then considered how to investigate the variants that we had been to understand how they could contribute to oncogenesis. With so many variants to consider, we first had to decide on which variants to focus on. Given that we had already used known variant databases to annotate our variants, this gave us the opportunity to disregard variants that had already been discovered to be benign in previous studies. Table 4 provides some variant details with noteworthy annotations obtained from our analysis.

One result of interest was that two variants had been found in very high frequency (at least 88%) for both data sets. A T>C transition at position 53,534,126 for the *HECT, UBA and WWE domain containing E3 ubiquitin protein ligase 1* (*HUWE1*) gene had been called in 51 of the 56 SK data samples, while the same variant had been called in all 24 of the NZ data samples, excluding CRC335/LM335 as previously described. The other variant was a A>G transition at position 154,532,439 for the *glucose-6-phopsphate dehydrogenase* (*G6PD*) gene, which was called in 50 for the SK samples and 23 of the NZ samples.

Both variants possessed a DBSNP (rs1021487491/rs477171 for *HUWE1* and rs2230037 for *G6PD*) and a ClinVar annotation (129251 and 470162, respectively). The evidence on the ClinVar website for the *HUWE1* gene variant suggested its frequency was too high to be a pathogenic mutation. Similar was said for the *G6PD* gene variant, whose allele frequency was at least 78.4% for the databases GNOMAD, ExAC, and the 1000 Genomes Project. Another variant had also been called in both data sets with a higher frequency for one of the data sets. A G>C transition at position 34,372,875 for the *KIAA1161* gene , also known as *myogenesis regulating glycosidase* or *MYORG*, was obtained from 11 SK samples. This variant was also called in the NZ data (four samples).

For only the SK data set, one other variant frequently observed (four samples) was for the *fatty acid synthase* (*FASN*) gene. However, this variant possessed a ClinVar annotation (462063) that explicitly described this variant as both germline and benign. For only the NZ data set, two other high frequency variants of note were a T>C transition at position 99,672,916 for the *cytochrome P450 family 3 subfamily A member 5* (*CYP3A5*, 20 samples), and a T>C transition at position 37,628,628 for the *DExD-box helicase 52* (*DDX52*, nine samples).

While the *KIAA1161*, *CYP3A5* and *DDX52* did not possess ClinVar IDs, all three did possess DBSNP IDs. The *KIA1161* annotation (rs4879782) suggested this is a minor allele, i.e. the second most frequent variant for a given SNP [190], from the Korean Genome Project at approximately 5.35%. Similarly, both the *CYP3A5* and *DDX52* gene variants were reported as having minor allele frequencies, or MAF, of between 11.8-40.0% and 23.2-46.3% respectively. Given these frequencies, we believe these variants are also unlikely to contribute to CRC pathogenesis.

The final frequently observed variant was a C>T transition at position 25,245,347 for the *Kirsten rat sarcoma 2 viral oncogene homolog* (*KRAS*). Like the *HUWE1*, *G6PD*, and *KIAA1161* gene

variants, this *KRAS* gene variant was also called across both data sets with five samples each. This variant possessed annotations from all three databases. However, unlike the other variants, the *KRAS* gene variant was described explicitly as both somatic and pathogenic by both its COSMIC (COSM1140132) and ClinVar (12580) annotations. While *KRAS* is a known oncogenic target, this frequency in in our results, across geographically distinct demographics, could suggest that this *KRAS* variant may play a more specific role in CRC pathogenesis.

While no one variant was as common as those described above, the *tumour protein 53* (*TP53*) gene was frequently affected in that it it was affected by multiple different variants. Of the five different high confidence variants called for the SK data, all but one possessed DBSNP, COSMIC, and ClinVar IDs. The remaining variant possessed only COSMIC IDs that had not been excluded from the COSMIC website. The NZ data obtained two variants called in high confidence, one with annotations from DBSNP, COSMIC, and ClinVar, while the remaining again possessed only COSMIC annotations not excluded from the website. Most of these variants had been observed as both somatic and pathogenic, although there were some variants that were both germline and pathogenic. *TP53* is one of the more quintessential oncogenic targets for any type of cancer, meaning the frequency of its variants, both in terms of the number of samples and the number of different variants, in our results was not surprising.

The remaining variants to be discussed where not observed as frequently as those above, having been observed in no more than two samples each. Of those with ClinVar IDs, all had been described as both germline and benign. Those with COSMIC IDs had been described as observed both somatic and pathogenic. This left us with several genetic variants that lacked annotations with more clinically explicit annotations, i.e. those that described variants as germline/somatic/benign/pathogenic. This included variants for the *kinesin family member 9 (KIF9), calpin 2 (CAPN2), scm polycomb group protein homolog 1 (SCMH1),* and *cytoplastic FMR1 interacting protein 1* (*CYFIP1)* genes seen in Table 4. Of those examples with DBSNP annotations, all possessed MAF's below 0.05.

Table 4 also reveals that some germline/benign variants had been called, paradoxically, only in high confidence for cancerous samples. Given the highly specific nature of our variant calling pipeline, and the knowledge that hard filters can arbitrarily exclude true positive calls, we were aware the same could happen here. It was possible that the same variant may have been called in one sample by all seven variant calling methods and called in other samples with fewer methods. This would mean the "less stringently" called variants would have been filtered out after applying the high confidence filter. It is unlikely that the same variant called with high confidence for one sample (and so more likely a true positive) could also be a false positive in another sample because it was called less stringently.

To investigate this, we returned to some of the VCF files that had been retained before applying our high confidence "set = Intersection" filter. For variants called in high confidence for the SK data's cancerous samples, we could search within the respective NRM VCF file, using that variant's nucleotide position, to confirm if it had been called only in cancerous samples. However, we could not do that same for the NZ data's variants, given this data sets lack of

normal samples. As a a result, in this regard we entirely reliant on known database annotations for any paradoxical variant calls.

This analysis revealed that of those SK variants detailed in Table 4, such as the *KIF9*, *CAPN2*, *ITCH*, and *CYFIP1* gene variants, all been called for less stringently in their respective NRM samples. This suggests that these variants were germline, but this otherwise does not suggest if they are benign or pathogenic. For example, it is possible that these putative germline variants are also minor alleles, i.e. those with MAF below 50%, which may increase the likelihood that they are also risk alleles for disease [191]. Therefore, some of these variants were considered further as part of our analysis.

4.6 – Enrichr showed an enrichment for oncogenic processes, with some oncogenic targets affected by novel variation.

Given the number of variants obtained without clinically explicit annotations, we decided to use gene set enrichment analysis as a method to identify which of these variants may be more likely involved with cancer. For each data set, a list of genes affected by variants was compiled and uploaded to Enrichr, a web-based gene-set enrichment analysis program. The SK data's gene set included 60 genes, while the NZ data provided a gene set of 46. These gene sets included variants without clinically explicit annotations, to ensure Enrichr was provided with enough genes to allow for meaningful enrichment analysis. Table 5 provides details on some functional categories that were observed as part of this analysis, with the p-value provided being one of lowest, if not the lowest, p-value seen for that functional category. We found several processes had become enriched for both gene sets, either with oncogenic features, or more specific to CRC (such as the Jensen TISSUES category "intestine").

As Enrichr allows users to observe which genes are involved in these functional categories, we then made counts for how often these genes were observed in these categories. The purpose of this was to find genes affected by variants which possessed a high number of counts. Those with higher counts could then be more involved in various cancerous processes.

Table 6 shows the results of this analysis. The *TP53*, *mitogen-activated protein kinase 1* (*MAPK1*), *MET* (also known as the *hepatocyte growth factor receptor* or *HGFR*), *SWI/SNF related matrix associated actin dependant regulator of chromatin subfamily A member 4* (*SMARCA4*), *Janus kinase 1* (*JAK1*), *RAC-alpha serine/threonine protein kinase 1* (*AKT1*), and *histone deacetylase 1* (*HDAC1*) genes all possessed some of the highest counts for our gene sets. Many are all genes with well documented oncogenic functions. Many also possessed known database annotations, however it was also interesting to find that some had been affected by variants which lacked database annotations. such as *MAPK1* and *AKT1*. These could be examples of new or "novel" oncogenic variants worthy of further investigation.

4.7 – Some novel variants had possible pathogenic significance.

Given that some novel variants had been found which lacked more clinically explicit annotations provided by the COSMIC or ClinVar databases, we considered the possibility that these could be new variants with possible implications in the pathogenesis of CRC. As a result, we investigated these variants more closely.

For all variants called as part of our pipeline, the SnpEff tool to provide some putative impacts which we could filter against. For example, we considered those annotated with "high" impacts as being more likely involved in cancer development. However, as with most of our analyses we were concerned that relying on SnpEff's functional predictions alone could be misleading, especially when considering novel variant calls not studied in-depth elsewhere. To address this issue, we used a second web-based deleterious prediction tool, PredictSNP2, to compliment the predictions provided by SnpEff. PredictSNP2 makes deleterious SNP predictions based on an ensemble of predictive tools, which includes the Functional Analysis Through Hidden Markov Models (FATHMM) used by COSMIC.

Initially, PredictSNP2's accuracy was assessed using two known pathogenic variants already called in our analysis. These variants were the *KRAS* gene variant (ClinVar 12580), and one *TP53* gene variant (ClinVar 12365). As can be seen in Table 7, both variants were correctly identified as overall deleterious, demonstrating PredictSNP2's ability to make accurate predictions of deleteriousness.

Table 7 also provides the PredictSNP2 results for three other variants (*MAPK1*, *SCMH1*, and *ALDH1A2*) called in our analysis. The *MAPK1* gene variant was included in this initial analysis as it was a putative germline variant without a clinically explicit annotation, and it affected a known proto-oncogene. The *SCMH1* and *ALDH1A2* gene variants where included as they were putatively somatic and lacked any clinically significant annotations.

While the *MAPK1* gene variant was described as unanimously deleterious by PredictSNP2, with the same expected accuracy found for the *KRAS* and *TP53* gene variants, both the *SCMH1* and *ALDH1A2* gene variants were overall predicted non-deleterious. That said, their results were not unanimous among PredictSNP2's encompassing tools. The *SCMH1* gene variant was supported as being deleterious by two of PredictSNP2's individual tools, while the *ALDHA1A2* gene variant was also predicted deleterious by one tool. Given these results, we further studied only those novel variants predicted by PredictSNP2 as unanimously deleterious.

We found that some of the variant calls (including the *HUWE1*, *G6PD*, *KIAA1161*, *CYP3A5* and *DDX52* gene variants) were incompatible with PredictSNP2. The error we received suggested that there were conflicting reference alleles between the GRCh38 references available to PredictSNP2 (which included the primary assembly and the p1, p2, and p6 patch releases) and our chosen reference. These conflicts may pertain to the EBV and human decoy sequences, previously described, which are not included in the primary assemblies.

Table 8 provides the PredictSNP2 tool's results for the eleven gene variants, across both data sets, that were predicted unanimously as deleterious. However, this table does exclude gene

variants for the *FLNA* an *MET* genes as both already possessed ClinVar annotations of uncertain significance (452140 and 411875, respectively). This suggests that these variants are already being studied more in-depth elsewhere.

For the SK data, those variants that remained were for: *multiple myeloma 1* (*MUM1*, also known as *interferon regulatory factor 4*, *IRF4*), *CAPN2*, *MAPK1*, *JAK1*, *KIF9*, *ovarian tumour deubiquitinase 7B* (*OTUD7B*), and *protein tyrosine phosphatase receptor type F* (*PTPRF)*. For the NZ data, this include variants for: *phytanoyl-coenzyme A 2-hydxoylase* (*PHYH*), *exportin 1* (*XPO1*), *plant homeodomain finger protein 23* (*PHF23*), and *shisa family member* 5 (*SHISA5*, also known as *SCOTIN*).

4.8 – More significant changes in gene expression were observed between normal and primary samples.

Gene expression was first quantified using the "featureCounts" tool available to the Subread package. Differential gene expression was then assessed using the DESeq2. More genes were found to have their expression disrupted significantly for the SK data's CRC vs. NRM comparisons. Fewer significant differences were seen when comparing either data sets' CRC vs. LM samples. This is not unexpected, given that the comparisons being made were between samples that had also undergone oncogenic transformations.

Tables 10 and 11 show the results of this differential expression analysis. However, only the SK variants (*MUM1*, *CAPN2*, *MAPK1*, *JAK1*, *KIF9*, *OTUD7*, and *PTPRF*, Table 10) showed adjusted p-values below that of the 0.05 significance threshold we set for our analysis. None of the NZ data's more novel variants (*PHF23*, *SHISA*, *XPO1*, and *PHYH*, Table 11) had adjusted p-values below this 0.05 threshold.

Interestingly for the SK data's CRC vs. NRM comparison, three of the gene affected by novel NZ variants (*SHISA5*, *XPO1*, and *PHYH*) did have their expressions significantly perturbed (supplementary material 20). Similarly, despite the NZ data's gene expression being compared against already cancerous states (CRC vs. LM), gene expression was also significantly perturbed for two genes which were affected by novel SK gene variants (*MUM1* and *KIF9*, supplementary material 21). This suggests that these genes could be targets for oncogenic disruptions beyond the scope of this study, e.g. more complex variants, epigenetic modifications, etc.

As only the novel SK variants were found to have their expressions perturbed significantly, we then compared the quantified expression values (given as tags per million or TPM) between the sample where that variant was obtained against the averages of that gene's expression for the SK data set. As all the SK variants were putatively germline, we did not suspect them of having grossly altered gene expressions for the NRM samples. However, we did consider that these could be risk alleles that somehow facilitate oncogenesis.

All but two of the samples that contained these gene variants displayed gene expression values within one standard deviation of the average of expression. The two exceptions pertained to the *CAPN2* and *JAK1* gene variants, with the *CAPN2* gene's NRM expression being just above that

of one standard deviation, whereas both the NRM and CRC samples with the *JAK1* gene variant showed expression below that of one standard deviation of the mean.

These results suggest that these SK novel variants do not seem to disrupt gene expression any more than what was seen in other cancerous samples, with possible exception for the *CAPN2* and *JAK1* gene variants. We will now consider each in more detail and speculate as to what role these variants and their genes may play in oncogenesis.

4.9 – Noteworthy novel variant calls made for the SK and NZ data.

As discussed, all the SK novel gene variants (*MUM1*, *CAPN2*, *MAPK1*, *JAK1*, *KIF9*, and *OTUD7B*) were putatively germline, i.e. they were either called in either in high confidence in NRM samples, or less stringently in NRM samples and in high confidence for cancerous samples. Unlike the SK data, the lack of normal samples for the NZ data means it was difficult to ascertain if the NZ novel variants were germline or somatic. However, as both types of variation could influence carcinogenesis in different ways, the NZ variants were considered regardless of knowledge of their germline/somatic status.

For example, germline variants could be risk alleles that function normally until subsequent cellular stress impedes this function. As a result, germline variants can lead to a predisposition towards diseases like CRC. Somatic variants, meanwhile, are a direct result of carcinogenesis taking place, with somatic variants serving some oncogenic purpose being maintained as tumorous cells proliferate. Somatic variants could serve as important biomarkers for clinical diagnoses and prognoses.

That said, the lack of normal samples also means it is difficult to understand how the expression may have changed as a tissue transitions from being normal to cancerous. This is likely why none of the NZ novel gene variants (*PHYH*, *XPO1*, *PHF23*, and *SHISA5*) showed a significant log fold change in gene expression as the comparison was being made between two already cancerous samples. We similarly saw little difference when comparing the SK data's CRC and LM samples, with more perturbed expression seen when comparing the SK data's CRC and NRM samples.

With that in mind, had NRM samples been available to the NZ data, some or all the genes affected by novel variants could have seen significant changes in expression. Additionally, we found that for the SK data, the *PHYH*, *XPO1*, and *SHISA5* genes did show perturbed expression (supplementary material 20). This was despite high confidence variants for these genes only being called in the NZ data and may support these genes as being oncogenic targets for diseases like CRC. Future work would clearly benefit from the inclusion of normal samples and will be discussed further in section 4.11.

While we have noted 11 novel variants called during our analysis (*MUM1*, *CAPN2*, *MAPK1*, *JAK1*, *KIF9*, *OTUD7B*, *PHYH*, *XPO1*, *PHF23*, and *SHISA5*), for brevity we will only discuss one variant from each data set (*MUM1* for the SK data and *SHISA5* for the NZ data). A more in-depth discussion for all novel variants can be found in supplementary material 24.

The *multiple melanoma 1* (*MUM1)* gene was originally identified as a myeloma-associated proto-oncogene activated as a result of a translocation between chromosomes 6 (p25) and 14 (q32). This gene, also known as *interferon regulatory factor 4* or *IRF4*, can both positively and negatively regulate the expression of other genes [192].

In B-cell lymphoma/leukaemia, for example, *MUM1* co-operates with the myeloid transcription factor PU.1, upregulating the *monokine induced by interferon gamma* gene (*MIG*). Also known as *chemokine C-X-C motif ligand 9* (*CXCL9*), *MIG* itself plays contradictory roles in tumour suppression and progression. Regarding CRC specifically, high expression of *MIG* has been predicted to have better overall CRC survival rates [193].

The *MUM1* gene variant found in this study was annotated by SnpEff with a "stop gained" mutation. This suggests that its protein product would be truncated and likely made non-functional. This would normally be subject to nonsense-mediated decay (NMD) whereby the aberrant mRNA is degraded before protein translation. However, our *MUM1* gene lacked any explicit NMD annotations as seen for other variants called as part of our analysis.

MUM1 gene expression was found to be higher for NRM samples than for CRC (log fold change 1.81, adjusted p-value $3.09*10^{-6}$, Table 10, Figure 10). While the CRC *MUM1* variant's gene expression did exist within one standard deviation of its CRC mean, it appeared to be one of the lower values for this deviation (Figure 10b). Additionally, we did not find that gene expression was significantly perturbed for the *MIG* gene (adjusted p-value 0.07 for *CXCL9* in supplementary material 20).

Given the lack of any NMD annotations, the early stop codon for this *MUM1* gene variant could suggest that is somehow not subject to NMD. If this is the case, the resulting truncated protein may function in normal conditions but could become further repressed under stressful cancerous conditions. This could then impede the over-expression of more tumour suppressive genes. For example, we found that *MIG* expression for the relevant sample did not appear upregulated. This may suggest our *MUM1* variant is worth investigating further.

Literature pertaining to *shisa family member 5* (*SHISA5*, also known as *SCOTIN)*, has shown some involvement in oncogenesis. One study suggested that *SHISA5* had pro-apoptotic properties and induced *TP53* expression under conditions of DNA damage or cellular stress [194]. More recently, *SHISA5* expression was found to be absent from a study using glioblastoma cell lines. This contrasted the expression seen in normal astrocytes, which regulate electrical signals in the brain. This study also found that *SHISA5* modulated both apoptosis and autophagy as a result of DNA damage independent of *TP53* [195].

The *SHISA5* gene variant was annotated by both SnpEff and PredictSNP2 as having gained a stop codon. SnpEff also predicted this as having high functional impact while PredictSNP2 described this variant as having occurred within an exonic region. The *SHISA5* variant was annotated with both a loss of function (LOF) and, unlike the *MUM1* gene variant above, also possessed an NMD annotation. We also found that, while no *SHISA5* gene variant was called in high confidence for the SK data, this gene's expression was significantly perturbed for this data set (log fold change -0.56, adjusted p-value $1.66*10^{-6}$, supplementary material 20).

If the *SHISA5* gene is involved in tumour suppressive functions like those detailed above, then this variant could clearly exacerbate CRC pathogenesis. Loss of functional *SHISA5* may no longer contribute to processes like apoptosis and terminate cancerous cells, nor induce remedial proteins like TP53 when a cell is subject to stress. This would clearly benefit cancerous cells by facilitating further somatic mutations, promoting cellular survival processes, and may be worth investigating further.

4.10 - Future Work.

Our efforts to make high confidence variant calls heavily compromised the current pipeline's sensitivity and is perhaps one of the first issues that should be considered. While we were fortunate enough to make some putative novel variant discoveries after variant filtering, other studies in the future may not be as fortunate. Future work may then benefit from developing a more sensitive pipeline, as even those variants we have highlighted may be inconsequential in better understanding CRC.

One change that we could make to improve sensitivity is to have a range of confidences regarding intersecting consensus calls. This is juxtaposed with the current pipeline that only considers variants of the "highest" confidence, i.e. called by all seven of our methods. Instead, variants called by five or six could still be considered variants of "high" confidence, especially when compared to calls made by only one tool, etc.

Additionally, we also encountered instances were highest confidence calls were made for some samples and not others. This was because, although the same variant had been called, it lacked the same level of support as the highest confidence variant calls. As discussed, these variants are likely true positives supported by the fact that the highest confidence call had already been made in one sample. Also, samples where a variant had been called less stringently were often related to the sample where the high confidence call had been made, e.g. a tumour-normal pair.

Like the known databases we have used to support a variant as being a true positive call, we could compile a VCF file containing all our highest confidence calls. This file could then be used to identify the same variants that may have been called less stringently in other samples.

As well as considering variants called less stringently in other samples, it may also be worth being suspicious of samples with an apparently inordinate number of variant calls. For example, we found that three of the four NZ novel variants discussed were all present in the CRC/LM337 samples. We then discovered that, after our filtering steps had been applied, 29 high confidence calls had been called in of these samples. Juxtaposed with this, an average of 8.96 had been called across all the SK samples (supplementary material 13). We also found that, before filtering based on a variant's QUAL value or being supported by multiple samples, both CRC337 and LM337 contained many more variants than all over samples (96 for CR337 and 90 for LM337, average 23.27).

We performed a similar analysis for the SK data. Although we did not observe any obvious patterns of hypermutation for the high impact SNPs (supplementary material 12), for moderate

impact SNPs (data not shown) we did find two related CRC and LM samples with possible signs of hypermutation. Both samples, SRR975564 and SRR975600, possessed 498 and 472 high confidence moderate variants, respectively. This was more than double that of the average number of high confidence moderate impact SNPs obtained for this data set (210.06). Interestingly, the related NRM sample (SRR975582) returned only 186 high confidence moderate impact SNPs.

This discovery made us reflect on some of the COSMIC annotations we have observed that had been removed from the COSMIC website. As discussed, these variants had been excluded as they came from samples suspected as "hypermutated" (defined initially as any samples with over 15,000 variants [187]). This made it more difficult for COSMIC to distinguish any variants from background noise. That said, when considering the total number of variant calls regardless of intersections for both data sets (supplementary materials 2 and 4), we did not see any obvious signs of hypermutation. This may warrant the development of a better methodology for identifying hypermutated samples in future work.

Other considerations for future work are which variant calling tools to use, and for which purpose. For example, the variants we ended up discussing for the SK data set were all called in their respective NRM samples, both in high confidence and less stringently. This might suggest there is some bias in our pipeline towards making germline variant calls, rather than calling somatic variants. This result is unsurprising given one of the primary focuses of this study was to make high confidence RNA-seq variant calls from the NZ data, which lacked normal samples.

The lack of normal samples for the NZ data set meant we had to make several comprises regarding which tools we could use. We could only include tools that did not require normal samples, were compatible with RNA-seq data (preferably with literature supporting the tools use in this regard) and were simple to use to and so easier to automate with the Snakemake workflow engine. For example, any study interested in making variant calls from RNA-seq data is likely to consider the HC tool, given the Broad Institute's "GATK best practises for calling variants on RNAseq". However, unlike the MT2 tool, HC is not explicitly a somatic variant calling tool.

With that in mind, seeking consensus between somatic variant callers and others could present conflicts further reducing variant calling sensitivity. For example, any call made by a non-somatic variant caller also had to appear "somatic" enough to be detected by MT2. Even some of our putative somatic variants, including those observed as somatic in a known variant database, does not mean they were also somatic in our samples. A possible solution for future work would be to develop two separate variant calling pipelines: one pipeline specific for germline calls using tools like HC, and another for somatic calls using tools like MT2.

From what we have seen so far, we would suggest using both HC and PT, coupled with both the GATK and OP pre-processing methods, as part of the germline pipelines methodology. To make RNA-seq variant calls, the data in question needs to be pre-processed and presently, these are the two pre-processing methods available. Fortuitously, both pre-processing methods also have variant calling tools to which they are specifically designed.

This would provide any future studies with a variant calling pipeline that includes an ensemble of at least three variant calling methods: GATKHC, OPHC, and OPPT. We would not consider using the FB tool at least until some filters have been formally developed for the tool. This is because the FB tool's high sensitively suggests a lack of specificity and likely would result in many false positive calls.

Developing both a germline and somatic variant calling pipelines presents us with another consideration for future work. As already discussed, the lack of normal samples for the NZ data compromised how many somatic variant callers we could use. That said, there is still merit from using an ensemble of variant callers, as we have seen given the sensitivity of tools like FB in making variant calls. Any future studies whereby using RNA-seq data to make variant calls would then benefit greatly from the inclusion of normal samples, which would have various benefits.

Having access to normal samples would allow us to use other somatic variant callers. These tools often require normal samples as input, along with a cancerous sample. Providing these tools with normal samples would serve a similar purposes as to when we would manually inspect a sample's NRM VCF file, verifying if a variant was absent from those samples. Using multiple somatic callers would also allow us to see which of these calls intersected and so provide more support for those variants being true positives in the absence of support from known variant databases.

While the NZ data did lack normal samples, the increased depth at which it was sequenced would be beneficial for future studies. For example, while the OPPT method obtained enough variants to be compared to the HC and MT2 methods for the SK data, both HC and MT2 obtained roughly double the number of variants called by OPPT for the NZ data. OPPT's performance for the SK data was interesting, given it had access to only the OP pre-processed BAM files. Meanwhile, both HC and MT2 had access to both OP and GATK BAM files. The result seen for the NZ data, therefore, was more in line with what we expected. This might also suggest that the PT tool's performance does not improve with this increase in depth and may be relevant depending on the depth of sequencing for future work.

Another critique is the fact that our current methodology is both computationally expensive, requiring a significant investment in runtime. This appeared especially true when calling variants for the more deeply sequenced NZ data when using MT2. This may be because somatic tools, like MT2, make better use of this increased depth when making variant calls. As expected, this increased depth did increase average runtime (Figures 1a and 2a). However, the increased runtime for the NZ data's OPMT2 variant calling method was substantial. Given the overwhelming number of MT2 variants that were only called after OP pre-processing (Figures 8e and 9e), this requires further investigation. The above result regarding the OPMT2 method may call into question the suitability of using pre-processing methods with somatic variant callers and may undermine any somatic variant calls we can make from RNA-seq data.

Another possibility for future work would be to make variant calls from both RNA-seq and DNA-seq data. Concordant variant calls made in both RNA-seq and DNA-seq data would

provide support for each other's calls as being true positives. Additionally, Sanger sequencing can also be used to verify that variant as a true positive, although this may be best reserved for the variants of the most interest to researchers.

Regarding our pipeline's runtime, it may be worthwhile prioritizing one germline and one somatic variant calling method, unlike the current pipeline which requires the whole ensemble to finish before obtaining data. For example, allowing the already quick OPPT method to make its variant calls and then annotated them with known variant databases, etc. would provide some preliminary data that could be analysed. Meanwhile, the rest of the ensemble of variant callers could continue to work in the background, providing support for some calls already made with OPPT. This may be especially useful in a more clinical setting where a quick turnaround of results could impact a patients' treatment and prognosis significantly.

Other limitations of this study include a focus on SNP variants, and the exclusion of more complex variants such as MNP's, indels, etc. SNPs were initially chosen because their simplicity meant that more robust filters have been designed for making their calls. While SNPs may be easier to call, complex variants are also more likely to have more significant pathological affects because of the scale to which they can alter a genome [63].

To conclude our thoughts for future work, another consideration regarding SNP variants specifically is a biological process known as "RNA editing". After a RNA molecule is generated within a cell, that cell may then make discrete post-translational edits or modifications to the molecule. Examples of such modifications include making insertions, deletions, and base substitutions, such as the common deamination of cytidine and adenosine to uridine and inosine, respectively [196]. At a surface level, this is not far removed from what a variant calling tool may interpret as a genetic variant or mutation.

Unfortunately, there appears to be no RNA editing resource akin to the known variants databases we included in this study, that could be easily implemented into our pipeline. The closest we could find to such a resource was a text file, rather than a VCF file with which we could annotate our variants, from the Rigorously Annotated Database of A-to-I RNA editing website. Developing an RNA editing VCF file like the DBSNP, COSMIC, and ClinVar resources, which were all proven to be very useful in our study, could both limit any false positive variant calls and help identify aberrant RNA editing events.

5.0 – Conclusion.

Developing an RNA-seq variant calling pipeline, without a reliance on normal samples, has proven to be an interesting academic exercise. However, the issues pertaining to RNA-seq variant calling have not been resolved in this study. Of the two RNA-seq pre-processing methods implemented in our pipeline, the Genome Analysis Toolkit/GATK best practises and the Opossum pre-processing tool, GATK appeared less specific/more sensitive in comparison to Opossum.

The combination of Opossum with the Platypus variant caller (the tool for which Opossum was specifically designed and is also incompatible with GATK pre-processing) appeared to be a remarkably efficient variant calling method, considering the speed at which it called variants. That said, this combination may not make the best use of data sequenced at deeper depths.

The GATK best practises, designed for their HaplotypeCaller, improved the number of concordant variant calls when using this tool following GATK and Opossum pre-processing. This likely limited the number of false positives we would have obtained otherwise.

Meanwhile, MuTect2 was one of the few somatic variant callers we could use as it did not require normal samples as part of its input. This was a limitation we were forced to consider when developing the pipeline. The overwhelming number of variants called following the combination of Opossum with MuTect2 suggests there is a bias which future work will need to investigate. This may also call into question if any RNA-seq pre-processing method is suitable when making somatic variant calls.

Freebayes appears to be an incredibly sensitive tool, not surprising given its reputation. Its lack of documented filters, akin to the GATK best practises or tools like Opossum, may burden researchers with false positives until this is addressed.

Of the known variant databases used in this study (the SNP known variant database/DBSNP, the Catalogue of Somatic Mutations in Cancer/COSMIC, the database of Clinical Variants/ClinVar, the Exome Aggregation Consortium/ExAC, and the Genome Aggregation Database/GNOMAD), the DBSNP, COSMIC, and ClinVar databases proved most useful for our study's aims. We also found that often the IDs annotated by those databases, and their accompanying websites, provided information made available by both ExAC and GNOMAD.

Future work would benefit from various changes to the existing pipeline. For example, we could make separate germline and somatic variant calling pipelines. We could also prioritize one variant calling method, such as the quick Opossum/Platypus combination complete with known variant annotations. This would provide researchers with some preliminary variant calls which may be useful in situations where timely variant calls could have significant consequences.

Also, we would recommend any future work to include normal samples, permitting the use of more somatic variant calling tools in any developed pipeline. We would also consider using RNA and DNA sequencing data to make variant calls. This would allow one set of results to support the other where the same variant call had been made.

Finally, we could develop a database of known RNA editing sites, akin to the known variant databases above, which could be included in our pipeline to further reduce any false positive variant calls.

6.0 – References

1. Juul, J.S., et al., *Differences in diagnostic activity in general practice and findings for individuals invited to the danish screening programme for colorectal cancer: a population-based cohort study.* Scand J Prim Health Care, 2018. **36**(3): p. 281-290.
2. Saeinasab, M., et al., *SNHG15 is a bifunctional MYC-regulated noncoding locus encoding a lncRNA that promotes cell proliferation, invasion and drug resistance in colorectal cancer by interacting with AIF*. J Exp Clin Cancer Res, 2019. **38**(1): p. 172.
3. Yang, Y., et al., *Whole transcriptome sequencing identifies crucial genes associated with colon cancer and elucidation of their possible mechanisms of action.* OncoTargets and therapy, 2019. **12**: p. 2737-2747.
4. Bao, Y., et al., *Long noncoding RNA BFAL1 mediates enterotoxigenic Bacteroides fragilis-related carcinogenesis in colorectal cancer via the RHEB/mTOR pathway.* Cell Death & Disease, 2019. **10**(9): p. 675.
5. De Almeida, C.V., et al., *Role of diet and gut microbiota on colorectal cancer immunomodulation.* World journal of gastroenterology, 2019. **25**(2): p. 151-162.
6. Ai, D., et al., *Using Decision Tree Aggregation with Random Forest Model to Identify Gut Microbes Associated with Colorectal Cancer.* Genes, 2019. **10**(2): p. 112.
7. Ma, H., et al., *Correlation between microbes and colorectal cancer: tumor apoptosis is induced by sitosterols through promoting gut microbiota to produce short-chain fatty acids.* Apoptosis, 2019. **24**(1-2): p. 168-183.
8. Aghabozorgi, A.S., et al., *Role of adenomatous polyposis coli (APC) gene mutations in the pathogenesis of colorectal cancer; current status and perspectives.* Biochimie, 2019. **157**: p. 64-71.
9. Thomas, A.M., et al., *Metagenomic analysis of colorectal cancer datasets identifies cross-cohort microbial diagnostic signatures and a link with choline degradation.* Nat Med, 2019. **25**(4): p. 667-678.
10. Kolodziejczyk, A.A., D. Zheng, and E. Elinav, *Diet–microbiota interactions and personalized nutrition.* Nature Reviews Microbiology, 2019. **17**(12): p. 742-753.
11. Tarrado-Castellarnau, M., et al., *Glyceraldehyde-3-phosphate dehydrogenase is overexpressed in colorectal cancer onset.* Translational Medicine Communications, 2017. **2**(1): p. 6.
12. Hannigan, G.D., et al., *Diagnostic Potential and Interactive Dynamics of the Colorectal Cancer Virome.* mBio, 2018. **9**(6): p. e02248-18.
13. McLeod, M., et al., *Colorectal Cancer Screening: How Health Gains and Cost-Effectiveness Vary by Ethnic Group, the Impact on Health Inequalities, and the Optimal Age Range to Screen.* Cancer Epidemiol Biomarkers Prev, 2017. **26**(9): p. 1391-1400.
14. Hurtado, C.G., et al., *Roles for Interleukin 17 and Adaptive Immunity in Pathogenesis of Colorectal Cancer.* Gastroenterology, 2018. **155**(6): p. 1706-1715.
15. Chen, E., et al., *MiR-92a promotes tumorigenesis of colorectal cancer, a transcriptomic and functional based study.* Biomed Pharmacother, 2018. **106**: p. 1370-1377.
16. Feng, Q., et al., *Gut microbiome development along the colorectal adenoma–carcinoma sequence.* Nature Communications, 2015. **6**: p. 6528.
17. Ferlay, J., et al., *Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012.* Int J Cancer, 2015. **136**(5): p. E359-86.
18. Wylie, N., et al., *From scratch: developing a hepatic resection service for metastatic colorectal cancer.* ANZ J Surg, 2018. **88**(5): p. E377-e381.

19. Teng, H., et al., *Identification of recurrent and novel mutations by whole-genome sequencing of colorectal tumors from the Han population in Shanghai, eastern China.* Molecular medicine reports, 2018. **18**(6): p. 5361-5370.
20. Heiken, J.P., *Screening for colon cancer.* Cancer Imaging, 2006. **6**: p. S13-21.
21. Sze, M.A. and P.D. Schloss, *Leveraging Existing 16S rRNA Gene Surveys To Identify Reproducible Biomarkers in Individuals with Colorectal Tumors.* mBio, 2018. **9**(3): p. e00630-18.
22. Zackular, J.P., et al., *The human gut microbiome as a screening tool for colorectal cancer.* Cancer prevention research (Philadelphia, Pa.), 2014. **7**(11): p. 1112-1121.
23. Bader, J.E., et al., *Macrophage depletion using clodronate liposomes decreases tumorigenesis and alters gut microbiota in the AOM/DSS mouse model of colon cancer.* American journal of physiology. Gastrointestinal and liver physiology, 2018. **314**(1): p. G22-G31.
24. Gómez-Moreno, R., et al., *The Presence of Gut Microbial Genes Encoding Bacterial Genotoxins or Pro-Inflammatory Factors in Stool Samples from Individuals with Colorectal Neoplasia.* Diseases (Basel, Switzerland), 2019. **7**(1): p. 16.
25. Handley, S.A. and S. Devkota, *Going Viral: a Novel Role for Bacteriophage in Colorectal Cancer.* mBio, 2019. **10**(1): p. e02626-18.
26. Faïs, T., et al., *Colibactin: More Than a New Bacterial Toxin.* Toxins, 2018. **10**(4): p. 151.
27. Weinhold, N., et al., *Genome-wide analysis of noncoding regulatory mutations in cancer.* Nat Genet, 2014. **46**(11): p. 1160-5.
28. Chen, Y. and W. Song, *Wnt/catenin β1/microRNA 183 predicts recurrence and prognosis of patients with colorectal cancer.* Oncology letters, 2018. **15**(4): p. 4451-4456.
29. Li, J., et al., *The oncogenic role of Wnt10a in colorectal cancer through activation of canonical Wnt/beta-catenin signaling.* Oncol Lett, 2019. **17**(4): p. 3657-3664.
30. Xiao, C.H., et al., *Long non-coding RNA TUG1 promotes the proliferation of colorectal cancer cells through regulating Wnt/β-catenin pathway.* Oncology letters, 2018. **16**(4): p. 5317-5324.
31. Lee, H.K., et al., *Ubiquitylation and degradation of adenomatous polyposis coli by MKRN1 enhances Wnt/beta-catenin signaling.* Oncogene, 2018. **37**(31): p. 4273-4286.
32. Ding, M. and X. Wang, *Antagonism between Hedgehog and Wnt signaling pathways regulates tumorigenicity.* Oncol Lett, 2017. **14**(6): p. 6327-6333.
33. Zhan, T., N. Rindtorff, and M. Boutros, *Wnt signaling in cancer.* Oncogene, 2017. **36**(11): p. 1461-1473.
34. D'Elia, G., et al., *APC and MUTYH Analysis in FAP Patients: A Novel Mutation in APC Gene and Genotype-Phenotype Correlation.* Genes (Basel), 2018. **9**(7).
35. Leoz, M.L., et al., *The genetic basis of familial adenomatous polyposis and its implications for clinical practice and risk management.* The application of clinical genetics, 2015. **8**: p. 95-107.
36. Møller, P., et al., *Cancer risk and survival in path_MMR carriers by gene and gender up to 75 years of age: a report from the Prospective Lynch Syndrome Database.* Gut, 2018. **67**(7): p. 1306-1316.
37. Lee, S.E., et al., *Pyloric gland adenoma in Lynch syndrome.* The American journal of surgical pathology, 2014. **38**(6): p. 784-792.

38. Le, D.T., et al., *PD-1 Blockade in Tumors with Mismatch-Repair Deficiency.* The New England journal of medicine, 2015. **372**(26): p. 2509-2520.
39. Lee, Y.K., et al., *The Protective Role of Bacteroides fragilis in a Murine Model of Colitis-Associated Colorectal Cancer.* mSphere, 2018. **3**(6).
40. Poirion, O., et al., *Using single nucleotide variations in single-cell RNA-seq to identify subpopulations and genotype-phenotype linkage.* Nature Communications, 2018. **9**(1): p. 4892.
41. Yang, S.Y., M.S. Cho, and N.K. Kim, *Difference between right-sided and left-sided colorectal cancers: from embryology to molecular subtype.* Expert Rev Anticancer Ther, 2018. **18**(4): p. 351-358.
42. Sadanandam, A., et al., *A colorectal cancer classification system that associates cellular phenotype and responses to therapy.* Nat Med, 2013. **19**(5): p. 619-25.
43. De Sousa, E.M.F., et al., *Poor-prognosis colon cancer is defined by a molecularly distinct subtype and develops from serrated precursor lesions.* Nat Med, 2013. **19**(5): p. 614-8.
44. Hutchinson, L., *Turning up trumps for new CRC subtypes.* Nature Reviews Clinical Oncology, 2013. **10**: p. 303.
45. Guinney, J., et al., *The consensus molecular subtypes of colorectal cancer.* Nat Med, 2015. **21**(11): p. 1350-6.
46. Simoneaux, R., *The Four Colorectal Cancer Consensus Molecular Subtypes*. Vol. 40. 2018. 10-11.
47. Zhao, D., et al., *A reliable method for colorectal cancer prediction based on feature selection and support vector machine.* Med Biol Eng Comput, 2019. **57**(4): p. 901-912.
48. Stintzing, S., et al., *Consensus molecular subgroups (CMS) of colorectal cancer (CRC) and first-line efficacy of FOLFIRI plus cetuximab or bevacizumab in the FIRE3 (AIO KRK-0306) trial.* Ann Oncol, 2019. **30**(11): p. 1796-1803.
49. Aguirre-Portolés, C., L.P. Fernández, and A. Ramírez de Molina, *Precision Nutrition for Targeting Lipid Metabolism in Colorectal Cancer.* Nutrients, 2017. **9**(10): p. 1076.
50. Dai, L., et al., *Temporal expression and functional analysis of long non-coding RNAs in colorectal cancer initiation.* Journal of cellular and molecular medicine, 2019. **23**(6): p. 4127-4138.
51. Dai, Z., et al., *Multi-cohort analysis of colorectal cancer metagenome identified altered bacteria across populations and universal bacterial markers.* Microbiome, 2018. **6**(1): p. 70-70.
52. Schmidt, T.S., et al., *Extensive transmission of microbes along the gastrointestinal tract.* eLife, 2019. **8**: p. e42693.
53. Bundgaard-Nielsen, C., et al., *The presence of bacteria varies between colorectal adenocarcinomas, precursor lesions and non-malignant tissue.* BMC cancer, 2019. **19**(1): p. 399-399.
54. Shah, M.S., et al., *Re-purposing 16S rRNA gene sequence data from within case paired tumor biopsy and tumor-adjacent biopsy or fecal samples to identify microbial markers for colorectal cancer.* PloS one, 2018. **13**(11): p. e0207002-e0207002.
55. Burns, M.B., et al., *Colorectal cancer mutational profiles correlate with defined microbial communities in the tumor microenvironment.* PLOS Genetics, 2018. **14**(6): p. e1007376.

56. Miyamoto, D.T., et al., *RNA-Seq of single prostate CTCs implicates noncanonical Wnt signaling in antiandrogen resistance.* Science, 2015. **349**(6254): p. 1351-1356.
57. Jacouton, E., et al., *Elucidating the Immune-Related Mechanisms by Which Probiotic Strain Lactobacillus casei BL23 Displays Anti-tumoral Properties.* Frontiers in microbiology, 2019. **9**: p. 3281-3281.
58. Burns, M.B. and R. Blekhman, *Integrating tumor genomics into studies of the microbiome in colorectal cancer.* Gut Microbes, 2019. **10**(4): p. 547-552.
59. Purcell, R.V., et al., *Distinct gut microbiome patterns associate with consensus molecular subtypes of colorectal cancer.* Scientific reports, 2017. **7**(1): p. 11590-11590.
60. Giannoulatou, E., et al., *Verification and validation of bioinformatics software without a gold standard: a case study of BWA and Bowtie.* BMC bioinformatics, 2014. **15 Suppl 16**(Suppl 16): p. S15-S15.
61. Cummings, B.B., et al., *Improving genetic diagnosis in Mendelian disease with transcriptome sequencing.* Science Translational Medicine, 2017. **9**(386): p. eaal5209.
62. Wrzeszczynski, K.O., et al., *Analytical Validation of Clinical Whole-Genome and Transcriptome Sequencing of Patient-Derived Tumors for Reporting Targetable Variants in Cancer.* J Mol Diagn, 2018. **20**(6): p. 822-835.
63. Lappalainen, T., et al., *Transcriptome and genome sequencing uncovers functional variation in humans.* Nature, 2013. **501**: p. 506.
64. Tang, X., et al., *The eSNV-detect: a computational system to identify expressed single nucleotide variants from transcriptome sequencing data.* Nucleic acids research, 2014. **42**(22): p. e172-e172.
65. Piskol, R., G. Ramaswami, and J.B. Li, *Reliable identification of genomic variants from RNA-seq data.* American journal of human genetics, 2013. **93**(4): p. 641-651.
66. Zhao, Y., et al., *A high-throughput SNP discovery strategy for RNA-seq data.* BMC Genomics, 2019. **20**(1): p. 160.
67. Lai, Z., et al., *VarDict: a novel and versatile variant caller for next-generation sequencing in cancer research.* Nucleic acids research, 2016. **44**(11): p. e108-e108.
68. Sandmann, S., et al., *Evaluating Variant Calling Tools for Non-Matched Next-Generation Sequencing Data.* Scientific Reports, 2017. **7**: p. 43169.
69. Tukiainen, T., et al., *Landscape of X chromosome inactivation across human tissues.* Nature, 2017. **550**: p. 244.
70. Han, A., et al., *Butyrate decreases its own oxidation in colorectal cancer cells through inhibition of histone deacetylases.* Oncotarget, 2018. **9**(43): p. 27280-27292.
71. Ryu, T.Y., et al., *Downregulation of PRMT1, a histone arginine methyltransferase, by sodium propionate induces cell apoptosis in colon cancer.* Oncology reports, 2019. **41**(3): p. 1691-1699.
72. Kim, J.C., et al., *Complex Behavior of ALDH1A1 and IGFBP1 in Liver Metastasis from a Colorectal Cancer.* PLoS One, 2016. **11**(5): p. e0155160.
73. Currey, N., et al., *Mouse Model of Mutated in Colorectal Cancer Gene Deletion Reveals Novel Pathways in Inflammation and Cancer.* Cellular and molecular gastroenterology and hepatology, 2019. **7**(4): p. 819-839.
74. Conesa, A., et al., *A survey of best practices for RNA-seq data analysis.* Genome Biology, 2016. **17**(1): p. 13.
75. Jakhesara, S.J., et al., *Identification and quantification of novel RNA isoforms in horn cancer of Bos indicus by comprehensive RNA-Seq.* 3 Biotech, 2016. **6**(2): p. 259.

76. Hu, Y., et al., *PennDiff: detecting differential alternative splicing and transcription by RNA sequencing.* Bioinformatics (Oxford, England), 2018. **34**(14): p. 2384-2391.
77. Trapnell, C., et al., *Differential analysis of gene regulation at transcript resolution with RNA-seq.* Nat Biotechnol, 2013. **31**(1): p. 46-53.
78. Jakhesara, S.J., et al., *Identification and quantification of novel RNA isoforms in horn cancer of Bos indicus by comprehensive RNA-Seq.* 3 Biotech, 2016. **6**(2): p. 259-259.
79. Zhang, X. and S.R. Ellingson, *Computationally Characterizing Genomic Pipelines Using High-confident Call Sets.* Procedia Computer Science, 2016. **80**: p. 1023-1032.
80. Park, S.T. and J. Kim, *Trends in Next-Generation Sequencing and a New Era for Whole Genome Sequencing.* Int Neurourol J, 2016. **20**(Suppl 2): p. S76-83.
81. Heydari, M., et al., *BrownieAligner: accurate alignment of Illumina sequencing data to de Bruijn graphs.* BMC Bioinformatics, 2018. **19**(1): p. 311.
82. Zhu, P., et al., *OTG-snpcaller: an optimized pipeline based on TMAP and GATK for SNP calling from ion torrent data.* PLoS One, 2014. **9**(5): p. e97507.
83. Giordano, F., et al., *De novo yeast genome assemblies from MinION, PacBio and MiSeq platforms.* Scientific Reports, 2017. **7**(1): p. 3935.
84. Love, M.I., J.B. Hogenesch, and R.A. Irizarry, *Modeling of RNA-seq fragment sequence bias reduces systematic errors in transcript abundance estimation.* Nat Biotechnol, 2016. **34**(12): p. 1287-1291.
85. Head, S.R., et al., *Library construction for next-generation sequencing: overviews and challenges.* Biotechniques, 2014. **56**(2): p. 61-4, 66, 68, passim.
86. Xu, C., *A review of somatic single nucleotide variant calling algorithms for next-generation sequencing data.* Computational and structural biotechnology journal, 2018. **16**: p. 15-24.
87. Genomes Project, C., et al., *A global reference for human genetic variation.* Nature, 2015. **526**(7571): p. 68-74.
88. MacArthur, D.G., et al., *Guidelines for investigating causality of sequence variants in human disease.* Nature, 2014. **508**(7497): p. 469-476.
89. Gonsalves, S.G., et al., *Using exome data to identify malignant hyperthermia susceptibility mutations.* Anesthesiology, 2013. **119**(5): p. 1043-1053.
90. Gonorazky, H.D., et al., *Expanding the Boundaries of RNA Sequencing as a Diagnostic Tool for Rare Mendelian Disease.* Am J Hum Genet, 2019. **104**(3): p. 466-483.
91. Esfandi, F., et al., *Long noncoding RNAs expression in gastric cancer.* J Cell Biochem, 2019. **120**(8): p. 13802-13809.
92. Wei, J.W., et al., *Non-coding RNAs as regulators in epigenetics (Review).* Oncol Rep, 2017. **37**(1): p. 3-9.
93. Oczkowicz, M., et al., *Variant calling from RNA-seq data of the brain transcriptome of pigs and its application for allele-specific expression and imprinting analysis.* Gene, 2018. **641**: p. 367-375.
94. Vu, T.N., et al., *A fast detection of fusion genes from paired-end RNA-seq data.* BMC Genomics, 2018. **19**(1): p. 786.
95. Lowe, R., et al., *Transcriptomics technologies.* PLOS Computational Biology, 2017. **13**(5): p. e1005457.
96. Bohannan, Z.S., *Calling Variants in the Clinic: Informed Variant Calling Decisions Based on Biological, Clinical, and Laboratory Variables.* Computational and Structural Biotechnology Journal, 2019. **v. 17**: p. pp. 561-569-2019 v.17.

97. Prodduturi, N., et al., *Indel sensitive and comprehensive variant/mutation detection from RNA sequencing data for precision medicine.* BMC Medical Genomics, 2018. **11**(3): p. 67.
98. Mohammad, T.A., et al., *CeL-ID: cell line identification using RNA-seq data.* BMC Genomics, 2019. **20**(Suppl 1): p. 81.
99. Coudray, A., et al., *Detection and benchmarking of somatic mutations in cancer genomes using RNA-seq data.* PeerJ, 2018. **6**: p. e5362-e5362.
100. Adetunji, M., et al., *VARIANT ANALYSIS PIPELINE FOR ACCURATE DETECTION OF GENOMIC VARIANTS FROM TRANSCRIPTOME SEQUENCING DATA*. 2019.
101. Hong, J.H., Y.H. Ko, and K. Kang, *RNA variant identification discrepancy among splice-aware alignment algorithms.* PloS one, 2018. **13**(8): p. e0201822-e0201822.
102. Chepelev, I., et al., *Detection of single nucleotide variations in expressed exons of the human genome using RNA-Seq.* Nucleic Acids Research, 2009. **37**(16): p. e106-e106.
103. Yu, L.C., *Microbiota dysbiosis and barrier dysfunction in inflammatory bowel disease and colorectal cancers: exploring a common ground hypothesis.* J Biomed Sci, 2018. **25**(1): p. 79.
104. Kelly, D., L. Yang, and Z. Pei, *Gut Microbiota, Fusobacteria, and Colorectal Cancer.* Diseases (Basel, Switzerland), 2018. **6**(4): p. 109.
105. Coleman, O.I., et al., *Activated ATF6 Induces Intestinal Dysbiosis and Innate Immune Response to Promote Colorectal Tumorigenesis.* Gastroenterology, 2018. **155**(5): p. 1539-1552.e12.
106. Khan, S.A., et al., *Colorectal cancer in the very young: a comparative study of tumor markers, pathology and survival in early onset and adult onset patients.* J Pediatr Surg, 2016. **51**(11): p. 1812-1817.
107. Nakano, K., et al., *Clinicopathologic and Molecular Characteristics of Synchronous Colorectal Carcinoma With Mismatch Repair Deficiency.* Am J Surg Pathol, 2018. **42**(2): p. 172-182.
108. Li, L., et al., *KCTD12 Regulates Colorectal Cancer Cell Stemness through the ERK Pathway.* Scientific Reports, 2016. **6**(1): p. 20460.
109. Rimmer, A., et al., *Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications.* Nature Genetics, 2014. **46**: p. 912.
110. Giannakis, M., et al., *RNF43 is frequently mutated in colorectal and endometrial cancers.* Nat Genet, 2014. **46**(12): p. 1264-6.
111. Jain, M., et al., *Linear assembly of a human centromere on the Y chromosome.* Nature biotechnology, 2018. **36**(4): p. 321-323.
112. Genovese, G., et al., *Mapping the human reference genome's missing sequence by three-way admixture in Latino genomes.* Am J Hum Genet, 2013. **93**(3): p. 411-21.
113. Brouard, J.-S., et al., *The GATK joint genotyping workflow is appropriate for calling variants in RNA-seq experiments.* Journal of animal science and biotechnology, 2019. **10**: p. 44-44.
114. O'Rawe, J., et al., *Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing.* Genome Medicine, 2013. **5**(3): p. 28.
115. Linderman, M.D., et al., *Analytical validation of whole exome and whole genome sequencing for clinical applications.* BMC Med Genomics, 2014. **7**: p. 20.
116. Tan, A., G.R. Abecasis, and H.M. Kang, *Unified representation of genetic variants.* Bioinformatics, 2015. **31**(13): p. 2202-2204.

117. McCarthy, D.J., et al., *Choice of transcripts and software has a large effect on variant annotation.* Genome Med, 2014. **6**(3): p. 26.
118. Wang, C., et al., *RVboost: RNA-seq variants prioritization using a boosting method.* Bioinformatics, 2014. **30**(23): p. 3414-6.
119. Shao, M. and C. Kingsford, *Accurate assembly of transcripts through phase-preserving graph decomposition.* Nature Biotechnology, 2017. **35**: p. 1167.
120. Mose, L.E., C.M. Perou, and J.S. Parker, *Improved indel detection in DNA and RNA via realignment with ABRA2.* Bioinformatics, 2019. **35**(17): p. 2966-2973.
121. McKenna, A., et al., *The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data.* Genome research, 2010. **20**(9): p. 1297-1303.
122. DePristo, M.A., et al., *A framework for variation discovery and genotyping using next-generation DNA sequencing data.* Nature genetics, 2011. **43**(5): p. 491-498.
123. Team, T.G.D. *RNAseq short variant discovery (SNPs + Indels)*. [Best Practices Workflows] 2018 [cited 2019 10/09]; Available from: https://software.broadinstitute.org/gatk/best-practices/workflow?id=11164.
124. Lek, M., et al., *Analysis of protein-coding genetic variation in 60,706 humans.* Nature, 2016. **536**: p. 285.
125. Team, T.G.D. *(hotwo) Apply hard filters to a call set*. [Tutorial] 2013 27/11/2018 10/09/2019]; Available from: https://software.broadinstitute.org/gatk/documentation/article.php?id=2806.
126. Eberle, M.A., et al., *A reference data set of 5.4 million phased human variants validated by genetic inheritance from sequencing a three-generation 17-member pedigree.* Genome research, 2017. **27**(1): p. 157-164.
127. Mu, W., et al., *Sanger Confirmation Is Required to Achieve Optimal Sensitivity and Specificity in Next-Generation Sequencing Panel Testing.* J Mol Diagn, 2016. **18**(6): p. 923-932.
128. Baudhuin, L.M., et al., *Confirming Variants in Next-Generation Sequencing Panel Testing by Sanger Sequencing.* J Mol Diagn, 2015. **17**(4): p. 456-61.
129. Beck, T.F., et al., *Systematic Evaluation of Sanger Validation of Next-Generation Sequencing Variants.* Clinical chemistry, 2016. **62**(4): p. 647-654.
130. Garrison, E. and G. Marth *Haplotype-based variant detection from short-read sequencing*. arXiv e-prints, 2012.
131. Chiara, M., et al., *CoVaCS: a consensus variant calling system.* BMC genomics, 2018. **19**(1): p. 120-120.
132. Trubetskoy, V., et al., *Consensus Genotyper for Exome Sequencing (CGES): improving the quality of exome variant genotypes.* Bioinformatics, 2015. **31**(2): p. 187-93.
133. Rogier, O., et al., *Accuracy of RNAseq based SNP discovery and genotyping in Populusnigra.* BMC genomics, 2018. **19**(1): p. 909-909.
134. Narzisi, G., et al., *Genome-wide somatic variant calling using localized colored de Bruijn graphs.* Communications biology, 2018. **1**: p. 20-20.
135. Dezan, M.R., et al., *SMIM1 intron 2 gene variations leading to variability in Vel antigen expression among Brazilian blood donors.* Blood Cells Mol Dis, 2019. **77**: p. 23-28.
136. Taylor-Weiner, A., et al., *DeTiN: overcoming tumor-in-normal contamination.* Nat Methods, 2018. **15**(7): p. 531-534.

137. Ross, M.G., et al., *Characterizing and measuring bias in sequence data.* Genome Biol, 2013. **14**(5): p. R51.
138. Fu, Y., et al., *BACOM2: a Java tool for detecting normal cell contamination of copy number in heterogeneous tumor*. 2015.
139. Narasimhan, V.M., et al., *Health and population effects of rare gene knockouts in adult humans with related parents.* Science, 2016. **352**(6284): p. 474-7.
140. Rivas, M.A., et al., *Effect of predicted protein-truncating genetic variants on the human transcriptome.* Science, 2015. **348**(6235): p. 666-669.
141. Johnston, J.J., et al., *Individualized iterative phenotyping for genome-wide analysis of loss-of-function mutations.* American journal of human genetics, 2015. **96**(6): p. 913-925.
142. Zheng-Bradley, X., et al., *Alignment of 1000 Genomes Project reads to reference assembly GRCh38.* Gigascience, 2017. **6**(7): p. 1-8.
143. Ananthakrishnan, A., V. Gogineni, and K. Saeian, *Epidemiology of primary and secondary liver cancers.* Semin Intervent Radiol, 2006. **23**(1): p. 47-63.
144. Aronesty, E., *ea-utils: Command-line tools for processing biological sequencing data*. 2011.
145. Cox, M.P., D.A. Peterson, and P.J. Biggs, *SolexaQA: At-a-glance quality assessment of Illumina second-generation sequencing data.* BMC Bioinformatics, 2010. **11**: p. 485.
146. Kim, S.K., et al., *A nineteen gene-based risk score classifier predicts prognosis of colorectal cancer patients.* Mol Oncol, 2014. **8**(8): p. 1653-66.
147. Köster, J. and S. Rahmann, *Snakemake—a scalable bioinformatics workflow engine.* Bioinformatics, 2012. **28**(19): p. 2520-2522.
148. Abdouh, M., et al., *Colorectal cancer-derived extracellular vesicles induce transformation of fibroblasts into colon carcinoma cells.* Journal of experimental & clinical cancer research : CR, 2019. **38**(1): p. 257-257.
149. Chen, J., et al., *Single-cell SNP analyses and interpretations based on RNA-Seq data for colon cancer research.* Scientific Reports, 2016. **6**: p. 34420.
150. Meynert, A.M., et al., *Quantifying single nucleotide variant detection sensitivity in exome sequencing.* BMC Bioinformatics, 2013. **14**(1): p. 195.
151. National Center for Biotechnology Information. *README_analysis_sets.txt*. 2019 [cited 2019; Available from: ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/001/405/GCA_000001405.15_GRCh38/seqs_for_alignment_pipelines.ucsc_ids/.
152. Blue Collar Bioinformatics. *Validated variant calling with human genome build 38*. 2015 [cited 2019; Available from: https://libanswers.liverpool.ac.uk/faq/49511.
153. Guo, Y., et al., *Improvements and impacts of GRCh38 human reference on high throughput sequencing data analysis.* Genomics, 2017. **109**(2): p. 83-90.
154. Anderson, S., et al., *Sequence and organization of the human mitochondrial genome.* Nature, 1981. **290**(5806): p. 457-65.
155. Andrews, R.M., et al., *Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA.* Nat Genet, 1999. **23**(2): p. 147.
156. Li, H. *Which human reference genome to use?* 2017 [cited 2019; Available from: https://lh3.github.io/2017/11/13/which-human-reference-genome-to-use.
157. Broad Institute. *Picard Toolkit*. 2019 [cited 2019; Available from: http://broadinstitute.github.io/picard/.

158. Sherry, S.T., et al., *dbSNP: the NCBI database of genetic variation.* Nucleic Acids Research, 2001. **29**(1): p. 308-311.
159. Li, H., *Tabix: fast retrieval of sequence features from generic TAB-delimited files.* Bioinformatics, 2011. **27**(5): p. 718-9.
160. McKenna, A., et al., *The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data.* Genome Res, 2010. **20**(9): p. 1297-303.
161. Li, H., et al., *The Sequence Alignment/Map format and SAMtools.* Bioinformatics, 2009. **25**(16): p. 2078-9.
162. Engstrom, P.G., et al., *Systematic evaluation of spliced alignment programs for RNA-seq data.* Nat Methods, 2013. **10**(12): p. 1185-91.
163. Sahraeian, S.M.E., et al., *Gaining comprehensive biological insight into the transcriptome by performing a broad-spectrum RNA-seq analysis.* Nature Communications, 2017. **8**(1): p. 59.
164. Broad Institute. *Best Practise Workflows*. 2020 [cited 2019; Available from: https://software.broadinstitute.org/gatk/best-practices/.
165. National Center for Biotechnology Information. *GCA_000001405.15_GRCh38_full_analysis_set.refseq_annottion.gtf.gz*. 2019 [cited 2019; Available from: ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/001/405/GCA_000001405.15_GRCh38/seqs_for_alignment_pipelines.ucsc_ids/.
166. Dobin, A., et al., *STAR: ultrafast universal RNA-seq aligner.* Bioinformatics, 2013. **29**(1): p. 15-21.
167. Oikkonen, L. and S. Lise, *Making the most of RNA-seq: Pre-processing sequencing data with Opossum for reliable SNP variant detection.* Wellcome Open Res, 2017. **2**: p. 6.
168. Tarasov, A., et al., *Sambamba: fast processing of NGS alignment formats.* Bioinformatics, 2015. **31**(12): p. 2032-4.
169. Broad Institute. *Workflows for processing RNA data for germline short variant discovery with GATK (v3+v4) and related tools*. 2018 [cited 2019; Available from: https://github.com/gatk-workflows/gatk3-4-rnaseq-germline-snps-indels/blob/master/rna-germline-variant-calling.wdl.
170. Cibulskis, K., et al., *Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples.* Nat Biotechnol, 2013. **31**(3): p. 213-9.
171. Karczewski, K.J., et al., *The mutational constraint spectrum quantified from variation in 141,456 humans.* bioRxiv, 2020: p. 531210.
172. Danecek, P., et al., *The variant call format and VCFtools.* Bioinformatics (Oxford, England), 2011. **27**(15): p. 2156-2158.
173. Salatino, S. and V. Ramraj, *BrowseVCF: a web-based application and workflow to quickly prioritize disease-causative variants in VCF files.* Briefings in Bioinformatics, 2016. **18**(5): p. 774-779.
174. Cingolani, P., et al., *Using Drosophila melanogaster as a Model for Genotoxic Chemical Mutational Studies with a New Program, SnpSift.* Front Genet, 2012. **3**: p. 35.
175. Cingolani, P., et al., *A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3.* Fly (Austin), 2012. **6**(2): p. 80-92.
176. Landrum, M.J., et al., *ClinVar: improving access to variant interpretations and supporting evidence.* Nucleic acids research, 2018. **46**(D1): p. D1062-D1067.

177. Kobayashi, Y., et al., *Pathogenic variant burden in the ExAC database: an empirical approach to evaluating population data for clinical variant interpretation.* Genome medicine, 2017. **9**(1): p. 13-13.
178. Tate, J.G., et al., *COSMIC: the Catalogue Of Somatic Mutations In Cancer.* Nucleic Acids Research, 2018. **47**(D1): p. D941-D947.
179. Ho, H., et al., *NUP98-PHF23 fusion is recurrent in acute myeloid leukemia and shares gene expression signature of leukemic stem cells.* Leuk Res, 2016. **45**: p. 1-7.
180. Chen, E.Y., et al., *Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool.* BMC Bioinformatics, 2013. **14**(1): p. 128.
181. Kuleshov, M.V., et al., *Enrichr: a comprehensive gene set enrichment analysis web server 2016 update.* Nucleic Acids Res, 2016. **44**(W1): p. W90-7.
182. Bendl, J., et al., *PredictSNP2: A Unified Platform for Accurately Evaluating SNP Effects by Exploiting the Different Characteristics of Variants in Distinct Genomic Regions.* PLoS Comput Biol, 2016. **12**(5): p. e1004962.
183. Liao, Y., G. Smyth, and W. Shi, *FeatureCounts: An efficient general purpose program for assigning sequence reads to genomic features.* Bioinformatics (Oxford, England), 2013. **30**.
184. Liao, Y., G.K. Smyth, and W. Shi, *The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote.* Nucleic Acids Research, 2013. **41**(10): p. e108-e108.
185. Love, M.I., W. Huber, and S. Anders, *Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.* Genome biology, 2014. **15**(12): p. 550-550.
186. Schmeier, S. *Differential Gene Expressiom from BAM*. 2020 [cited 2020 2020]; Computation methodology provided for reproducibility]. Available from: https://gitlab.com/schmeierlab/tom/dgea-from-bam.
187. Catalogue of Somatic Mutations in Cancer, *Cancer Genome Annotation.* 2020.
188. Kim, D., et al., *Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype.* Nature Biotechnology, 2019. **37**(8): p. 907-915.
189. Li, H. and R. Durbin, *Fast and accurate short read alignment with Burrows-Wheeler transform.* Bioinformatics, 2009. **25**(14): p. 1754-60.
190. NCBI. *New SNP Attributes*. Available from: https://www.ncbi.nlm.nih.gov/projects/SNP/docs/rs_attributes.html#gmaf.
191. Kido, T., et al., *Are minor alleles more likely to be risk alleles?* BMC Medical Genomics, 2018. **11**(1): p. 3.
192. Uranishi, M., et al., *Multiple myeloma oncogene 1 (MUM1)/interferon regulatory factor 4 (IRF4) upregulates monokine induced by interferon-γ (MIG) gene expression in B-cell malignancy.* Leukemia, 2005. **19**(8): p. 1471-1478.
193. Ding, Q., et al., *CXCL9: evidence and contradictions for its role in tumor progression.* Cancer medicine, 2016. **5**(11): p. 3246-3259.
194. Bourdon, J.C., et al., *Scotin, a novel p53-inducible proapoptotic protein located in the ER and the nuclear membrane.* J Cell Biol, 2002. **158**(2): p. 235-46.
195. Qiao, B., K. Oneill, and N. Syed, *P69: SCOTIN EXPRESSION AND SURVIVAL IN GBM.* Neuro-Oncology, 2014. **16**(Suppl 6): p. vi11-vi12.
196. Peng, Z., et al., *Comprehensive analysis of RNA-Seq data reveals extensive RNA editing in a human transcriptome.* Nature Biotechnology, 2012. **30**(3): p. 253-260.