Development of computational tools for variant calling in single-cell RNAseq

Kinga Anna Zielińska

St. Peter's College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Michaelmas 2022

Abstract

Single-cell sequencing technologies have unsurprisingly become a favourable choice for studying key biological questions about cell heterogeneity, rare cell types or lineages. It is only cell-level resolution that allows for an accurate analysis of internal cell processes such as mutagenesis. Eventually, single-cell RNAseq could provide an explanation of mechanisms that lead to the ultimate transformation of healthy tissues into cancerous lesions. One of the main interests of my lab is Barrett's oesophagus. It is a highly clonal disease and a likely cancer precursor. We decided to take advantage of the single-cell RNAseq technology in order to attempt to identify the tissue of origin of the disease which, despite years of research, still remains unknown. However, the range of methods for identification of mutations in single-cells is very limited. In order to address that, we developed our own single-cell RNAseq variant caller. We validated it on a publicly available breast cancer dataset by achieving a reasonable intersection of our results with the output of commonly used bulk tools. Furthermore, we showed that our caller was capable of identifying expected data characteristics such as known breast cancer signatures and mutations in breast cancer genes. We then applied our method to the Barrett's dataset to investigate connections of Barrett's with surrounding tissues. Contrary to the previous transcriptomic analysis conducted on the same dataset and indicating a Barrett's-oesophagus connection, our results revealed a more likely link of Barrett's with the stomach.

Development of computational tools for variant calling in single-cell RNAseq

Kinga Anna Zielińska

St. Peter's College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Michaelmas 2022

Acknowledgements

First and foremost, I would like to thank my supervisors: Dr Benjamin Schuster-Böckler and Dr Francesco Boccellato. Dr Benjamin, for his expertise in best data science and bioinformatics practices, and for teaching me how to conduct my research independently. Dr Francesco, for assisting me in the final stages of my DPhil by providing constructive feedback and mental support. I would also like to extend my gratitude to the whole Ludwig Institute for Cancer Research, the Medical Sciences Division and St. Peter's College. It was the diverse community of highly motivated researchers that taught me how to think critically and inspired me to continue my career in Science.

I would also like to thank my friends and colleagues, especially Magdalena Drożdż, for sharing knowledge, experience and everyday struggles. All the members of the Triathlon, Cycling and Cross Country University Clubs, for providing a distraction from the academic life. I have always believed that a strong body is a requirement for a strong mind, and you created an environment that allowed me to push myself physically every day. You were not only great training partners but, most importantly, great friends sharing same passions and goals.

Last but not least, I would like to thank my family and friends from outside of the academic life. My Mother, for always checking up on me, my Father, for making sure I am open to new opportunities, my Sister, for being a good listener, and my Grandparents, for being the best supporters. Friends, for reminding me of who I am, why I ever decided to do a PhD, and for making sure I do not give up. Finally, huge thanks to my Partner for keeping my stress levels in check and providing stable support during the unsteady times of the Covid pandemic and war in Europe.

Abstract

Single-cell sequencing technologies have unsurprisingly become a favourable choice for studying key biological questions about cell heterogeneity, rare cell types or lineages. It is only cell-level resolution that allows for an accurate analysis of internal cell processes such as mutagenesis. Eventually, single-cell RNAseq could provide an explanation of mechanisms that lead to the ultimate transformation of healthy tissues into cancerous lesions. One of the main interests of my lab is Barrett's oesophagus. It is a highly clonal disease and a likely cancer precursor. We decided to take advantage of the single-cell RNAseq technology in order to attempt to identify the tissue of origin of the disease which, despite years of research, still remains unknown. However, the range of methods for identification of mutations in single-cells is very limited. In order to address that, we developed our own single-cell RNAseq variant caller. We validated it on a publicly available breast cancer dataset by achieving a reasonable intersection of our results with the output of commonly used bulk tools. Furthermore, we showed that our caller was capable of identifying expected data characteristics such as known breast cancer signatures and mutations in breast cancer genes. We then applied our method to the Barrett's dataset to investigate connections of Barrett's with surrounding tissues. Contrary to the previous transcriptomic analysis conducted on the same dataset and indicating a Barrett's-oesophagus connection, our results revealed a more likely link of Barrett's with the stomach.

Table of Contents

List of Figures

List of Tables

1. Introduction

1.1 DNA

1.1.1 Mutagenesis in healthy tissues

DNA was assumed to be a very stable molecule when it was first discovered as the carrier of genetic information (Watson and Crick, 1953). In reality, a genome is highly vulnerable to alterations due to oxidation, hydrolysis or alkylation (Lindahl, 1993). DNA damage involves physical modifications of its structure – ranging from breaks, depurination, depyrimidination to modified bases and crosslinks (Zhang and Vijg, 2018). The apparent stability of the genome can be attributed to a highly conserved system of genome maintenance mechanisms (Hoeijmakers, 2001). Every day, thousands of modifications are introduced in a typical cell. If it was not for the complex genome maintenance systems playing a crucial role in their elimination, DNA damage would not be as sporadic as it is in a typical mammalian organism (Collins et al., 2004).

DNA damage is not the only form of genomic alterations. DNA changes that arise during DNA damage repair, replication and cell division are collectively termed DNA mutations and occur naturally in normal cells (Zhang and Vijg, 2018). Common types of genetic variation include single-nucleotide variants (SNVs), copy-number variants (CNVs), and other structural variants (SVs). Unlike DNA damage, DNA mutations are not recognized by repair enzymes and are, therefore, irrevocable (Lindahl and Wood, 1999).

Genetic heterogeneity can be described as the one occurring in a group of individuals (population-level) or specific to one organism (individual-level). Population-level

heterogeneity arises before the formation of a zygote, due to germline mutations. Those are inherited by all progenitor cells. Somatic mutations, an expression of individual-level heterogeneity, exist only in a subpopulation of cells as they occur post-zygotically (Dou et al., 2018). The presence of multiple populations of cells with distinct genotypes in an individual have been termed as "somatic mosaicism" (De, 2011).

Because germline mutations provide a constant genetic variation in organisms, together with natural selection they drive evolution. Mutation rates differ between species (Baer et al., 2007), are well balanced between being too high and too low, and are subject to natural selection (Sturtevant, 1937). Both excessively high and low mutation rates, preventing a species from adapting to environmental changes, would lead to extinction. Therefore, genomic instability in the germline is a necessary phenomenon in order for a species to survive (Corbett et al., 2018).

Somatic mutations accumulate spontaneously throughout a person's lifetime. While most of them are harmless, they can occasionally have a phenotypic consequence as a result of interference with a gene or its regulatory element (Martincorena and Campbell, 2015). If a somatic mutation gives selective advantage to a cell, it can lead to preferential growth or survival of a clone. Such mutations, usually under positive selection, are termed "driver mutations" (Stratton et al., 2009). The end result of somatic evolution, apart from ageing, is cancer. In this disease, an autonomous clone of cells evades normal behaviour and gradually accumulates alterations in cell physiology that dictate malignant growth (Hanahan and Weinberg, 2000).

The germline mutation rate in humans can be estimated in a relatively straightforward way. It can be done by tissue sequencing from parents and their offspring, and the differences

will indicate an occurrence of de novo mutations (Zhang and Vijg, 2018). The human germline mutation rates are being approximated at around 1.0-1.5×10^{-8} per nucleotide per generation (Conrad et al., 2011, Rahbari et al., 2016). Consortia such as the 1,000 Genomes Project (The 1000 Genomes Project Consortium et al., 2015) have been launched to investigate differences between individuals. They concluded that two typical human genomes would differ by 4-5 million sites, and most divergences would take the form of single nucleotide polymorphisms (SNPs). Short indels or structural variations, although detected, were much less frequent.

De novo somatic mutations are much more difficult to estimate. Because they arise in individual cells and spread slowly, after sequencing they highly resemble low frequency sequencing errors (Zhang and Vijg, 2018). Approximations of normal mutation rates in mammals range from 1 to 5×10^{-10} mutations per base pair per cell division (Lee-Six et al., 2018,Werner et al., 2020). This is a seemingly low value until the size of the genome is considered – in that case, 3 billion variants are expected in humans on average (Mustjoki and Young, 2021). There has recently been a number of studies investigating the somatic mutation landscape in different healthy human tissues, mainly via deep DNA sequencing. Examples of such studies include tissues such as liver (Brunner et al., 2019), bronchus (Yoshida et al., 2020), brain (Lodato et al., 2018, Bae et al., 2018), blood cells (Lee-Six et al., 2018, Watson et al., 2020), colon and rectum (Lee-Six et al., 2019), endometrial epithelium (Moore et al., 2020), skin (Martincorena et al., 2015, Tang et al., 2020) and oesophagus (Martincorena et al., 2018, Yokoyama et al., 2019). Li et al. argue that while the aforementioned studies contribute greatly to the knowledge of mutation rates, driver genes and mutagenic factors, cross-organ comparison is unreliable due to the fact the samples came

from different donors with distinct germline backgrounds and life histories (Li et al., 2021). In order to compare the organs directly, they performed a comprehensive genomic analysis of over 1,700 normal tissue biopsies from 5 donors. They found widespread, but occurring to variable extents, somatic mutation accumulations and clonal expansions. In tissues such as rectum, colon and duodenum, somatic clones evolved independently and were microscopic in size, potentially limited by local tissue structures. On the other hand, macroscopic somatic clones in oesophagus and cardia were frequently expanded to hundreds of micrometres. The results highlight the importance of comparing mutation landscapes in samples from the same individual.

1.1.2 Calling variants from DNA sequencing

A pipeline for calling variants from DNA sequencing has now been well established, both for germline and somatic variants (**Figure 1**). The calling is usually preceded by extensive quality control and sample pre-processing (**Figure 1a**). The simplest mode of identifying mutations relies on calling variants from a single DNA sample of interest. However, more complex calling modes have been available in order to address a wide range of questions that follow the advances in next-generation sequencing (NGS) technologies. For example, joint calling of variants from parents and their child enables the phasing of variants (**Figure 1b**). Or when tumour variants are desired, somatic variant calling from tumour-normal matched pairs is performed (**Figure 1c**).

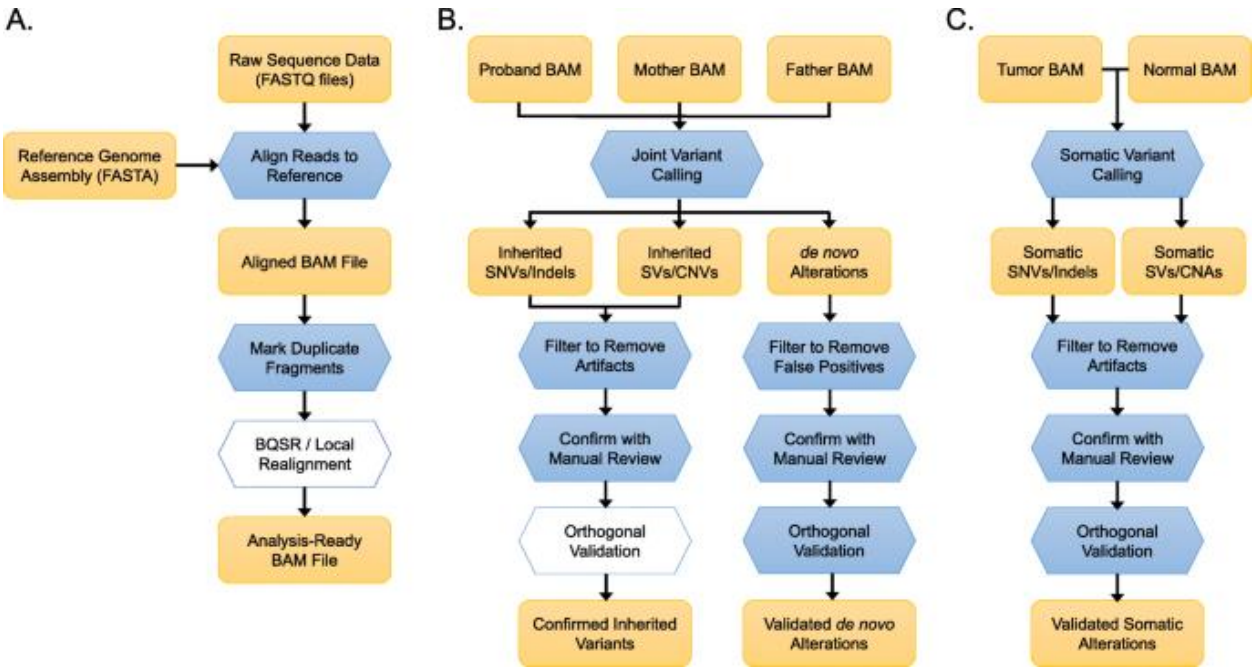


Figure 1. Standard pipeline for variant calling from DNA sequencing. a. Alignment and pre-processing of data. b. Variant calling in trio sequencing. c. Somatic variant calling in matched tumour-normal pairs. Source: Koboldt, 2020.

1.1.2.1 Germline variant calling

Dozens of germline variant callers have been published in the last decade, an even more have been developed for internal use by individual researchers. One of the most popular tools is the Genome Analysis Toolkit (GATK) HaplotypeCaller (McKenna et al., 2010), which consists of pre-processing steps, followed by variant calling and final filtering. The pre-processing workflow includes sample realignment, marking duplicate reads and recalculating base qualities. The HaplotypeCaller itself is a tool capable of calling SNPs and indels simultaneously via local de-novo assembly of haplotypes. The reassembly of reads at every position of interest enables the program to be more accurate in regions that are usually difficult to call, such as where different types of variants occur at close proximity. Another

example of a germline variant caller is Octopus (Cooke et al., 2018). Octopus is a mapping-based variant caller, which constructs a tree of haplotypes and dynamically prunes and extends the tree based on haplotype posterior probabilities. The ability to implicitly consider all possible haplotypes this way enables Octopus to find an optimal solution in reasonable time. FreeBayes (Garrison and Marth, 2012), Samtools/BCFtools (Li, 2011) and Platypus (Rimmer et al., 2014) further extend the list of most commonly used SNP callers. Because SNP detection in tools such as Haplotypecaller has been shown to have very high accuracy, selection of a single caller is usually sufficient in most research settings (Chen et al., 2020). However, combining the outputs of two callers using different approaches might result in a slight sensitivity advantage (Koboldt, 2020).

1.1.2.2 Somatic variant calling

Somatic variant calling is usually performed with a paired disease-normal sample strategy, for example by combining a tumour tissue with a matched normal sample (skin or blood) from the same patient. While disease-only variant detection has been adopted in some research settings in order to decrease the costs, the ability of this approach to detect somatic mutations is heavily compromised (Hiltemann et al., 2015). One of the most popular callers is the somatic variant calling mode of the GATK pipeline, MuTect2 (Cibulskis et al., 2013). MuTect2, and its predecessor MuTect, is a method that applies a Bayesian classifier to identify mutations with very low allelic frequencies. Thanks to carefully tuned filters, high specificity is ensured even when a variant is supported by just a few reads. MuTect2 is particularly effective at studying cancer subclones and their evolution with respect to normal tissues. Somatic variants can also be identified from DNA samples using the aforementioned

Octopus (Cooke et al., 2018) or tools such as Strelka (Saunders et al., 2012) and VarDict (Lai et al., 2016). Strelka uses a Bayesian approach which considers a normal sample as a mixture of germline variants with noise, and the tumour as a mixture of the normal with somatic variants. As a result of that, it is able to define continuous allelic frequencies for tumour and matched normal samples, while considering the expected genotype of the normal. Because of the way the model is structured, high sensitivity can be achieved even at high tumour impurity. VarDict, on the other hand, performs a local realignment on the fly and simultaneously calls a range of variants, such as SNVs, indels and complex structural alterations. This procedure allows it to estimate allelic frequencies more accurately. When using paired samples, VarDict is able to not only detect somatic mutations, but also loss of heterozygosity between the two tissues.

There have been a number of studies aiming to benchmark and compare the performance of somatic variant callers in different research settings (Krøigård et al., 2016, Xu et al., 2014, Wang et al., 2013). They found that because each caller had strengths and weaknesses, no tool appeared to offer superior performance. Therefore, an ensemble approach combining the output of multiple callers is often advised to reach an optimal balance of sensitivity and specificity (Callari et al., 2017, Fang et al., 2015). However, probably the most reliable method so far is purely experimental – and relies on the Sanger sequencing validation of the mutations (De Cario et al., 2020).

1.2 RNA

1.2.1 The definition and role of RNA

According to the central dogma of molecular biology, the information stored in genes as DNA is transcribed into RNA, to ultimately be translated into proteins (Crick, 1970). The transcription of specific genes into complementary RNA, dependent on environmental factors, forms the phenotype of an individual. In this form, it is responsible for specifying a cell's identity and regulating biological activities within the cell (Kim and Eberwine, 2010).

Historically, RNA molecules were considered a simple intermediate between genes and proteins. However, is has since been known that there is a high degree of variety within RNA molecule types and the analysis of RNA is not limited to mRNA molecules which encode proteins via the genetic code. Instead, there has been an increasing focus on the noncoding functional RNA molecules (ncRNA). Among those are ribosomal RNAs and transfer RNAs involved in mRNA translation, small nuclear RNA (snRNAs) associated with splicing, and small nucleolar RNAs (snoRNAs) active in the modification of rRNAs (Mattick and Makunin, 2006). Collectively termed the transcriptome, the aforementioned RNA molecules are indispensable to interpreting the functional regions of the genome and understanding mechanisms of development and disease (Kukurba and Montgomery, 2015).

1.2.2 RNA sequencing and its advantages

RNA sequencing (RNAseq) is a method that resulted from the development of the high-throughput next-generation DNA sequencing technologies. It has revolutionized the understanding of the compound nature of the transcriptome by enabling analysis through

the sequencing of complementary DNA (cDNA) (Wang et al., 2009). RNAseq has clear advantages over existing DNA approaches, as it provides a quantitative insight into gene expression, allele-specific expression and alternative splicing (Kukurba and Montgomery, 2015). Calling variants from RNAseq allows for combining both genomic and transcriptomic information in order to get a deeper understanding of the samples analyzed. RNAseq is cheaper than the sequencing of the whole genome, therefore, a large number of studies are limited to solely performing RNA-seq analysis. When no paired DNA is available, the ability to call genomic variants from RNA-seq alone is invaluable. Due to heterogeneity of diseases like cancers, SNP calling from WGS or WES can be challenging. In this case, variant calling from RNA-seq can be a useful validation method (Piskol et al., 2013).

1.2.3 Current methods in RNAseq variant calling

While calling variants from RNAseq offers significant advantages, there are obvious drawbacks such as being able to identify variants solely from regions that are expressed. In addition to that, there are a number of RNA-specific issues that need to be addressed. The main challenges involve splice junctions, management of duplicated reads or identifying variants in regions with low coverage due to, for example, poor gene expression (Brouard et al., 2019). One of the most commonly used tools for discovering and genotyping variants from the NGS RNAseq data is the GATK (McKenna et al., 2010), developed to call variants from DNA. There have been a number of early callers, such as SNPiR (Piskol et al., 2013), that would expand on the capacities of GATK and apply additional filtering to identify variants from RNAseq by matching its unique characteristics. However, it has since been

possible to accurately call mutations within the GATK pipeline itself by modifying some of its steps. An example of such adjustments is the SplitNCigarReads procedure. The main goal of this step is to reformat alignments that span introns. It is done by splitting reads with N in the cigar into multiple supplementary alignments and hard clipping mismatching overhangs. In addition to that, mapping qualities are reassigned to match DNA conventions (Brouard et al., 2019). Another example of tool that enables calling variants from RNAseq is the aforementioned Octopus (Cooke et al., 2018), which has a specific RNAseq mode.

It has been shown that RNAseq is a very accurate method of germline detection (Quinn et al., 2013). However, this has not been stated about somatic mutations, the calling of which the aforementioned methods do not cover. There have been attempts to identify somatic variants from RNAseq (García-Nieto et al., 2019), however, they take the form of filtering approaches (using the GATK statistics) rather than actual callers. To our knowledge, the only complete method so far developed for the purpose of somatic variant calling from RNA-seq is RNA-MuTect (Yizhak et al., 2018). RNA-MuTect consists of a set of filters that take advantage of the SNV calling from matched DNA samples in order to confirm the existence of mutations in RNAseq, but forms a complete pipeline. The key filtering steps involve the removal of mapping errors using both STAR and Hisat2 aligners, removal of sequencing errors by a site-specific error model built upon thousands of normal RNAseq datasets, and removal of RNA editing sites using known databases. RNA-MuTect is claimed to achieve high sensitivity and precision, and to outperform previous methods (Tang et al., 2014). However, it had been developed to target macroscopic clones rather than all somatic variants in general, especially those with lower variant allelic frequencies.

In conclusion, it has now been possible to identify germline variants from RNAseq with high accuracy, thanks to tools such as the GATK and Octopus. However, there is limited scope for calling somatic mutations, apart from RNA-MuTect which targets macroscopic clones. In addition to that, the existing RNAseq callers require matched DNA or WES, therefore, they cannot be applied to studies where only RNAseq data is available.

1.3 Single-cell RNA sequencing

1.3.1 The definition and reasons for calling variants from single-cell RNAseq

In the past decades, bulk RNA sequencing methods have been widely used to study gene expression patterns at population level. While they opened numerous opportunities to study what had previously not been accessible, they could only get the average of many cells and would inevitably lose cellular heterogeneity information (Chen et al., 2019).

Single-cell sequencing technologies refer to the sequencing of a single-cell genome or transcriptome, in order to obtain genomic, transcriptomic or other multi-omics information at the single-cell level. It is, therefore, not surprising that they have become a favorable choice for studying key biological questions about cell heterogeneity, identifying rare cell types or delineating cell maps (Tang et al., 2019). In fact, single-cell RNA sequencing proved to be so effective that it was named the “Method of the Year 2013” by Nature Methods (“Method of the Year 2013 | Nature Methods,”).

1.3.2 Pure bulk RNAseq callers cannot be used to call variants from single-cells

Due to the nature of single-cell data, calling variants requires certain adjustments, and pure bulk RNAseq variant callers are not suitable for identifying mutations from single-cells (Stegle et al., 2015). The main difference between single-cells and bulk RNAseq is the abundance of the starting material. Because single-cell samples contain data from individual cells, the initial number of RNA sequences will be very low, even despite artificial amplification using PCR. Moreover, many regions will not be covered at all, due to cell-specific expression or the state that the cell is in (Lähnemann et al., 2020). Any technical errors or biases will have a stronger impact on the final appearance of data, as they will be present in a greater fraction of the sparse sequences. The observed allelic frequencies of the variants will also be different. Because bulk samples consist of a large number of single-cells, the allelic frequencies of some variants will, in general, be lower. It is due to the fact that SNVs would be present in only a subset of cells, and when combining sequences from all cells, the number of reads with the SNV would be comparably low. In healthy single-cells, variants have an allelic frequency of either 100% (homozygous mutations) or 50% (heterozygous mutations). Therefore, bulk methods should not be used to call variants from single-cells, unless they are modified to suit the specific nature of single-cell data.

1.3.3 Current methods in single-cell RNAseq variant calling

Identifying variants from single-cell RNAseq is difficult, as there is a wide range of technical artefacts present such as reverse transcriptase errors, PCR errors or sequencing biases, all of which are discussed in the subsequent sections of the Thesis. There have been a number of approaches to distinguish variants from artefacts in single-cell RNAseq. Most

of them are based on tools developed specifically for bulk RNAseq, but could potentially be suitable for single-cells provided parameters are adjusted accordingly. Liu et al. performed a systematic comparison of seven bulk RNAseq tools (SAMtools, the GATK pipeline, CTAT (Fangal, 2020), FreeBayes, MuTect2, Strelka2, and VarScan2 (Koboldt et al., 2012)) in terms of their ability to call variants from single-cell RNAseq. While the specificities were generally high, predictably the sensitivities would dramatically decrease in regions with low read depths, low variant allele frequencies or in certain genomic contexts. All callers had performance trade-offs. SAMtools showed the highest sensitivity in regions with poor coverage, but it was not able to perform well in the presence of introns or high-identity regions. Similarly, FreeBayes showed high sensitivities with high allelic frequencies, but the specificities were inconsistent between different datasets. In conclusion, the results indicated the necessity of improving detection sensitivity in difficult regions that could only be achieved when developing callers specific for single-cell RNAseq (Liu et al., 2019).

SCmut is one of the first approaches to adjust bulk RNAseq callers to identify variants from single-cells specifically. Its output is based on somatic calls identified by other callers from bulk RNAseq with matched DNA normal. By applying an additional filter (two-dimensional local false discovery rate) that statistically detects somatic mutations at cell level, it removes a substantial amount of false positives commonly produced by bulk methods when calling variants from single-cells (Vu et al., 2019). Despite being quite accurate, variant discovery with SCmut is limited to whatever is detectable from the matched samples.

Developed specifically for single-cell RNA seq, Red Panda takes into consideration the unique nature of single-cell RNAseq data and, therefore, has a distinct advantage over other methods. It classifies all putative variants into three categories: homozygous-looking, bimodally-distributed heterozygous, and non-bimodally-distributed heterozygous variants. The main limitation of the method is its reliability on information from other callers (GATK Haplotypecaller). Furthermore, even the authors state that despite the 72.44% specificity, there is an ongoing need for improvement (Cornish et al., 2020).

1.3.4 Goals of the thesis

The main goal of the DPhil was to develop a stand-alone variant caller that would identify somatic mutations from single-cell RNAseq with high accuracy without the need for matched DNA or bulk RNAseq. The following thesis contains detailed descriptions of how the tool was created and validated. We then present its application to our Barrett's dataset, an attempt to resolve unanswered questions about the origins of the disease.

2. Development of the single-cell RNAseq caller

1.4 Introduction

While bulk RNA sequencing methods have been widely used to study gene expression patterns for many years, single-cell sequencing has proven to be a favourable choice for obtaining information at the single-cell level. It has since been applied to study key biological questions about cell heterogeneity, identifying rare cell types or delineating cell maps (Tang et al., 2019). The analyses have previously been conducted solely from the transcriptomic perspective. By developing a variant caller, we are hoping to combine transcriptomic and genomic information to get a novel perspective on phenomena occurring within and between cells. Specifically, we aim to investigate whether we could identify mechanisms of mutagenesis in healthy tissues that lead to diseases such as cancer.

The single-cell RNAseq workflow is a complex, but a well-established, protocol. The first step involves an extraction of suitable cells from the tissue of interest. Individual cells are lysed and the RNA molecules are captured. In some cases, poly[T]-primers to remove ribosomal RNA molecules are used, as a result of what only polyadenylated mRNAs are obtained (Haque et al., 2017). The next step involves conversion of the poly[T]-primed mRNA to complementary cDNA using a reverse transcriptase enzyme. If unambiguous marking of the sequence is desired, additional adaptor sequences or unique molecular identifiers (UMIs) are used (Kivioja et al., 2011). The minimal amounts of cDNA are not

sufficient for subsequent sequencing, therefore, PCR amplification is performed. It is recommended to execute multiple cycles of PCR, separated by fragmentation (Thermo Fisher Scientific - UK). In the final step of the workflow, the amplified and tagged cDNA is pooled and passed to a sequencing platform. There is a lot of variety in terms of reagents, machines and settings used. Those are usually dataset-specific, and can be modified depending on the experiment conducted (Haque et al., 2017).

1.4.1 Technical errors in single-cell RNAseq

Practically every step of the single-cell RNAseq workflow is likely to introduce technical biases and errors. They can start as early as cell extraction and lysis, where important cells are missed or cells are captured together (Hu et al., 2016). Errors can also arise while converting mRNA into cDNA due to infidelity of the reverse transcriptase, which lacks proofreading ability and, consequently, has a higher error rate (Li and Lynch, 2020). There are two main sources of errors that occur in subsequent PCR amplification: the error rate of the PCR polymerase and the thermal damage of the cDNA (Pienaar et al., 2006). In addition to that, sequences could be amplified at different levels of efficacy, which is dependent on factors like transcript length or GC content (Dabney and Meyer, 2012). Template switching is another contributor, occurring more frequently within the final amplification steps (Balázs et al., 2019). Depending on how early during amplification the errors occur, they have a varying impact on the rest of the sequences. For example, if a PCR error is introduced in the first amplification round, it will be present in as much as a half of the final amplicon set. On the other hand, if it is introduced during one of the last amplification steps, its frequency will be very low and most likely undetectable. Finally, errors can be introduced

at the sequencing stage. As only a small share of all amplicons is selected for sequencing, the error frequency might change again (we are not aware of any studies that determine the degree of bias, however). There have been a number of studies aiming to estimate the degree of bias and suggesting ways of reducing it, but they are not universally applicable to all experiments and, therefore, not really helpful in minimizing biases in custom datasets (Lahens et al., 2014). Sequencing errors are dependent on the sequencing method. It is estimated that on average 1 in 1,000 bases from Illumina contain a sequencing error (Pfeiffer et al., 2018), however, there is a lot of variation depending on data characteristics and technology used. While there is an up to 10-fold difference in error rates between Illumina sequencers alone, the difference between samples coming from the same sequencer is even more striking. Some studies suggest it might be due to oxidative damage introduced as a result of differential sample handling (Ma et al., 2019). Error rates are also variable at the level of individual reads, as they are highly correlated with the sequencing cycle and tend to increase towards the end of the read. Furthermore, they are highly dependent on trinucleotide contexts and often increase in the presence of certain sequence motifs (Stoler and Nekrutenko, 2021). Once the experimental steps are completed, it is the computational post-processing where the last errors occur. Alignment errors arise from incorrect mapping of the reads by an aligner. They are specific to the mapping software used, as there is a lot of variety in how different aligners deal with problematic areas (Alser et al., 2021). For example, many mismatches in a read would confuse aligners which require exact seed matching and extension (Sun and Buhler, 2006). Splicing or large indels are another such issue, as a way of dealing with gaps in the sequence is aligner-specific (Sahlin and Mäkinen, 2021). Furthermore, there are low-complexity regions, which are regions

very similar across different parts of a reference genome. They pose a challenge to all aligners, and often lead to different combinations of unique and non-unique alignments (Phan et al., 2015).

Because the amount of the starting material in single-cells is so minute, every mistake has a significant influence on the final appearance of the data. Errors introduced early will, in general, have higher frequencies, as they will undergo numerous rounds of amplification. An allelic frequency of such errors can sometimes be so high that they closely resemble real mutations. On the other hand, single-cell data processing has a tendency to result in "dropout" events, which indicate observed zeros (Qiu, 2020). Zeros are ambiguous and pose significant challenges to transcriptomic analyses. They can either be attributable to methodological noise (expressed genes not detected by technology used) or they can express genuine biological absence (Lähnemann et al., 2020). Therefore, they have serious implications in variant calling, as they make mutations undetectable or make them appear at different frequencies than they really are.

1.4.2 Existing methods

Existing single-cell RNA sequencing variant callers strongly rely on sample pre-processing and quality control. Strict quality thresholds, applied in order to minimize the number of false positives often identified in the more ambiguous regions of the genome (such as poor coverage), are usually the only method used. Examples of such thresholds include keeping only reads with maximum alignment scores or setting minimum limits for the numbers of supporting reads. Red Panda (Cornish et al., 2020) and SCmut (Vu et al., 2019) take it a step further, as instead of setting hard thresholds, they attempt to adjust them

based on mutation profiles observed in the data. As aforementioned, mappers have different strategies of handling problematic aspects of RNAseq. Such areas include multiple mismatches in a read, splicing or large indels, and low-complexity errors. Therefore, to reduce the mapping error rate, intersection of results from multiple aligners is recommended (Liu et al., 2019).

1.4.3 Our approach

The main goal of our work was to create a single-cell RNAseq-specific variant caller, which would be an uncomplicated independent tool. Specifically, we did not want it to rely on any bulk callers and aimed to limit the need for other tools. We hoped to achieve high accuracy while using only single-cell data, as obtaining and processing bulk or matched DNA samples is troublesome and expensive. Finally, we wanted to limit the number of steps the user would have to complete to process their samples by compiling our method as a complete Python program.

In order to develop a single-cell RNAseq caller, we decided to address the single-cell-specific issues from the very beginning, rather than improve on the output of bulk callers as most single-cell callers do. In the next sections of this chapter, we describe the development of our method. We begin with detailed explanations of technical errors that arise during single-cell RNAseq and outline simple steps required to eliminate some of them. We then describe how we used relationships between the remaining, more difficult to remove, errors, to ultimately create a complete caller. Finally, we show that combining calls from multiple single-cells substantially improves calls in individual samples.

1.5 Data and methods

The following section describes a dataset used to develop the single-cell RNAseq caller and methods applied to prepare it for the calling.

1.5.1 Data

The lab of Professor's Xin Lu (Ludwig Institute for Cancer Research, Oxford branch) generated a dataset consisting of bulk RNA-sequencing and single cell RNA-sequencing, further referred to as the "Barrett's dataset". In the analysis, we used data from 4 patients, separated into 6 batches as outlined in **Table 1**. While the original dataset included more individuals, only those patients had bulk RNAseq complemented with single-cell RNAseq data, and they all experienced symptoms of Barrett's. There were 4 bulk RNAseq replicates for every tissue and patient.

Table 1. Number of single-cells per batch, patient and tissue in the Barrett's dataset

Batch	Patient	Tissue	Number of single-cells
4	GEN02021	Barrett's	86
4	GEN02021	Gastric	86
4	GEN02021	Oesophagus	7
5	GEN02021	Barrett's	171
5	GEN02021	Gastric	46
6	GEN02021	Duodenum	90
6	GEN02021	Gastric	90
6	GEN02023	Barrett's	90
6	GEN02023	Gastric	90
6	GEN02024	Barrett's	90
6	GEN02024	Oesophagus	90
6	GEN02025	Duodenum	90
6	GEN02025	Oesophagus	90
7	GEN02021	Barrett's	93
7	GEN02021	Oesophagus	90
7	GEN02023	Duodenum	90
7	GEN02023	Oesophagus	65
7	GEN02024	Duodenum	82
7	GEN02024	Gastric	90
7	GEN02025	Barrett's	90
7	GEN02025	Gastric	90
8	GEN02021	Oesophagus	92
8	GEN02023	Oesophagus	92
8	GEN02024	Gastric	93
8	GEN02025	Gastric	93

Single cell RNAseq was prepared with a custom adaptation of the smart-seq2 method. Bulk RNAseq was created using the mir-Vana miRNA Isolation Kit (Thermofisher). ERCC spike-ins were added for quality control and the sequencing was done using the Illumina HiSeq 4000 system. A detailed description of data preparation has previously been published (Owen et al., 2018).

1.5.2 Alignment to the reference genome

The following section contains information about the reference genome used and describes the alignment of the samples.

1.5.2.1 Reference genome

The human reference genome used was hg38. Sequences from the Epstein-Barr virus (EBV) and Helicobacter pylori (HP) were included in the FASTA file, but only the human chromosomal regions were used in the analysis. Information about functional regions (known genetic variation, gene annotations) was downloaded from GENCODE and the UCSC genomic browser.

1.5.2.2 Single-cell RNAseq

The raw FASTQ files were aligned using two different aligners: STAR 2.6.0c and Hisat 2.0.4. The parameters used with STAR *were --readFilesCommand zcat --runThreadN 4 --outSAMtype BAM SortedByCoordinate --outSAMmapqUnique 60*. The parameters used with Hisat2 were *--min-intronlen 20 --max-intronlen 500000 -k 5 -X 800*. In order to maximize the quality of the alignment, all unmapped or multimapped reads were removed. Furthermore, only reads with the maximum mapping quality (MAPQ = 60, less than 5 mismatches) were used in the analysis.

1.5.2.3 Bulk RNAseq

The raw FASTQ files were aligned using STAR 2.6.0c by Dr Ruud G.P.M. var Stiphout. The parameters used were *--readFilesCommand zcat --runThreadN 4 --outSAMtype BAM SortedByCoordinate --outSAMmapqUnique 60*.

1.5.3 Quality control

The quality of the samples was assessed using a variety of methods, both before and after alignment.

1.5.3.1 Quality examination of raw reads with FastQC

FastQC was used to examine the quality of raw reads. It included the analysis of sequence length and quality, base quality and content, GC and N content, duplication levels, overrepresented sequences and adapter content.

1.5.3.2 Quality examination of aligned sequences

The quality of the aligned sequences was assessed using standard tools used in the analysis of BAM files, namely output from the aligners and the *samtools* toolkit. Reports produced by the alignments were used to evaluate general statistics for each sample, such as alignment rates (ratio of aligned reads to total reads) and percentages of uniquely mapped, multi-mapped and unmapped reads. The quality of individual reads and bases was examined using *samtools*. Metrics taken into account included mapping and base qualities, number of mismatches and length of the soft-clipping regions.

1.5.4 Ground truth variant calling (germline variants)

For every patient, 4 bulk RNAseq samples from every non-disease tissue (Gastric, Oesophagus and Duodenum) were collected. Each of the samples was processed separately. The aligned BAM files were processed according to the GATK 3.7.0. Best Practices and the germline variants were called with Haplotypecaller (default settings). The final variants were defined as those that passed the additional GATK VariantFiltration (default settings) step. After that, variants called for the same patient and tissue were grouped. All calls were taken into account, even if they were called in only one out of four samples from the same tissue. In order to create the final list of germline mutations per patient, calls from healthy tissues (Gastric, Oesophagus and Duodenum) were merged.

1.6 Development of the single-cell RNAseq caller

This section contains detailed information about the development of the single-cell RNAseq caller. It begins with a description of technical errors that are the easiest to remove, and presents measures taken to eliminate them. It then continues to the main part of the caller, and finally explains additional filters applied during the last steps of the calling in order to maximize the quality of the output.

1.6.1 Most technical error types can be eliminated with standard quality control measures

The following subsection explains the characteristics of technical errors in single-cell RNAseq and shows how they can be eliminated with standard quality control measures.

1.6.1.1 Alignment errors

Alignment errors arise from incorrect mapping of reads by an aligner. They are specific to the mapping software used, as there is a lot of variety in how different aligners deal with problematic areas.

As already described, all non-uniquely mapped reads and reads with poor mapping quality were removed to reduce the mapping error rate. For most samples in our work, different mappers were used (STAR and Hisat2, specifically) and a variant was only called if it was supported by the output of both aligners.

1.6.1.2 Sequencing errors

Sequencing quality scores, provided by Illumina for each base, measure the probability that a base is called incorrectly (Illumina, Inc. U.S). Therefore, some sequencing errors can be identified by a poor base quality score. We excluded all bases with quality <20, or any positions with average base quality <20. However, this is not sufficient to remove all sequencing errors, as the scores only express error probabilities and a risk exists that some errors are not marked.

Considering that 1 in 1000 sequenced bases are incorrect and the coverage in single-cell RNAseq hardly ever exceeds 1,000 reads, it is quite unlikely for two sequencing errors to be present at the same position, and even more unlikely for them to be of the same type. To make sure that we do not call a sequencing error in case of a very unlikely event that two sequencing errors of the same type are present at the same position, we excluded any variants that were not supported by at least 3 reads.

1.6.1.3 PCR errors

PCR errors arise during PCR amplification due to inaccuracy of the PCR polymerase at a rate of approximately 10^{-4} (Potapov and Ong, 2017). If there is only one PCR amplification step, those errors can be easily removed. They can be identified by their presence in only one PCR duplicate group, therefore, removing PCR duplicates with *samtools markdup* should be enough to filter them out. However, in single-cell RNAseq there is a pre-adapter-ligation (pre-fragmentation) amplification, which complicates the problem. In this case, two independent PCR amplification steps are performed and each of them carries technical errors of unlike characteristics. As described, errors from the pre-library-amplification (second round of PCR) can be removed using *samtools markdup*. In the next section, we explain how we estimated the frequency of the pre-adapter-amplification errors, and our attempts to eliminate them.

1.6.2 Errors from the first PCR amplification round (pre-adapter-amplification errors) are difficult to eliminate

Because amplicons from the pre-adapter-amplification undergo fragmentation, it is impossible to tell which duplicate group they originally came from. Therefore, tools like *samtools markdup* are not suitable to remove PCR errors from the pre-adapter-amplification (further referred to as PCR1 errors).

We made an attempt to estimate the frequency of such errors in order to estimate the scale of their influence on our calling. The crucial question is about the regularity of PCR1 errors that could resemble real variants. If 3 out of 10 reads supported an alternative allele, is it likely to be a PCR1 error? Is it still likely if 6 out of 10 reads support the variant?

We found the calculations to be challenging due to a large number of unknowns. Among the information we did not have was the initial number of transcripts, the probability of an amplicon to be passed to the pre-library-amplification PCR, or the probability / number of amplicons from the pre-library-amplification PCR to be chosen for sequencing. We hence had to make rough approximations, simulating a "bad" scenario (i.e. resulting in the highest allelic frequency of the PCR error).

Let us assume that there is one initial transcript, as in this case, the number of observed reads will be the lowest and the frequency of the error will be the highest. Even though PCR efficiency is dependent on factors like GC content and amplification round, in our calculations we assume that all amplicons are amplified at maximum rate (i.e., every amplicon is amplified at every round). After the pre-adapter-amplification is finished, all amplicons undergo fragmentation and are passed on to the pre-library-amplification, where they are, again, amplified with maximum efficiency. Up to this point, the original allelic frequency of the PCR1 error from the pre-adapter-amplification is unchanged. The next and final step to consider before sequencing is amplicon selection, as not all amplicons will be sequenced. We do not know how many amplicons with the PCR1 error are going to be in this group, but in order to estimate this, we can use the binomial distribution. The number of trials can be expressed as the total number of observed reads at the position of interest, number of successes is the observed number of reads with the PCR1 error, and the probability of success is the allelic frequency of the PCR1 error. The probability values output by the distribution should give us an indication about which combinations of reads with and without PCR1 errors are possible, hence answering our original question (given

that 3 out of 10 reads in our data support the alternative allele, is it likely to be a PCR error from the pre-adapter-amplification)?

The earlier a PCR error occurs during amplification, the higher its allelic frequency. In order to estimate the frequency of such errors in our data, we needed to evaluate the maximum number of amplification rounds the PCR1 error could be introduced by. In other words, the error would only be detectable if it had been introduced early enough (as otherwise, its frequency would be too low). We started our calculations with determining the allelic frequency of a PCR error at each PCR amplification round (**Table 2**).

Table 2. Allelic frequency of PCR errors from the first round of PCR at each amplification round.

Amplification round	Total number of amplicons	AF of PCR error
1	2	0.5000000
2	4	0.2500000
3	8	0.1250000
4	16	0.0625000
5	32	0.0312500
6	64	0.0156250
7	128	0.0078125

The minimum number of reads supporting an alternative allele to be considered during our calling is 3 and the allelic frequency is 0.1. As shown in **Table 3**, we considered 3 extreme scenarios - 3 out of 30 reads supporting the alternative variant (minimal coverage required to call a variant with AF = 0.1), plus 3/10 and 3/5 as most positions in our data have lower coverage and higher allelic frequencies. Using the binomial distribution

(cumulative) as described earlier, we were able to calculate the probability of observing at least 3 reads for each coverage value. For example, for the case with 3 out of 30 reads supporting the alternative allele with a PCR error introduced during the second round of pre-library-amplification, the number of trials will be 30, number of successes 3, and the probability of success will be 0.25 (AF of PCR error corresponding to 2 rounds of amplification, taken from Table 2.). The probabilities of observing at least 3 reads were higher for lower allelic frequencies (3/30 versus 3/5), what is not surprising as one would expect to see 3 desired reads among a pool of 30 more often than among 5.

Table 3. Cumulative binomial probabilities of detecting a PCR1 error for different combinations of amplification rounds and supporting reads.

	Number of supporting reads / coverage		
Amplification round	**3/30**	**3/10**	**3/5**
1	>0.99	0.9450	0.500
2	0.989	0.4740	0.104
3	0.742	0.1200	0.016
4	0.288	0.0210	<0.016
5	0.066	<0.021	<0.016
6	0.011	<0.021	<0.016

The PCR errors can be introduced in all rounds of amplification. Therefore, if 18 rounds are performed, the number of positions in which potential errors can be introduced is 2^{17}. Assuming that the PCR error rate is 10^{-4}, one could expect 13 PCR errors at each genomic position.

A PCR1 error's detectability is dependent on two factors: probability of observing a sufficient number of reads (**Table 3**) and, fundamentally, the error being introduced in the

first N amplifications. This can be estimated using the binomial distribution, as in the previous case. The success rate can be expressed as the number of PCR errors versus the number of all amplicons in which they could be introduced ($13/2^{17}$), and the number of trials corresponds to the number of opportunities the error could be introduced in by the N amplification round (number of amplicons produced at the N-1 amplification round). For example, at the second round of amplification, the number of trials will be 2 (errors could only be introduced in 2 amplicons from the first round). The probabilities of observing 1 PCR error at N amplification round are presented in **Table 4**. Interestingly, the probability of observing more than 1 PCR error at each of the considered N amplification rounds was always 0, suggesting that it was not likely to observe more than 1 PCR error per genomic position.

Table 4. Probability of observing 1 PCR error in the first N amplification rounds

Amplification round (N)	Probability
1	0.0001
2	0.0002
3	0.0004
4	0.0008
5	0.0016
6	0.0032

Based on probabilities of detecting variants from **Table 3**, we were able to calculate the probability of observing a PCR1 error at a genomic position (probability of the error occurring in the first N amplification rounds from **Table 4** multiplied by the probability of observing a sufficient number of reads from **Table 3**). The resulting probabilities are

presented in **Table 5.** As previously, the "3/30" case carried the highest probability of occurrence. However, the probability did not decrease with the amplification rounds N, and it was more likely to observe PCR errors introduced within the first 2 or even 3 amplification rounds than in the first one, despite a lower allelic frequency. This was due to a greater number of amplicons the errors could be introduced in (greater difference between probabilities of a PCR error occurring in the N amplification rounds than probabilities of detecting a sufficient number of reads for N=2 or 3).

Table 5. Cumulative binomial probabilities of observing a PCR1 error for different combinations of amplification rounds and supporting reads.

	Number of supporting reads / coverage		
Amplification round (N)	**3/30**	**3/10**	**3/5**
1	>0.000099	0.000094500	0.000050000
2	0.0001978	0.000094800	0.000020800
3	0.0002968	0.000048000	0.000006400
4	0.0002304	0.000016800	0.000001776
5	0.0001056	0.000004960	0.000000464
6	0.0000352	0.000001344	0,000000128

The values from **Table 5** could be directly translated to the number of observable PCR errors per sample. Estimating that there were, on average, 120,000 chromosomal positions of high quality per cell in which the errors could be introduced, we calculated the number of PCR1 errors per sample, relative to the N amplification rounds that an error was introduced in, by multiplying the probabilities by the number of chromosomal positions (**Table 6**).

Table 6. Expected number of PCR1 errors per sample, with the corresponding amplification round that the error was introduced by

	Number of supporting reads / coverage		
Amplification round (N)	**3/30**	**3/10**	**3/5**
1	12	11	6
2	24	11	2
3	36	6	<1
4	28	2	<1
5	13	<1	<1
6	4	<1	<1

Beyond 6 amplification rounds the number of PCR1 errors becomes negligible. Therefore, we believed that setting the threshold for the number of amplification rounds N to 6 was feasible. Because we expected only 1 PCR error per genomic position, the final probability of observing a PCR1 error could be obtained by summing probabilities of observing the error at each amplification round N (mutually exclusive events), or, alternatively, summing the expected numbers of errors per sample. Therefore, the expected numbers of PCR1 errors per sample were 117, 30 and 8 for the cases “3/30”, “3/10” and “3/5”, respectively. While different allelic frequencies and combinations of supporting reads were possible, the case “3/30” was the one with the highest number of expected errors (lowest allelic frequency and fewest supporting reads). Therefore, up to 117 observable PCR1 errors were expected per sample. While we have not managed to produce a method to remove the PCR1 errors directly, we are going to show how we handle them in the later sections, and how we minimize their contribution to our final list of calls.

1.6.3 Real mutations have at least the same allelic frequency as reverse transcriptase errors

Reverse transcriptase is an enzyme used for translation of RNA into cDNA and is very prone to errors as it has no proofreading ability. There has been a limited number of attempts to estimate the error rate of the reverse transcriptase, and the values are not consistent between publications (most likely somewhere in the range of 10^{-5} (Gout et al., 2017)). These kinds of errors are the most difficult to identify, as they appear the earliest and are the most similar to real mutations.

Real mutations, due to the fact they are already present before reverse transcriptase errors occur, have at least the same or higher allelic frequency than transcriptase errors. Similarly, reverse transcriptase errors will have at least the same allelic frequency as PCR errors from the first round. Therefore, determination of the reverse transcriptase allelic frequency should allow us to determine an allelic frequency threshold above which all remaining errors (PCR1 and reverse transcriptase) are unlikely.

1.6.3.1 Estimation of error allelic frequency threshold from spike-ins cannot be translated to chromosomal regions

In the following section, we present our method of estimating the expected allelic frequency threshold above which technical errors are not expected and, ultimately, all variants are real mutations.

The allelic frequency of the reverse transcriptase error is dependent on the number of original transcripts in the sample and the number of transcripts it is introduced in (most likely introduced in only 1 original transcript, as both the error rate of the reverse transcriptase and coverage are very low in our data). It has been shown that the final number

of reads covering a position is proportional to the number of the original transcripts in our data (Owen, 2018). Therefore, one could expect the errors to occur at similar frequencies when comparing them across positions with similar depths, equal to 1 divided by the number of original transcripts. A position with one transcript would result in a reverse transcriptase error frequency equal to 1, a position with two transcripts equal to 0.5, a position with three transcripts equal to ~0.3 and so on.

In order to validate this statement, we decided to use the spike-in sequences included in our single-cell data (as we did not expect them to have any mutations). We removed the hypothetical sequencing, mapping and PCR2 errors using techniques described in the earlier sections of this chapter. By plotting allelic frequency of the remaining technical errors versus coverage , we aimed to determine the allelic frequency threshold above which no technical errors occur. As expected, we observed a decrease in allelic frequency of technical errors as coverage increased (**Figure 2**).

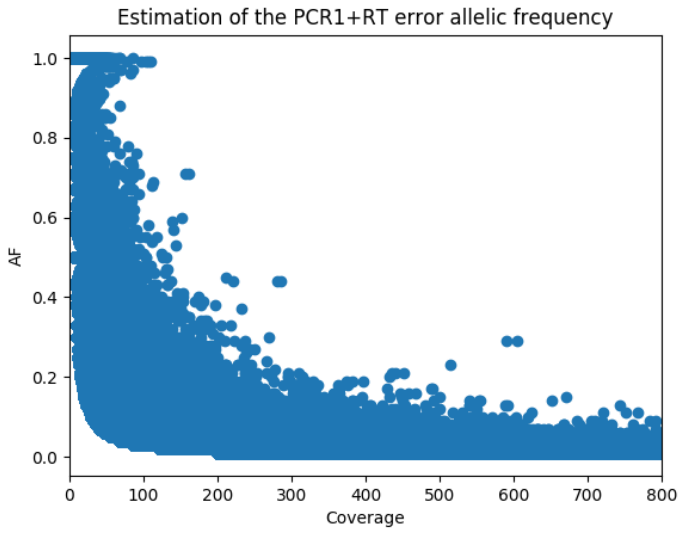


Figure 2. Allelic frequency of technical errors versus coverage in ERCC spike-ins.

We made rough approximations of the allelic frequency thresholds for different coverage ranges (**Table 7**). Applying those thresholds to positions different from the reference in the spike-in regions would be sufficient to remove most of the remaining errors in the spike-in regions (subject to noise).

Table 7. Allelic frequency thresholds for PCR1 and reverse transcriptase errors in the ERCC spike-ins.

Coverage (DP)	Max AF of PCR1+RT errors
DP>600	0.2
300<DP<=600	0.3
200<DP<=300	0.5
150<DP<=200	0.7
DP<=150	1.0

Naturally, we would proceed to apply the same procedure to chromosomal regions. Removing all variants below the just defined threshold frequency would, theoretically, ensure the sole preservation of real mutations. However, most chromosomal positions in our dataset have low coverage (<100) and considering that the allelic frequency threshold in this coverage range has been defined as 1.0, subjecting to this restriction would result in no variants being called. Therefore, while setting a fixed allelic frequency threshold could be a reasonable option for positions with high coverage, finding a different method for identifying mutations in regions with fewer reads was necessary.

1.6.4 The “linkage method” can distinguish between reverse transcriptase errors and real mutations in specific conditions

Because it was not an effective solution to set a fixed allelic frequency threshold above which no errors occur, especially in regions with poor coverage, we developed an

alternative strategy to distinguish between errors and real mutations, which from now on will be referred to as the "linkage method".

Figure 3 presents the main idea behind the method. For now, let us assume that only real variants and reverse transcriptase errors are present (no PCR1 errors). The goal is to determine the origin of the position different from the reference shown in red. There are two possible variants present at a nearby position (A and C), that share reads with our position of interest. Because a transcriptase error would only be introduced in one original transcript, it should always be present on reads with one version of the other allele (in other words, always with an A or a C). Therefore, if the error is introduced on a read with an A, it is not possible for it to be found on a read with a C. Because we can never see a full set of reads, even if we only see our variant present on reads with the A allele, it does not mean that our variant is never present on reads with a C. In this case, we are not able to determine the origin of our position. However, on the other hand, if it is present on the read with a C, it is an indication that it must have been "introduced" either at the same time as the "A/C" position or earlier – which basically means that it would be higher in the hierarchy, where homozygous SNVs are the highest, heterozygous are lower and RT errors are equal to or lower than the heterozygous SNVs (because they occur the latest). Therefore, if our position is present on reads with both A and C alleles, it is called as a real mutation.

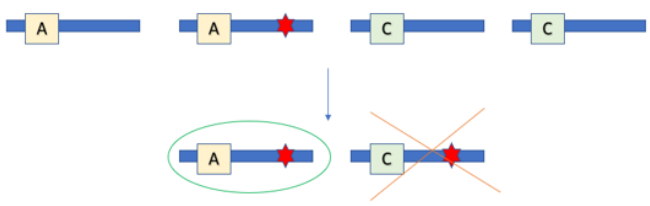


Figure 3. A visual representation of the linkage method.

An IGV screenshot in **Figure 4** presents a real-life application of the linkage method. Two variants share a number of reads. Variant on the right is present on all reads that the left variant is present on, but not vice versa. Therefore, the variant on the right is higher in the hierarchy than the left variant (i.e. appearing earlier in sample processing). Knowing that the only errors present are reverse transcriptase errors, the variant on the right must be a real mutation. Therefore, the variant on the right will be called and the variant on the left will be put aside. There are two scenarios for the variant on the left now. If no other positions different from the reference that share reads with the variant on the left exist, it is going to be discarded. This would be a correct decision if the variant was a reverse transcriptase error. However, it is also possible that the variant is a heterozygous mutation - and it in this case, it should not be removed. If another position exists that is evident to be lower in the hierarchy than the variant on the left, the variant on the left is going to be called as a heterozygous mutation (not shown).

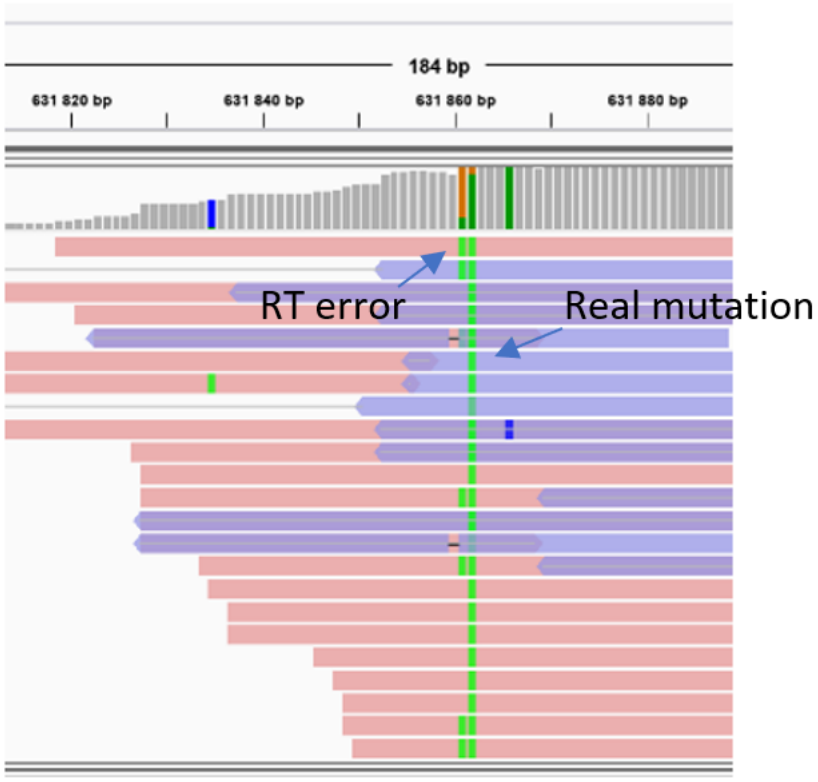


Figure 4. IGV screenshot presenting the linkage method in practice.

The problem is relatively simple if only reverse transcriptase errors and real mutations are present. The reconstruction of hierarchy, however, becomes much more difficult when errors from the first round of PCR are present as well. As shown in **Figure 5**, PCR1-RT pairs can be confused with RT-SNV pairs. In this case, when comparing a reverse transcriptase error to a PCR1 error instead of a reverse transcriptase error to a mutation, the reverse transcriptase error would be called a mutation.

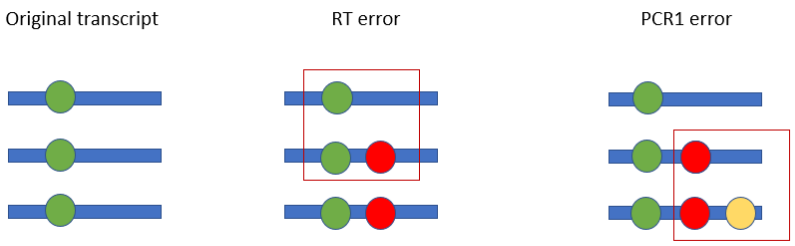


Figure 5. PCR1-RT and RT-SNV pairs can easily be confused.

1.6.4.1 The false positive rate of calls in the spike-in regions is low

We were wondering how often reverse transcriptase errors would be wrongly called as real mutations in the cases when PCR1-RT pairs were confused for RT-SNV pairs, and whether it was a significantly common phenomenon. In order to estimate that, we needed to investigate the frequency of PCR and reverse transcriptase errors occurring close to one another. As mentioned in the previous sections, the expected error rates for both of those errors was in the range of 1 per 10,000 bases. Considering that transcripts are only a few thousand bases long and usually only a few dozens of positions are considered (reads need to overlap in order for the linkage method to work), we suspected it was very rare for the PCR and RT errors to occur simultaneously, and therefore, for the RT error to be called a mutation.

To check whether our results agreed with that assumption, we looked at variants called from the spike-in regions. As we did not expect there to be any real mutations, all calls must have been reverse transcriptase errors that were misidentified as real mutations due to their proximity to PCR errors.

In total, there were 331,766 candidates in the spike-in regions that could be called as they were of good quality and located in regions with sufficient coverage. Out of those,

however, only 19,032 (5.7%) were paired and could be used as input for the linkage method. Even fewer were ultimately selected as variants by our caller, as presented in **Figure 6**.

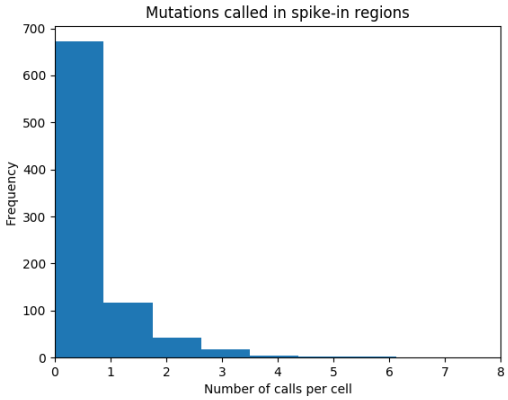


Figure 6. Histogram of calls made from the spike-in regions per cell.

Our method called 299 spike-in positions among 839 cells (only samples with SNV calls were considered), which gives a very low rate of 0.36 false positive calls per cell (within the spike-in region). Considering that there are around 300 times as many chromosomal positions with sufficient coverage as in the spike-in regions per cell, we could expect around 100 such cases per cell. However, the quality of positions (coverage and number of supporting reads) is higher in the spike-in than in the chromosomal regions. In effect, calls are made more often per region in the spike-ins, and therefore, it is not correct to use the frequency of false-positive calls from the spike-in regions to estimate their frequency in the chromosomes. We observed that the frequency of potential calls (PCR1 and reverse transcriptase errors plus real mutations) is over 4 times lower in the chromosomal regions than in the spike-ins (~$7*10^{-4}$ versus $3*10^{-3}$ potential variants per position covered by at least 3 reads, accordingly). Therefore, the number of expected PCR1-

RT error pairs in the chromosomal regions could be estimated not around 100, but rather closer to 25.

1.6.5 RNA editing sites

RNA editing is a molecular process that is one of the most evolutionarily conserved characteristics of RNAs. Occurring in all living organisms, it generates RNA and protein diversity by specific amino acid substitutions, deletions, and changes in gene expression levels (Li and Mason, 2014). Adenosine-to-inosine (A->G) substitutions represent the most important class of editing in humans, and are of particular interest during variant calling (Lo Giudice et al., 2020). The reason for that is that they occur at high allelic frequencies and can highly resemble the characteristics of real mutations.

There are a number of existing databases that contain information about known RNA editing sites, such as REDi and DARNED. Initially, we used them to filter out such sites post-calling. However, we realized that RNA editing sites had an influence on our calls as they were used during the linkage step, and therefore had to be removed before it. Additionally, we noticed that there still were numerous RNA editing-like variants left in our set of potential variants (A->G variants with high AF and occurring in many cells, most likely sites not present in the databases). Therefore, in the end we decided to remove all positions with A->G (if the transcript's orientation was "+") and T->C (if the transcript's orientation was "-") substitutions, regardless of whether they were in the databases or not. While this action decreased the number of mutations we could detect, it was preferable to calling RNA editing sites or using them to decide about other positions different from the reference.

1.6.6 Known variants can be used to identify others within the transcript

Successful identification of real mutations using the linkage method relies on two variants sharing the same reads. However, there are numerous positions that are not found at close proximity to other variants and which could potentially be real mutations. As explained in the previous sections, reverse transcriptase errors are not expected to have a higher allelic frequency than real mutations. Therefore, one could treat a mutation with the lowest allelic frequency identified using the linkage method as the allelic frequency threshold required to call other positions in the transcript. In other words, if we find a mutation with a particular allelic frequency, we would call all other variants within the transcript with allelic frequency above that.

1.6.7 Coverage and frequency thresholds exist above which all variants are real

As explained in the section describing reverse transcriptase errors, the higher the coverage, the lower the allelic frequency of the reverse transcriptase error. If one could determine the allelic frequency threshold of the reverse transcriptase for each coverage range, calling would be quite straightforward, as all variants with allelic frequency above the threshold would be real mutations. While this approach could work for regions with high coverage, it is not a reasonable solution for those with poorer, in our case most, regions.

However, we believed that there must be a reasonable combination of coverage and allelic frequency thresholds above which variants are real. In order to define them, we plotted allelic frequency histograms for different coverage ranges of SNPs (identified from

RNA bulk) present in single-cells versus all other variants present in those cells in the chromosomal regions. As shown in **Figure 7**, while the distributions are relatively flat in the coverage range of 10 to 20 reads, there is a clear increase of SNPs around the allelic frequency of 0.5 when more than 50 reads are present. Peaks around 0.5 and 1.0 AF are what we would expect to see in the distribution of SNPs, considering that heterozygous and homozygous mutations are present in the data. Furthermore, there is a large increase of variants with lower allelic frequencies, suggesting that this is where most errors are located.

We decided to use this case to determine our conservative threshold and, therefore, treated all variants with allelic frequency of at least 0.4 found in regions covered by more than 50 reads as real mutations.

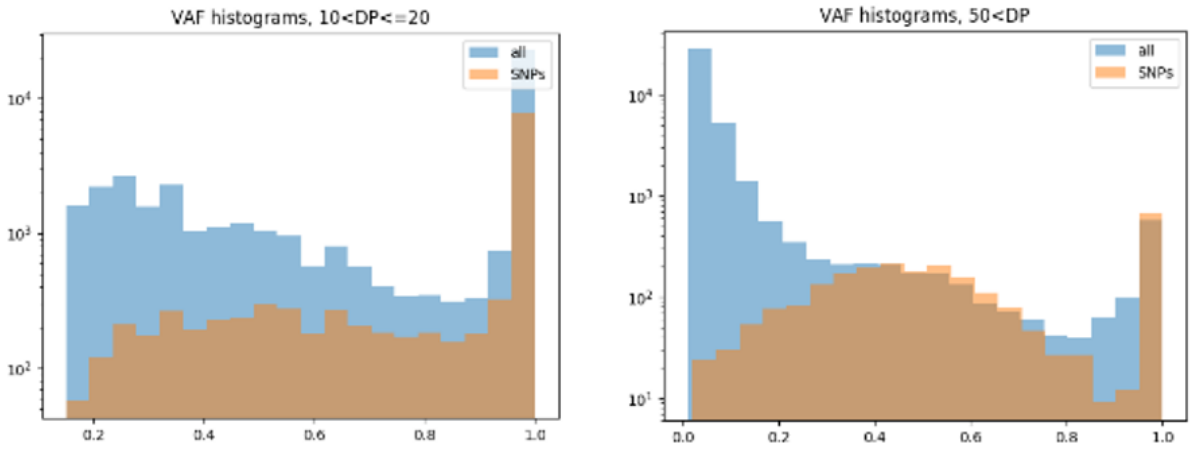


Figure 7. Variant allele frequencies of known SNPs versus other positions of interest for different coverage ranges.

1.6.8 Combining calls from multiple samples greatly improves results in individual single-cells

The linkage method, using known mutations to make calls from the same transcript or identifying "high-confidence" variants all rely on coverage, quality of reads around positions of interest and potential variants sharing the same reads. Those criteria

significantly decrease the number of calls we were able to make. This is why we decided to add the recalling step in the end of the calling.

The recalling step involves gathering all calls made from samples from the same cohort (we grouped the samples per patient) and checking whether there is any evidence for them in the samples in which they were not called. It would often happen that despite the fact that a call was made in only one cell, the same mutation was present in a number of other samples but not called for example due to lack of other variants sharing the same reads (the linkage method could not be applied). In order for a variant to be recalled, it still had to pass the standard quality criteria (minimal AF, coverage and number of supporting reads).

Figure 8 compares the numbers of calls per single-cell before and after recalling. After the first round of calling, the average number of calls per cell was 17. After the recalling, it increased to as many as 152. The results highlight the importance of the recalling step, and therefore the benefit of using information obtained from all samples in order to improve the calling in each individual cell.

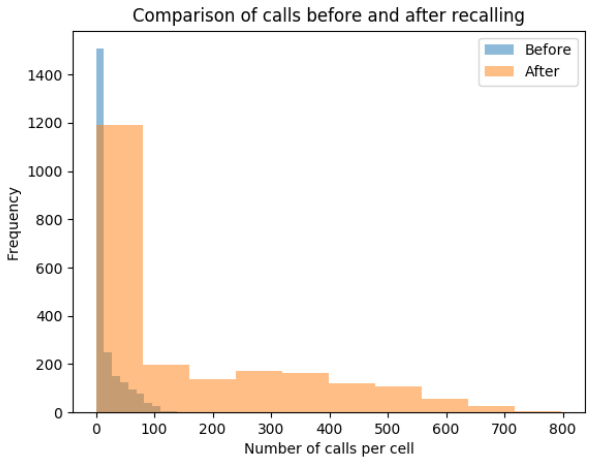


Figure 8. Distributions of the numbers of calls per single-cell before and after recalling.

1.7 Conclusions

In this chapter, we provided a detailed description of our single-cell RNAseq caller. We explained how we handled each type of technical errors and how we focused on relationships between them in order to select mutations with high confidence. Finally, we discussed additional measures we applied to increase the quality and number of our calls.

3. Validation of the single-cell RNAseq caller

1.8 Introduction

A comprehensive validation of a new technique is an essential final step to make the results reliable. However, proving that a method works is not always straightforward. This is the very issue we have been struggling with in our work. First of all, not many datasets exist that would satisfy all criteria required to perform variant calling from single-cells. It is common knowledge that unless data of sufficient quality is available, no trustworthy calls can be made. Because calling variants from single-cells is a novel approach, most single-cell data does not fulfil the required quality criteria. Secondly, in order to compare calls from single-cells to gold standard methods, paired DNA is required. Even if this is accessible, comparison to other methods is problematic in itself. It has long been known that there is poor consensus between commonly used variant callers (Cornish and Guda, 2015), and the degree of the overlap of called variants is highly dependent on the caller, or even aligner, used. Therefore, sole intersection of variant calls from single-cells with those made by gold standard callers from DNA or bulk RNAseq might not be sufficient to prove the method works. Alternatively, a method could be validated by proving that it successfully identifies phenomena that are known to be present in the data. In the case of cancer, those could include known mutation profiles or cancer driver genes. This approach would limit the need for additional tools, which are prone to errors.

We decided to start with the first method of validation. In the first section of the following chapter, we compare our single-cell RNAseq caller output to variants detected in bulk DNA and RNA sequencing data from the same patient. We conduct our analysis on a publicly available breast cancer dataset (described in subsequent sections). There are two reasons for that. Firstly, calling somatic variants from tumour and normal DNA is the most reliable way to identify cancer mutations - and DNA was not included in our Barrett's data. The breast cancer dataset was the only dataset that we could find that had whole-transcriptome single-cell RNAseq with a DNA-based match, and had spike-ins. Secondly, we believe that validation should be conducted on a different dataset than the method had been developed on.

We begin the following section with a comprehensive comparison of both germline and somatic calls made from whole exome sequencing (WES) and bulk RNAseq data using common tools. We then continue to the processing of the single-cell RNAseq samples with our method and evaluate our results with regards to the calls from gold standard tools. Apart from intersection statistics of our findings with the output of other callers, we provide a detailed analysis of the variants that our caller missed or the calls that were only made from single-cells. Finally, we search for known cancer driver genes and mutation profiles linked to breast cancer among our single-cell calls in order to expand our validation beyond what other variant callers are able to identify.

1.9 Data

The following section contains a description of the breast cancer dataset, which was used to validate the single-cell RNAseq caller.

1.9.1 Data

The breast cancer dataset consists of single-cell RNAseq data paired with bulk RNAseq and WES. The RNAseq data (single-cell and bulk) has been deposited in the NCBI Gene Expression Omnibus database under the accession code GSE75688, and the WES data can be found in the NCBI Sequence Read Archive under the accession code SRP067248. We chose it because apart from the aforementioned fact it was one of the few single-cell RNAseq datasets with paired WES, it had previously been shown to be of good quality (Chung et al., 2017).

1.9.2 Alignment to the reference genome

The following section contains information about the reference genomes used and describes alignment of the samples.

1.9.2.1 Reference genome

Two reference genomes were used in the analysis of the breast cancer dataset. Most of the analysis was conducted using the hg38 version of the human reference genome (already described along the Barrett's dataset). Somatic variant calling from bulk RNAseq was done using RNA-MuTect with the hg19 version of the human reference genome, as it was the only version that the software was compatible with. Published analysis of the breast cancer dataset was also performed using hg19 (Chung et al., 2017). In the cases where data was obtained in the hg19 format, it was lifted over to hg38 using the USCS *liftOver* tool.

1.9.2.2 Single-cell RNAseq

The raw FASTQ files were aligned using two different aligners: STAR 2.6.0c and Hisat 2.0.4. The parameters used with STAR *were --readFilesCommand zcat --runThreadN 4 --outSAMtype BAM SortedByCoordinate --outSAMmapqUnique 60*. The parameters used with Hisat2 were *--min-intronlen 20 --max-intronlen 500000 -k 5 -X 800*. In order to maximize the quality of the alignment, all unmapped or multi-mapped reads were removed. Furthermore, only reads with the maximum mapping quality (MAPQ = 60 and less than 5 mismatches) were used in the analysis.

1.9.2.3 Bulk RNAseq

The raw FASTQ files were aligned using STAR 2.4.2a. The parameters used were *--runThreadN 4 --outSAMtype BAM SortedByCoordinate --outSAMmapqUnique 60 --readFilesCommand zcat --outFilterType BySJout --outFilterMultimapNmax 20 --alignSJoverhangMin 8 --alignSJDBoverhangMin 1 --outFilterMismatchNmax 999 --outFilterMismatchNoverReadLmax 0.04 --alignIntronMin 20 --alignIntronMax 1000000 --alignMatesGapMax 1000000*.

1.9.2.4 WES

The raw FASTQ files were aligned using BWA 0.7.15 (*bwa mem*). We used *--M* parameter for compatibility with Picard, while all other settings were left as default.

1.9.3 Ground truth variant calling

We used gold standard tools to call germline and somatic variants from WES and bulk RNAseq data, in order to compare them to our calls from single-cells.

1.9.3.1 Methods

The following section contains information about how the ground truth variant calling was conducted.

1.9.3.1.1 SNP calling from WES

SNPs were called from WES using two callers - Haplotypecaller and Octopus. Calling variants with Haplotypecaller (default settings) involved the calling itself, preceded by pre-processing of the aligned BAM files according to the GATK 4.1.7.0 Best Practices. The final variants were defined as those that passed the additional GATK VariantFiltration (default settings) step. No pre-processing steps were applied before calling variants with Octopus, which was ran with the germline forest file (v0.6.3-beta) provided with the Octopus software. A final list of calls was created from an intersection of results from the two callers and restricted to the SureSelect All Exon V5 regions.

1.9.3.1.2 SNP calling from RNA bulk

The aligned BAM files were processed according to the GATK 4.1.7.0 Best Practices and the germline variants were called with Haplotypecaller (default settings). The final

variants were defined as those that passed the additional GATK Variant Quality Score Recalibration (default settings) step.

1.9.3.1.3 SNV calling from WES

SNVs were called from WES using two callers - Mutect2 and Octopus. Calling variants with Mutect2 (default settings) involved the calling itself, preceded by pre-processing of the aligned BAM files according to the GATK 4.1.7.0 Best Practices. The final variants were defined as those that passed the additional GATK VariantFiltration (default settings) step. No pre-processing steps were applied before calling variants with Octopus, which was ran with the germline (v0.6.3-beta) and somatic (v0.6.3-beta) forest files provided with the Octopus software. The additional parameter used was *--sequence-error-model PCR.HISEQ-2500*. A final list of calls was created from an intersection of results from the two callers and restricted to the SureSelect All Exon V5 regions.

1.9.3.2 Investigation of germline calls

There were two reasons for calling SNPs from the breast cancer dataset. Firstly, they gave us a good indication of what real mutations look like in our single-cells, in addition to less frequent SNVs. Secondly, because our single-cell RNAseq caller does not discriminate between somatic and germline variants, we needed a list of SNPs in order to remove them from our calls. Calling SNPs from both WES and RNAseq allowed us to identify more variants than if we only used either. Furthermore, it gave us an opportunity to explore the limitations of calling SNPs from each of those data types.

The following section begins with an evaluation of variants identified with different variant callers from WES. It then continues with a comprehensive comparison of SNP calls from WES and bulk RNAseq. This information is required not only to understand the limitations of each of the tools and data types, but also to have a good benchmark of the degree of overlap of gold standard callers.

1.9.3.2.1 High concordance in SNPs called from WES using different variant callers

As described in the previous sections, SNPs were called from WES using Haplotypecaller and Octopus. The output of both callers was concordant in terms of the numbers (**Figure 9a.**) and experienced a high degree of intersection of over 90% (**Figure 9b.**) for most patients. A poorer overlap for patients BC02, BC05 and BC06 resulted from a greater number of SNP calls made by Octopus, rather than from a disparity in results produced by the two callers.

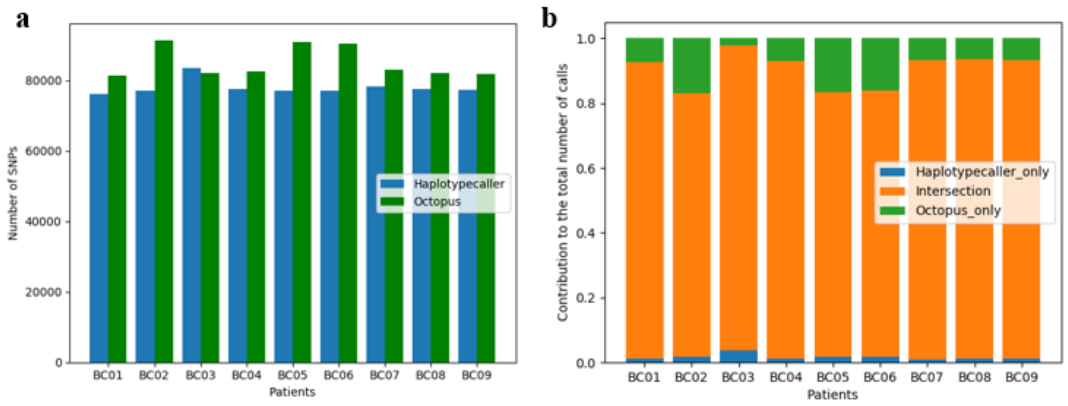


Figure 9. SNP calling from WES with Haplotypecaller and Octopus. a. Comparison of the numbers of SNPs from the two callers per patient. b. Intersection of the SNPs from the two callers per patient.

Closer examination of the reasons why some SNVs were only called by Octopus revealed that 23% of them were removed during the Haplotypecaller's Variant Quality Score Recalibration, which is a final filtering step that uses machine learning to eliminate probable artefacts. The remaining variants were either removed during the pre-processing steps such as marking duplicates and base quality score recalibration or during the main calling itself.

1.9.3.2.2 Agreement of SNP calls from WES and bulk RNAseq

Having obtained a satisfactory intersection of the WES SNP calls from different callers, which would from this time forth form the eventual list of SNP calls from WES, we progressed to its intersection with SNPs called from bulk RNAseq.

Before comparing SNP calls from WES and RNAseq, we needed to consider shared coverage. We expected RNAseq reads to span regions outside of the transcriptome, and this was something we could correct for by restricting coverage to the SureSelect regions. However, this would not solve the problem of the unexpressed DNA regions, not captured in RNAseq. Furthermore, raw coverage alone was not sufficient to compare the calls in a fair way, as they were also highly impacted by the presence of the SNPs in each data type in the first place, and an adequate number of supporting reads. Indeed, we discovered that only 6% of SNPs called from WES and missed in bulk RNAseq calling were covered in the bulk

RNAseq (**Figure 10**). Similarly, only 9.7% of SNPs called from bulk RNAseq and missed in the WES calling were covered in WES.

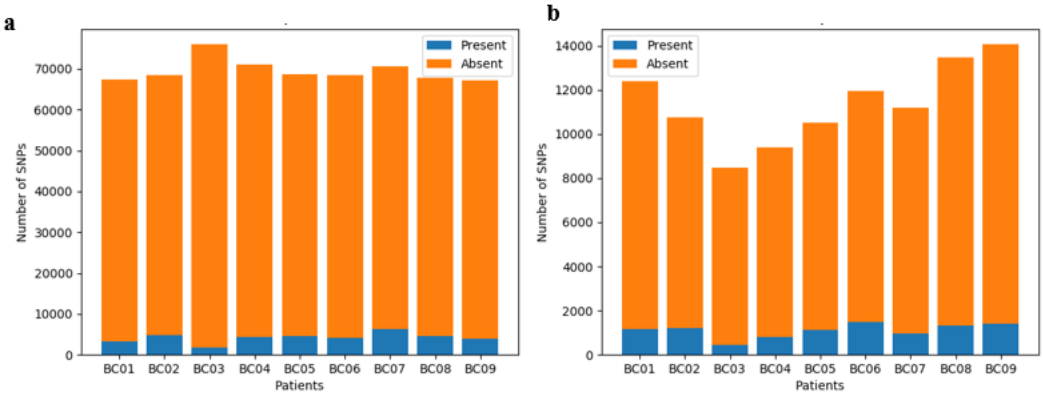


Figure 10. Presence of SNPs called from one data type but missed in the other. a. Presence of WES SNPs in bulk RNAseq. b. Presence of bulk RNAseq SNPs in WES.

Analysis of SNPs missed by either caller revealed that 43% of calls made from WES but missed in RNAseq were covered by less than 2 reads, and 50% by less than 3 reads in RNAseq (**Figure 11a**). Similarly, 56% of calls made from RNAseq but missed in WES were covered by less than 2 reads, and 73% by less than 3 reads in WES (**Figure 11b**). To focus on the effect of the callers, and not the difference in technology, we restricted to regions in which there was coverage of at least 3 reads in both RNAseq and WES. Furthermore, in order to ensure that variants had sufficient support, we restricted the comparison to variants covered by at least 3 reads (simultaneously matching our single-cell RNAseq calling requirements).

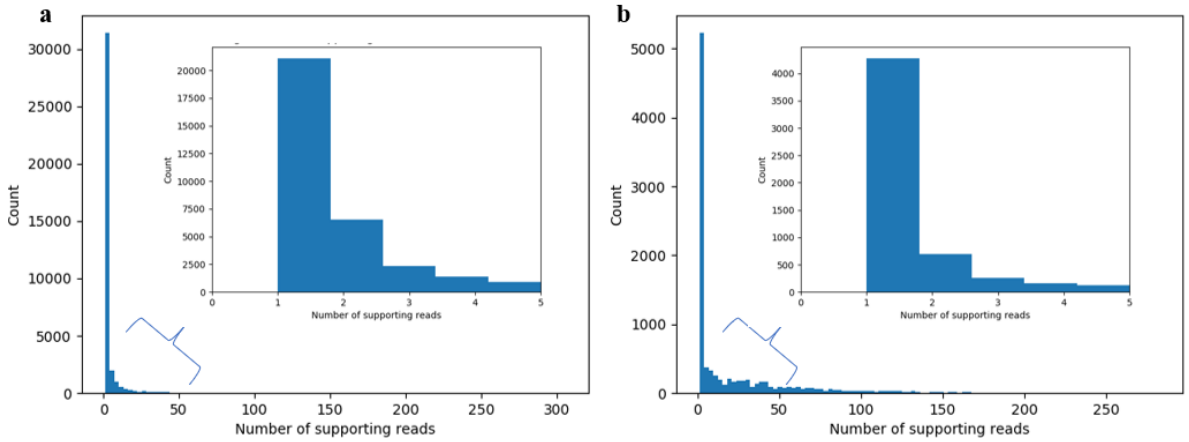


Figure 11. Number of reads supporting SNPs a. called from WES but missed in bulk RNAseq, b. called from bulk RNAseq but missed in WES.

The intersection of SNPs called from WES and bulk RNAseq after the initial coverage restriction (regions shared in both WES and RNAseq) contributed to 57.2% of all calls (**Figure 12a**). The value further increased to 80.9% when at least 3 reads supporting an SNP were required (**Figure 12b**).

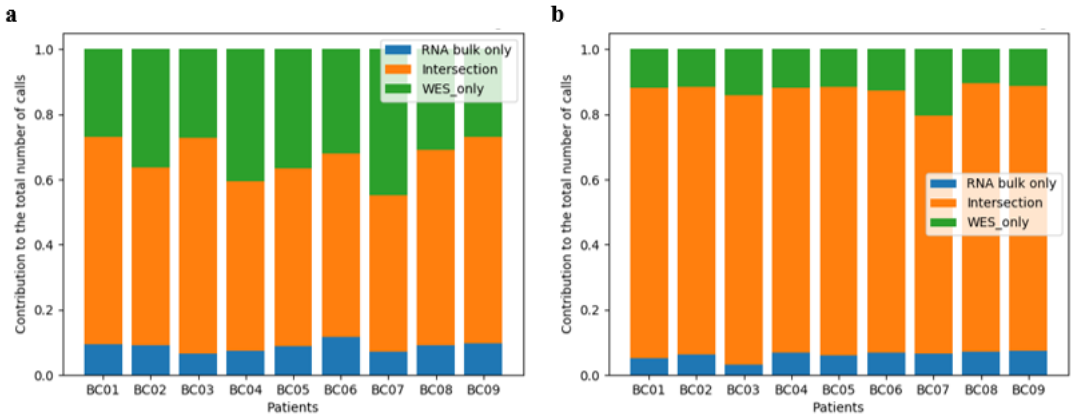


Figure 12. Comparison of SNPs called from WES and bulk RNAseq, restricted to regions only covered in both tissues (a.), with a further requirement for at least 3 supporting reads (b.).

Coverage alone was enough to explain most differences in the SNP calls from WES and bulk RNAseq. While additional aspects such as allelic frequency and quality of regions

would most likely be strong contributors as well, the results we obtained so far were sufficient to conclude that in regions with adequate quality, SNP calling from WES and bulk RNAseq was in relative agreement.

1.9.3.3 Investigation of somatic calls

Calling the same somatic mutations from single-cells as those identified using common tools would provide a sufficient validation of our method. Identification of SNVs from RNAseq is not yet a well-established method and tools such as RNA-MuTect rely mainly on identification of macroscopic clones rather than all somatic mutations. Therefore, we decided to treat only variants identified from WES as our benchmark list.

In the following section, we provide a comparison of calls from WES using Octopus and Mutect2.

1.9.3.3.1 Overlap of somatic calls made by Mutect2 and Octopus

In order to estimate the consensus of Mutect2 and Octopus in calling somatic variants from WES, we compared the number and the intersection of SNVs per patient (**Figure 13.**). The total number of SNV calls was similar between both callers, with slightly more calls being made by Octopus in general. However, this was not the case for patients BC03 and BC04, which not only featured a greater number of calls made by Mutect2 (**Figure 13a**), but also a poorer intersection of calls from both callers (**Figure 13b**). We investigate reasons for the differences between SNV calls made by Mutect2 and Octopus in the Supplement (Supplementary section 1.14).

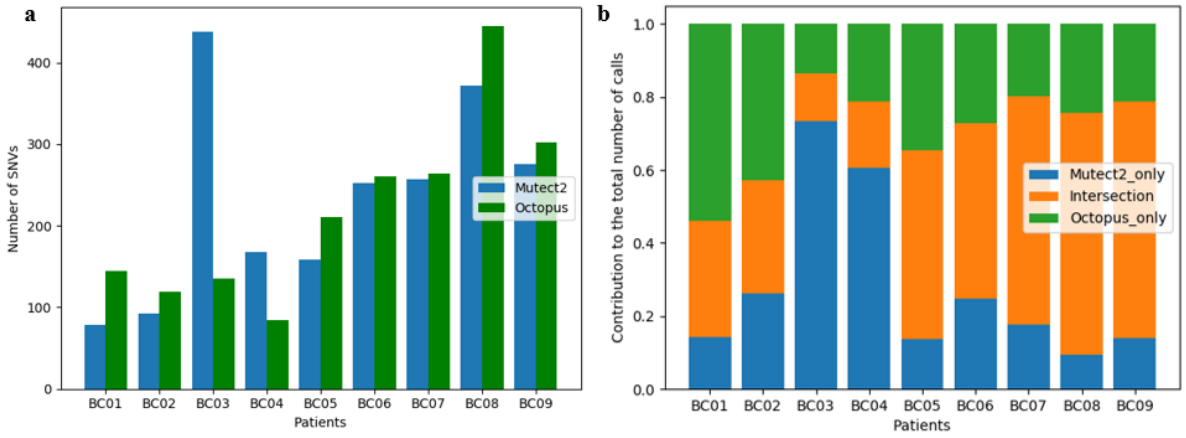


Figure 13. SNV calling from WES with Mutect2 and Octopus. a. Comparison of the numbers of SNVs from the two callers per patient. b. Intersection of the SNVs from the two callers per patient.

1.9.4 Conclusions

In the previous section, we compared germline and somatic mutations identified from WES and bulk RNAseq samples. We concluded that in regions with sufficient quality, the (germline) calls from the two data types were concordant. The results made us confident about the reliability of the consensus of the callers in the context of validating our calls from single-cell RNAseq.

1.10 Single-cell RNAseq variant calling

The following section begins with an analysis of mutations identified by our single-cell RNAseq caller, and follows with a comprehensive validation of our method.

1.10.1 Most SNPs called from WES and bulk RNAseq successfully identified in single-cells

We used our single-cell RNAseq caller to identify variants from the breast cancer single-cells (**Figure 14**). A large number of calls per cell was substantially reduced after removing germline variants previously identified from bulk RNAseq and WES. The size of the overlap was an indication that our method was successfully calling real mutations, and we considered that as the first indication that our tool is able to identify real genetic variants.

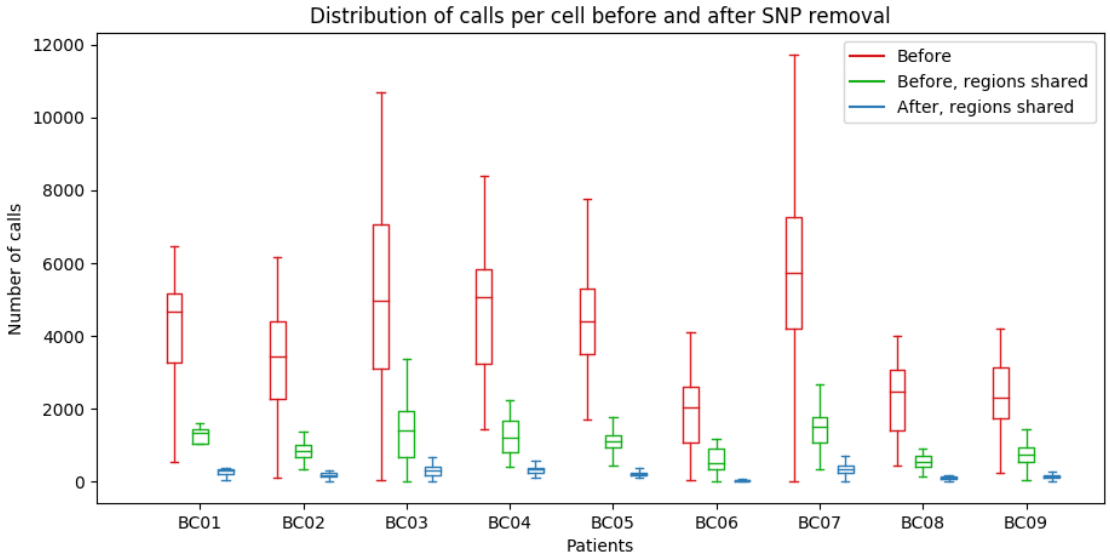


Figure 14. Distribution of calls per breast cancer single-cell before SNP removal (showing separately all calls per cell [red] and calls in regions covered by at least 3 reads in WES [green]) and after SNP removal, in regions covered by at least 3 reads in WES [blue].

Analysis of SNPs identified from bulk RNAseq and WES but missed by our tool provided an opportunity to explore the limitations of our calling. On average, our caller identified 6.45 +- 3.62% of SNPs from the consensus WES and bulk RNAseq calling. While the number might appear low, it was due to coverage as 74.55 +- 10.79% of the benchmark SNPs were not detectable in the single-cells at all. Out of the approximately 25% that were present in the single-cell data but missed by the caller, only 0.64 +- 0.16% were considered

during the calling. The reason for that was that they had not passed the necessary quality criteria (coverage, allelic frequency) or were detected in the alignment from only one mapper.

As coverage was the main factor limiting the potential of calling variants from single-cell RNAseq, estimation of positions missed by our caller required restriction to only positions with sufficient quality. The intersection indicated that the single-cell RNAseq caller only missed 2% of SNPs identified from WES and bulk RNAseq (**Figure 15**). Those were not called due to lack of suitable neighbours during the application of the linkage method (none of the missed SNPs passed the main calling stage, despite passing the necessary quality criteria).

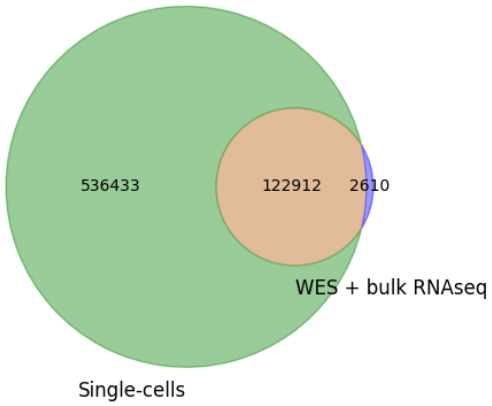


Figure 15. Intersection of variants called from single-cell RNAseq and SNPs from WES + bulk RNAseq.

1.10.2 Over a third of single-cell SNV calls shared between multiple cells

The statistics of single-cell RNAseq calls per patient after removal of SNPs are presented in **Table 8**. On average, the single-cell RNAseq caller identified 2,272 mutations per cell. 38% of the calls were shared between at least two single-cells. Because the calling was done independently for every single-cell, we treated variants identified in multiple cells as likely real SNPs or SNVs. The fact that over a third of our de-novo calls were shared between single-cells could be treated as a further validation of the caller.

Table 8. Statistics of single-cell RNAseq calls per patient after removal of SNPs.

Patient	Number of cells	Mean number of calls per cell	Median	STD
BC01	26	2089	2543	1139
BC02	56	1765	1833	778
BC03	92	2678	2591	1270
BC04	59	2636	2778	959
BC05	77	2362	2326	684
BC06	25	587	601	334
BC07	104	3173	3221	1255
BC08	23	1228	1295	527
BC09	60	1267	1204	608

The substantial number of calls shared between multiple cells made us wonder about the number of SNPs still remaining among the calls.

1.10.3 Calls shared between different cell types identified as unfiltered SNPs

We used cell type labels from the original publication to group the single-cells into three groups: tumour, stromal and immune cells. We noticed that the absolute numbers of variants called per single-cell differed per cell type, with the tumour cells having the most calls (**Figure 16**). Unfortunately, we could only investigate that in patients BC01-8, as the single-cell set from patient BC09 did not contain any tumour cells (or at least, no cells were identified as tumour).

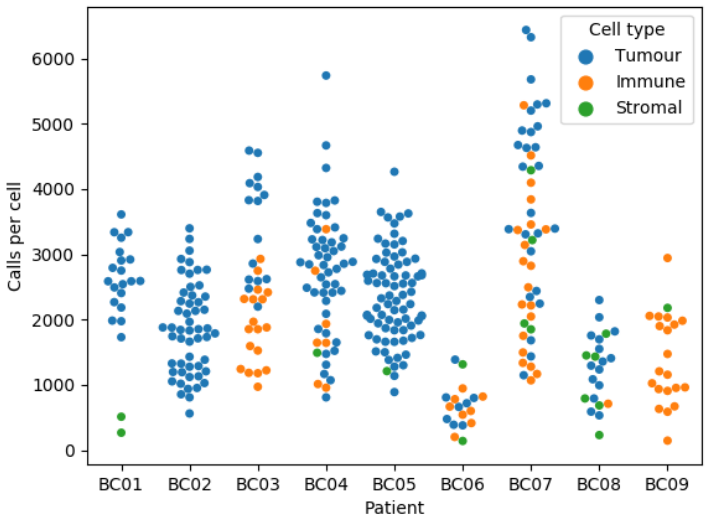


Figure 16. Swarm plots of calls per single-cell, separated by cell type and patient.

In order to search for the yet unfiltered SNPs among our single-cell calls, we intersected calls from different cell types (**Figure 17**). Our assumption was that any variant shared between multiple cell types would be germline. We found that 67% of calls were unique to tumour cells, while 22% were unique to either of the other cell types. Therefore, the remaining 11%, shared by at least 2 cell types were the unfiltered SNPs still present

among our calls. The fraction of shared calls was greater for patients in which more cells from each type were present (for example, 18% for patient BC03 with a 15:18 split of tumour to immune cells versus 1.6% for patient BC05 with a 75:1 split of tumour to stromal cells).

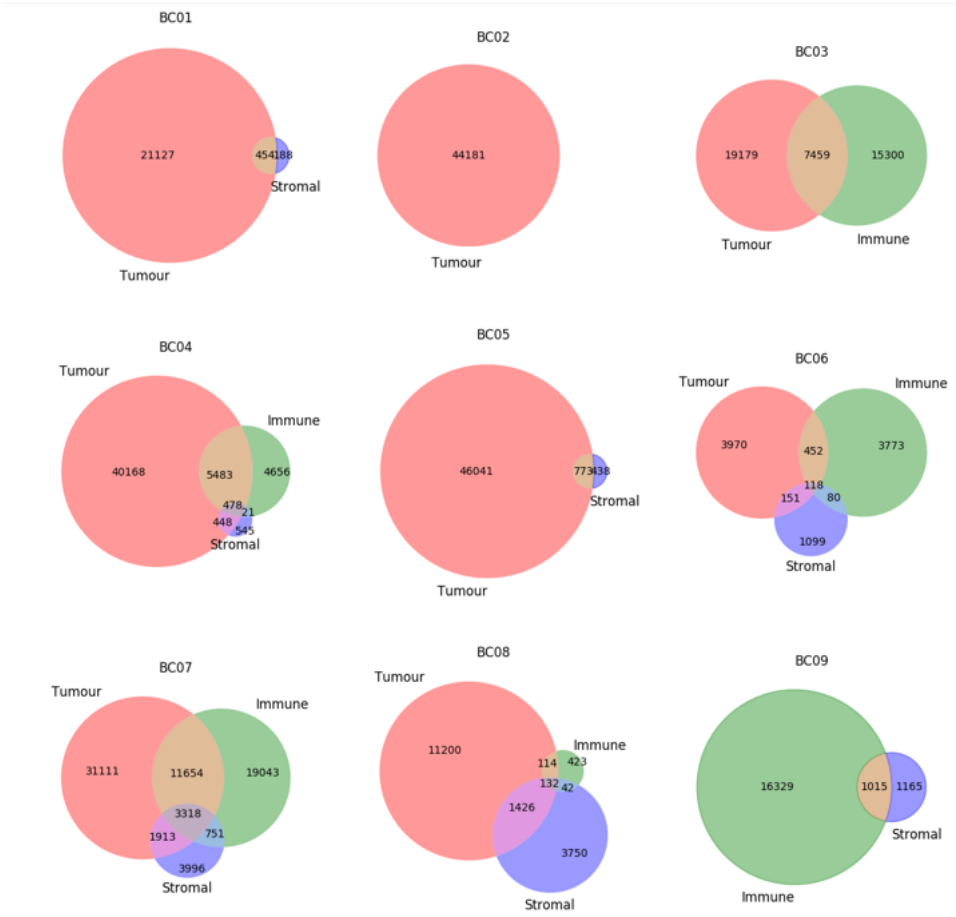


Figure 17. Intersection of calls from tumour, immune and stromal cells.

The increased contribution of shared calls in patients with a greater number of non-tumour single-cells indicated a high likelihood of the further rise in the fraction of shared calls if more non-tumour single-cells were present. In the case of patients like BC06, the final set of SNVs was substantially smaller than for BC02 with no non-tumour cells. Therefore, our ability to detect unfiltered SNPs was reliant on the presence of non-tumour

single-cells - the more single-cells of various types available, the better the ability to remove SNPs and the higher the accuracy of the final SNV set.

1.10.4 Cancer Signature 3 identified in tumour, and not in stromal, cells

COSMIC, the Catalogue Of Somatic Mutations In Cancer (Tate et al., 2019), is currently the most comprehensive resource for exploring the functional effects of somatic mutations in cancer. Mutation signatures, included in the catalogue, characterize different processes active throughout cancer development. Apart from the signatures, the catalogue also contains information about proposed aetiology, tissue distribution of each signature and potential connections with other signatures.

Breast cancer has not only been well characterized in terms of gene mutations, but also in terms of mutation signatures. **Figure 18** presents the most common signatures in breast cancer. Those include Signatures 1,2,3,5,8 and 13. Other signatures are also present (Nik-Zainal and Morganella, 2017), but less frequently (in less than 10% of samples).

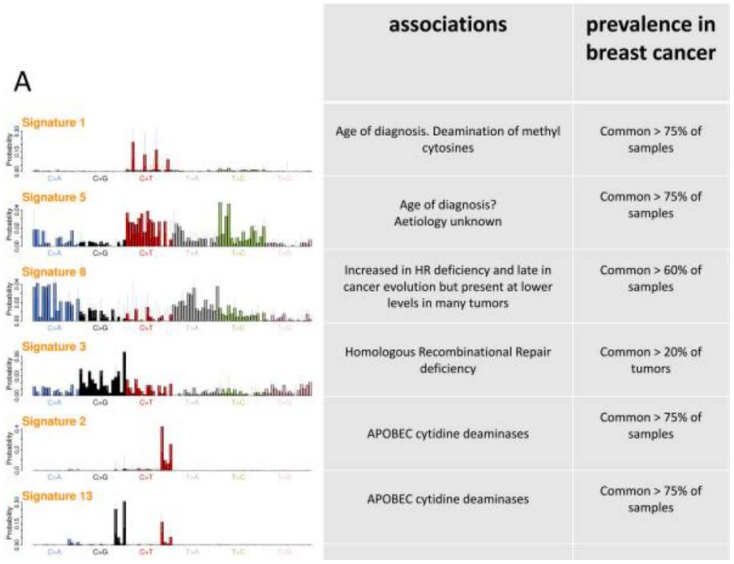


Figure 18. The most common mutation signatures in breast cancer (Nik-Zainal and Morganella, 2017).

While the overlap of the listed signatures with those identified by us from the single-cell data would be a reasonable further validation of our method, we were primarily focused on investigating differences between tumour and non-tumour cells. Identifying such signatures within the disease tissues and not in the healthy cells would prove that our caller had the ability to identify the crucial intracellular mechanisms correctly.

In order to obtain the mutation signatures in the breast cancer dataset, we grouped all SNVs identified from single-cells per patient and cell type (keeping only cell type-specific mutations). We then calculated the frequency of each mutation type, as shown in **Figure 19.**

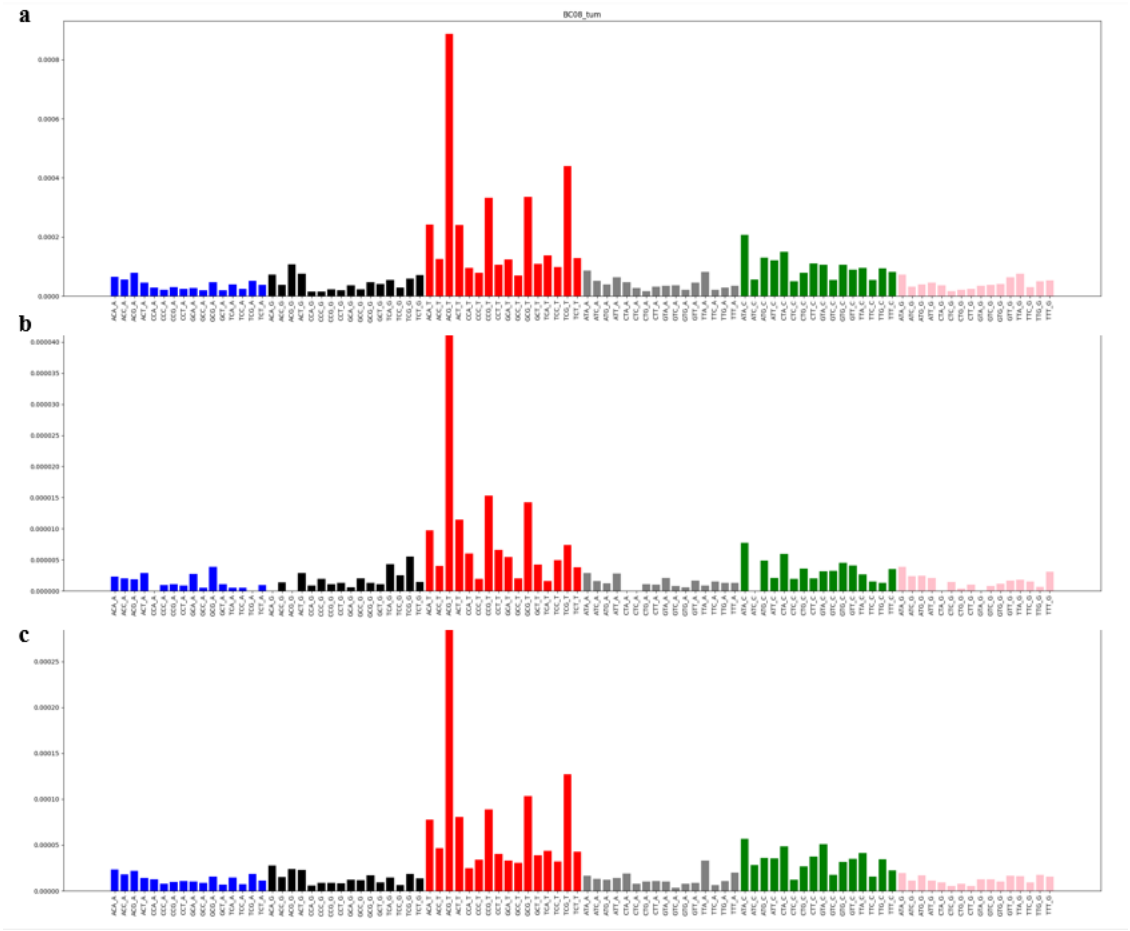
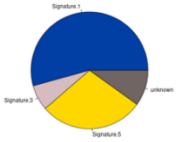
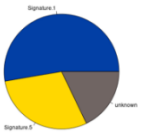
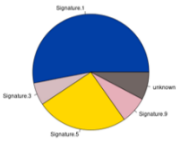
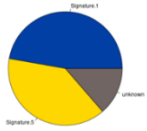
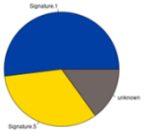
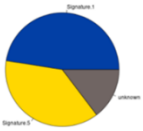
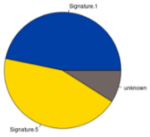
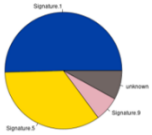
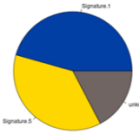
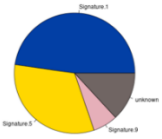
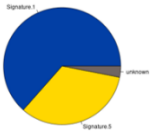
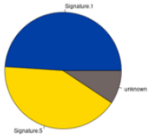


Figure 19. Sample mutation frequency plot of the SNVs identified from the breast cancer single-cells, grouped by patient (BC08) and cell type (a. tumour, b. immune, c. stromal).

There is a range of software available to identify mutation signatures from pre-calculated mutation profiles. In our study, we used an R package *deconstructSigs* (Rosenthal et al., 2016) due to its popularity in recent years. It uses a multiple linear regression model to determine the linear combination of mutation signatures and accurately reconstructs the mutation profile of a sample. We calculated mutation profiles for each patient and cell type separately, as presented in Table 9.

Table 9. Mutation signatures reconstructed from the single-cell somatic calls (cell type-specific). The signatures include Signature 1 (blue), Signature 3 (beige), Signature 5 (yellow), Signature 9 (rose) and unknown (gray).

Patient	Tumour cells	Stromal cells	Immune cells
BC01			
BC02			
BC03			
BC04			
BC05			
BC06			
BC07			
BC08			
BC09			

We identified 3 out of 6 most common breast cancer signatures (1,3 and 5). In addition to that, Signature 9 was detected. While less frequent, it had been identified in breast cancer samples before (**Figure 20).** The unknown signatures contributed to roughly 15% of the profiles. While Signatures 1 and 5 were detected in all cell types, Signature 9 was not discovered in stromal cells and Signature 3 was only found in tumour cells (with the exception of patient BC09 for whom no tumour cells were identified). The fact that no tumour cells were present in the data from patient BC09, and that the immune cells carried characteristics of tumour cells, made us aware of potential issues with the cell type assortment, and therefore potentially even greater differences between the cell types in reality.

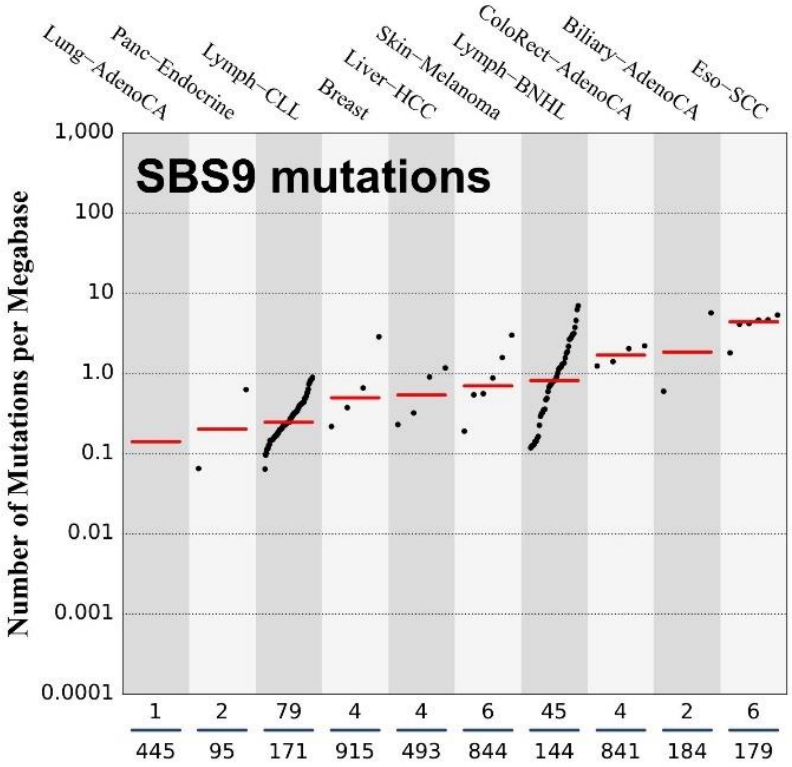


Figure 20. Frequency of Signature 9 in different cancer types. Source: COSMIC.

1.10.5 WES SNVs identified in all cell types

The intersection of our calls from single-cells with the SNVs identified from WES would provide a solid validation of our method. The following section contains information about the concordance of WES SNVs and the tumour-specific calls detected from our single-cells (all calls minus calls found in either stromal or immune cells), and an analysis of the calls that were either uniquely called from single-cells or missed with respect to WES.

Analysis of the single-cell RNAseq pre-processing output revealed that over 80% of the SNVs called from WES but missed in single-cells did not pass the quality criteria, and were therefore not considered during the calling. Out of those, 96.9% were not detectable in single-cells at all, while 2.1% had insufficient read support and 1% had low allelic frequency or either mapping or base qualities below required thresholds. Therefore, only around 20% of WES calls could in fact be called from single-cells.

Similarly, 82% of variants called from single-cells were not detected in WES at all. Among the detectable alleles, 74% did not pass the quality thresholds that we set for single-cells. Therefore, less than 5% of SNVs identified solely from single-cells had reasonable read and coverage support in WES. We believe the actual values could be even lower as our quality thresholds might not be strict enough for bulk samples (caller-dependent). Furthermore, the 5% of callable variants still included variants with low base and mapping qualities, and potentially only had sufficient support until duplicate reads were removed. We would normally eliminate such positions as well but because such thresholds are caller-specific, we decided to keep those positions at this stage.

The number of WES SNVs and the fraction called in tumour single-cells, restricted to variants only present in both data types and passing the quality criteria, is presented in

Table 10. No WES SNVs of sufficient quality were found in tumour single-cells from patients BC06-BC09. In total, 11.5% of WES SNVs were detected in the tumour single-cells. Interestingly, the percentage of detected WES SNVs would increase to 23.5% if all cell types were considered. There were a few potential reasons that we could think of that could explain this phenomenon. Firstly, the mislabeling of SNPs as SNVs. This is a scenario that we had seen before, which would explain the presence of variants in different cell types. Secondly, the wrong assignment of cells into cell types. Considering how similar immune and tumour cells were (what was not only evident in our mutation signatures but also repeatedly mentioned in the original publication), we did not exclude this possibility. While we were not able to prove any of those hypotheses directly (apart from proving that some variants were called as both germline and somatic and indicating a similarity between tumour and immune cells), we had the capability to investigate the reasons for removing the WES SNVs during the single-cell calling using our method.

Table 10. WES SNVs called from tumour single-cells, restricted to variants present in both WES and single-cells, in regions covered by at least 3 reads and with at least 3 supporting reads.

Patient	Number of WES SNVs	% of WES SNVs called from tumour single-cells
BC01	8	0
BC02	7	0
BC03	8	0
BC04	6	50
BC05	25	16
BC06	0	0
BC07	0	0
BC08	0	0
BC09	0	0

1.10.5.1 Lack of suitable pairings for the linkage method as the main reason for missing WES SNVs during single-cell variant calling

The quality analysis of variants called uniquely in WES but missed by our caller (all cell types considered) revealed that as much as 54% did not pass the main calling due to lack of suitable pairings for the linkage method (**Figure 21**. Reasons for removal of WES SNVs during the calling of variants from single-cells). A further 7% of calls were removed during the final filtering, which included removal of RNA editing sites and variants from outside the transcriptome region. The intersection of outputs from two aligners (STAR and Hisat2) resulted in the elimination of the subsequent 13% of WES SNVs. The remaining 26% of variants were removed during the exclusion of SNPs called from bulk RNAseq and WES. This suggested that somatic variants called from WES must have been simultaneously called as SNPs by germline callers. This was not surprising, as we had already described such a case when analyzing differences between SNV calls made by Mutect2 and Octopus. Therefore, if it had not been for the 33 WES SNVs filtered out during the single-cell variant calling as a result of being called as SNPs from bulk RNAseq, our caller would have successfully identified 29.3% of WES calls from single-cells.

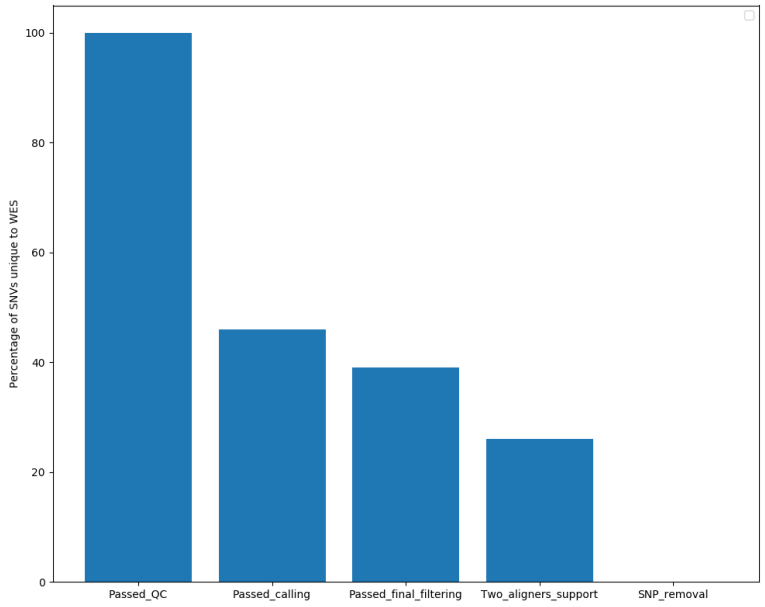


Figure 21. Reasons for removal of WES SNVs during the calling of variants from single-cells

1.10.5.2 Most single-cell SNVs removed from the callable WES positions during base recalibration

Because less than 1% of the single-cell SNVs were called from WES, we wanted to investigate the reasons for missing the single-cell calls by the WES callers as well. As above, we restricted our analysis to single-cell variants that were supported by at least 3 reads in WES.

Due to the fact that the gold standard list of SNVs called from WES was an intersection of Mutect2 and Octopus, we suspected that the overlap with our calls would be greater if the calls from only one caller were considered as well. In order to confirm this, we intersected the single-cell variants filtered out during the SNV calling from WES with outputs of Mutect2 and Octopus. We found that only 16 were identified solely from Mutect2 and 6 from Octopus. Those values were much lower than we had been expecting, therefore,

we continued to investigate the reasons for removing the single-cell calls by the WES callers.

We chose Mutect2 as our benchmark because it is not only the most popular caller, but also it provides filter tags for the most likely, but eliminated, variant candidates. Furthermore, we selected it because it consists of a number of steps, which allow for tracking of variants after every checkpoint of the pipeline. The first steps of the pre-processing pipeline were marking duplicates and base recalibration (BQSR), which relies on machine learning to correct over-optimistic base quality scores by fitting them to an expected distribution. Out of single-cell SNVs present but missed by Mutect2 in WES from patient BC05, only 54.3% passed base recalibration. The number of detectable SNVs decreased further by 72% after the second round of recalibration (Apply BQSR), therefore, only 16% of the SNV variants were available for Mutect2 to call for patient BC05. Considering all patients, less than 5% (none for patient BC05) passed the main Mutect2 calling step. While we were not able to investigate why the variants were eliminated at this stage, we could explore the reasons for filtering out the variants using the filter tags that variants eliminated in the last filtering step were supplied with.

We had been expecting that single-cell calls still contained SNPs, which was indeed confirmed by the “normal artifact” Mutect2 filter given to almost a third of the calls that passed Mutect2 (27.5%). The other filters included strand bias (22.5%), clustered events (20.0%), Panel of Normals (10%), haplotype (10%), weak evidence (7.5%) and base quality (2.5%). We found those filters to be reasonable, as they indicated aspects that we were not able to manage from the level of single-cells. For example, strand bias is a tag given to variants in which evidence for the alternative allele consists solely of reads from one strand.

While our caller does remove variants present on one strand only, the low count of reads in single-cells does not give us statistical power to confirm strand bias if the majority of the alleles are on one strand (or at least not for most positions, as the coverage is insufficient to apply any statistical tests). When it comes to the "clustered events" filter, it is applied when two mutation candidates are close to one another. Because single-cells cover only a small fraction of the genome, it is difficult to estimate the frequency of mutations accurately, and especially to determine whether any variants are statistically too close to be real. Panel of Normals is a list of common artefactual or germline sites. It is a part of the inner Mutect2 pipeline, and takes a form of a set of positions regularly called when using pairs of (unrelated) tumour-free samples as if they were tumour-normal pairs. Positions that were marked as variants repeatedly in different pairs allow for the identification of common variant calling artefacts. However, construction of a Panel of Normals would be difficult in the case of our data, as it does not contain any tumour-normal pairs. Finally, the remaining "haplotype", "weak evidence" and "base quality" filters are strictly linked to the quality of the data, which is heavily compromised in single-cells.

1.10.5.3 Conclusions

The intersection of single-cell RNAseq calls and variants identified by bulk callers revealed that our caller had the capacity to identify nearly 30% of the WES SNVs. The main reason for missing the remaining WES SNVs was insufficient single-cell data quality. On the other hand, an excess of variants called from single-cells in comparison to WES was explained by additional information that could not be obtained from single-cells alone. In

conclusion, we showed that in regions of sufficient quality and coverage, our single-cell RNAseq caller was capable of calling mutations with reasonable accuracy.

1.10.6 Majority of our calls shared with Red Panda

We used Red Panda, currently the most advanced of the few single-cell-specific variant callers available, as a direct benchmark for comparison of our calls. Because we were particularly interested in the number of unique tumour calls per patient and the number of WES SNVs detected by each tool, we restricted our analysis to calls from tumour cells only, and removed all variants called in the other cell types from the same patient.

The two methods have slightly different approaches – while our caller is tuned towards producing high-confidence calls, Red Panda heavily relies on the output of the GATK Haplotypecaller to validate its decisions. The creators of Red Panda show that the method is able to efficiently call "homozygous-looking" variants in comparison to other tools. However, they admit that the majority of calls are "heterozygous-looking" and because their fate is mainly determined by the Haplotypecaller, the number of false positives among the final set of calls is significant (Cornish et al., 2020). A large number of calls per cell output by Red Panda, most of which not shared between multiple cells, was something we observed when we intersected them with our calls (**Figure 22**). While 74.5% of our calls were shared with Red Panda, 89.3% of all called variants were unique to the latter.

Figure 22. Overlap of our and Red Panda calls, restricted to tumour-specific calls in tumour cells.

A comparison of our and Red Panda calls to WES revealed that both tools correctly identified 28 SNVs, with Red Panda additionally calling 69 variants. There were no SNVs that our caller would call that would not be found among the Red Panda calls. Therefore, it was Red Panda that correctly identified more WES SNVs than our caller. However, when taking into account the number of calls produced (and potentially large numbers of false positives present especially among Red Panda calls), the actual concordance of calls with

WES was higher in the case of our caller (0.012% of calls shared with WES versus 0.005% of those output by Red Panda). We considered this as a reasonable result – firstly, we obtained a solid intersection of calls with those produced by Red Panda without using the aid of the other tools such as Haplotypecaller. Secondly, we achieved our goal of maximizing the number of real variants while limiting the amount of false positives, which still are a difficult problem even for the most advanced single-cell RNAseq callers such as Red Panda.

1.10.7 Known breast cancer genes among single-cell calls

To validate the results of our caller, we decided to move beyond the comparison to mutations discovered by other tools from the same dataset. Specifically, we were interested to see whether our method identified mutations in genes known to be commonly mutated in breast cancer. In order to do this, we annotated the variants called from single-cells with the Variant Effect Predictor tool (McLaren et al., 2016). We then searched for the common breast cancer genes from *breastcancer.org.*

Interestingly, our caller did not identify any mutations in the BRCA1 and BRCA2 genes, which account for most of the inherited cases of breast cancer. However, we did discover at least one point mutation in all other genes mentioned. Specifically, those included high risk gene mutations (PALB2, PTEN, TP53), moderate to high risk gene mutations (ATM, CDH1), moderate risk gene mutations (CHEK2, NBN, NF1, STK11) and even some mutations with an uncertain breast cancer risk (BRIP1, MLH1, MSH2, RAD51C and RAD51D).

1.10.8 Conclusions and Discussion

In conclusion, we showed that the disagreement of bulk RNAseq and WES callers resulted mainly from the lack of or poor quality of variants in either tissue. The situation was similar when comparing the ground truth calls to those identified using our method from single-cells. However, additional limitations are evident when calling variants from single-cells alone. For example, due to low coverage, estimation of strand bias or sufficient read support is highly problematic. Furthermore, in order to ensure real mutations are identified, our caller heavily relies on their proximity to other positions differing from the reference (other mutations or technical errors) during the linkage method. Therefore, even if a variant is of very good quality and occurs at high frequency, it will not be called unless the linkage method confirms that it had been introduced at the very beginning of the sample preparation. This was very evident when comparing our calls for Red Panda, which identified more WES SNVs, but also appeared to have produced substantial amounts of false positives at the same time.

Despite the numerous obstacles related to the identification of variants from single-cells, our called still managed to identify nearly 30% of the gold standard calls from WES and provided reasons for not calling the remaining ones. In addition to that, our calls were highly concordant with the output of Red Panda, and a higher fraction of our, and not Red Panda's, calls overlapped with WES. Furthermore, we found cancer mutation profiles in tumour and not in stromal cells, and provided evidence for mutations in known cancer driver genes. We believed that, despite the evident limitations of working with single-cells, our results were sufficient to validate our caller, therefore, we decided to progress to the application of the method to our Barrett's dataset.

4. Application of the single-cell RNAseq caller

1.11 Introduction

1.11.1 Recap

In the previous chapters of the Thesis, we described the development of our single-cell RNAseq caller. The tool consists of three steps: pre-processing, the main calling stage, and final filtering. The pre-processing involves a removal of problematic reads, including multi-mappers, reads with numerous mismatches, or PCR duplicates, and elimination of positions with insufficient quality, such as regions with poor coverage or not supported by an acceptable number of reads. This step removes most technical errors present in single-cell data, such as sequencing and mapping errors, and errors occurring as a result of PCR amplification (or the second round of amplification, if two rounds are performed). Once the quality of callable regions in ensured, the linkage method is applied. It is a step which reconstructs the hierarchy of positions different from the reference, based on their relative positions on the reads, to distinguish the real mutations from the remaining reverse transcriptase errors and those introduced during the first PCR amplification round. Once only the strongest candidates remain, final filtering is performed. It includes the removal of RNA editing sites and positions outside of the transcriptome region.

We validated our method against a breast cancer dataset with matched exome-seq and scRNAseq data. We showed that not only our caller was able to identify mutations called

from WES, but it also identified mutations in known cancer driver genes and allowed for a successful reconstruction of the breast cancer mutation signatures.

The following chapter describes the application of the single-cell RNAseq caller to the Barrett's dataset. In the first section of the chapter, we introduce the Barrett's oesophagus in detail and elaborate on the unknowns of the origins and the mechanisms of the disease. We then present the general results of the single-cell RNAseq variant calling. Finally, we show how we used the calls to address the outstanding questions about Barrett's.

1.11.2 Barrett's Oesophagus

Intestinal metaplasia is the ectopic growth of intestine-like tissue in the oesophagus or stomach and has been commonly linked to the development of oesophageal and gastric adenocarcinomas (Xian et al., 2019). Its occurrence is associated with the acid reflux disease that leaves the oesophagus in the state of chronic inflammation - oesophageal adenocarcinoma - (Souza, 2016) or *H.pylori* infections in the case of gastric adenocarcinoma (Wroblewski et al., 2010).

Intestinal metaplasia at the distal oesophagus was initially observed and characterized in 1950, and termed Barrett's oesophagus after the author (Barrett, 1950). Barrett's oesophagus is a common condition, found in approximately 10-15% of patents with gastroesophageal reflux disease (Ouatu–Lascar et al., 199)). It is not only a serious disease in itself, manifesting reflux symptoms a couple of times per week (Zagari et al., 2008), but also carrying a 30-fold increased risk of oesophageal adenocarcinoma (Haggitt, 1994). The 5-year overall survival rate for this type of cancer is less than 20% (Pohl et al., 2010).

There is a strong motivation to improve current treatment or develop novel therapeutic strategies in order to improve the perspectives for oesophageal cancer patients. A preferred strategy aims to prevent the initial development of the adenocarcinoma in the first place, however, better understanding of its pathogenesis is still required. Barrett's oesophagus is the only known precursor of the tumour, and once Barrett's is established, increased stress from chronic exposure to acidic bile salt can lead to dysplasia and eventually to cancer (Rhee and Wang, 2018).

There have been numerous studies aiming to define how exactly benign tissues progress to malignant tumours. Agrawal et al., 2012 performed the first genome-wide study of mutations in oesophageal cancer, which consisted of exome sequencing data from two matched Barrett's oesophagus and oesophageal adenocarcinoma samples. They found that roughly 80% of the mutations identified in cancer samples had already been present in the neighbouring Barrett's epithelial tissue.

In order to characterize a gradual accumulation of mutations across different disease stages, Weaver et al., 2014 screened for the most recurrently mutated genes in Barrett's in cohorts who had never had dysplasia and those with high-grade dysplasia. Interestingly, they found multiple mutations occurring in never-dysplastic individuals at considerable allelic frequencies (>10%). Furthermore, the most widespread mutations present in adenocarcinoma were also occurring at similar rates in both non-dysplastic and high-grade dysplasia Barrett's samples, among which were mutations present in known cancer-associated genes, ARID1A and SMARCA4. Their results were a good example of complex mutational processes that occur even in tissues with a low risk of malignant progress. On the other hand, the authors argued that considering the mutations in ARID1A and SMARCA4

were present in patients without any signs of dysplasia over multiple years of follow-up, they could not play a causal role in disease progression. Instead, they indicated TP53 and SMAD4 mutations as more likely candidates, because they were specific to patients with high-grade dysplasia or oesophageal adenocarcinoma (Ross-Innes et al., 2015). A similar relationship was discovered by Paulson et al., 2022, whose multi-sample WGS revealed that the critical difference between individuals with Barrett's who either do or do not progress to adenocarcinoma was clonal expansion of TP53−/− cell populations and subsequent chromosomal structural events. Structural rearrangements and copy number changes are another aspect common for adenocarcinoma but rarely occurring in Barrett's epithelium. This was something that Ross-Innes et al. found in their later study (Ross-Innes et al., 2015), aiming to decipher the clonal architecture of Barrett's tissues over time. In addition, they found that while numerous point mutations and small indels would occur at all stages of disease progression, the specific SNVs were overlapping poorly between Barrett's and adenocarcinoma. Instead, it was the mutational context of the point mutations that was most frequent throughout the disease, suggesting that chronic exposure to similar mutagens could be the reason for disease progression.

Several endoscopic therapies exist that target Barrett's oesophagus in order to prevent its progression to adenocarcinoma, but they are costly and usually require multiple interventions (Shaheen et al., 2009). Understanding mechanisms driving the development of a highly heterogenous disease such as Barrett's has a potential to result in novel approaches that could be a reasonable supplement or, even, alternative to the endoscopic therapies. Furthermore, it would also give valuable insights into diverse biological processes such as wound healing, embryogenesis and cancer (Rhee and Wang, 2018).

The connection of Barrett's with healthy tissues remains unknown, and identification of the cell of origin is currently under intense study. Numerous hypothesis include dissemination from bone marrow, reparative emergence of oesophageal submucosal glands, or migration of gastric cardiac epithelium (Xian et al., 2019).

The bone marrow theory explained the origins of Barrett's oesophagus by the colonization of the acid-damaged oesophagus by circulating bone marrow stem cells. One evidence for that included the formation of Barrett's metaplasia after the transplantation of bone marrow info a wild-type mouse, followed by surgical esophagojejunostomy (Hutchinson et al., 2011). Similar results were found by Sarosi et al., 2008, who gave female rats a lethal dose of irradiation, followed by tail vein injection of bone marrow cells from male rats. Their results suggested a contribution of multi-potential progenitor cells of bone marrow origin to oesophageal regeneration and metaplasia. However, it is argued that the potential of bone marrow stem cells to form epithelial populations has not been established. Furthermore, the incorporation of bone marrow stem cells proposed in this way does not explain the foundation of Barrett's exclusively at the gastrointestinal junction (Xian et al., 2019).

Because Barrett's is known to be linked to gastroesophageal reflux, the migration of gastric cardiac epithelium to repair the reflux-mediated damage in the neighbouring oesophageal epithelium has been suspected to lead to the disease. It has previously been confirmed that the upward expansion of gastric cardiac epithelium was possible in the presence of Barrett's oesophagus (Bremner et al., 1970). Nowicki-Osuch et al., 2021 identified undifferentiated gastric cells which would eventually transform into Barrett's by certain transcriptional programs. Despite strong similarities between the undifferentiated

gastric cardia cells and Barrett's, causality has not been proven explicitly (Geboes and Hoorens, 2021). However, the results provide a strong support for the previous evidence of shared characteristics of Barrett's and gastric cardiac epithelium, such as the proposed gastric-origin goblet cell migrating up from the stomach to the oesophagus (Jin and Mills, 2018).

Oesophageal submucosal glands and ducts have also been proposed as potential sources of Barrett's progenitor cells (Lörinc and Öberg, 2012, Van Nieuwenhove and Willems, 1998, von Furstenberg et al., 2017, Leedham et al., 2008, Coad et al., 2005, Lörinc et al., 2015, Owen et al., 2018) . Other evidence suggests that Barrett's might originate directly from the native oesophageal squamous epithelium (Hu et al., 2007, Quante et al., 2012). Despite the fact that Barrett's always originates in the oesophagus, similar concentrations of oesophageal submucosal glands are found in proximal and distal regions (van Nieuwenhove et al., 2001). Furthermore, oesophageal submucosal glands express p63 (Glickman et al., 2001), which is absent in the stem cells of Barrett's oesophagus (Yamamoto et al., 2016). On the other hand, single-cell RNAseq performed by Owen et al., 2018 revealed a profound transcriptional overlap of Barrett's cell populations marked by LEFTY1 and OLFM4 with oesophageal submucosal glands, but not with gastric or duodenal cells. In addition to that, they found SPINK4 and ITLN1 cells that preceded morphologically identifiable goblet cells in colon and Barrett's, potentially facilitating the detection of metaplasia.

The best way to determine the cell of origin of Barrett's would be to trace the lineage from precursor cell to Barrett's using a genetic marker that is not under positive selection, such as passenger mutations or mutational signatures.

1.11.3 Goals

In the following chapter, we present the application of our caller to the Barrett's dataset, in which connections of Barrett's to oesophageal submucosal glands have previously been described. Our goal is to combine earlier transcriptomic with novel genomic information, to provide a new insight into the origins of Barrett's. Specifically, we are hoping to find new evidence for either Oesophagus or Stomach as the source of Barrett's precursor.

1.12 Single-cell RNAseq variant calling

1.12.1 General results

1.12.1.1 Only a third of single-cell calls detectable in bulk RNAseq

Before progressing to the analysis of single-cell SNVs, we wanted to explore the uniqueness of our calls. Specifically, we wanted to investigate the number of single-cell SNVs that could be identified from the bulk RNAseq. We found that only around 30% of our variants passed the QC criteria in bulk RNAseq, which were coverage of at least 3 and at least 3 supporting reads (**Figure 23**). While some variants were supported by 1 or 2 reads in bulk RNAseq, most were not detectable at all.

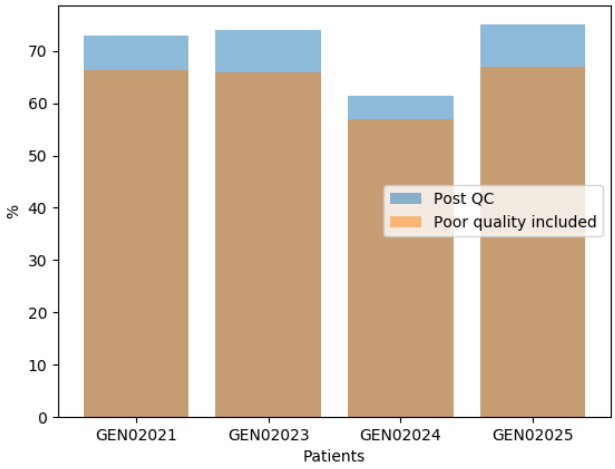
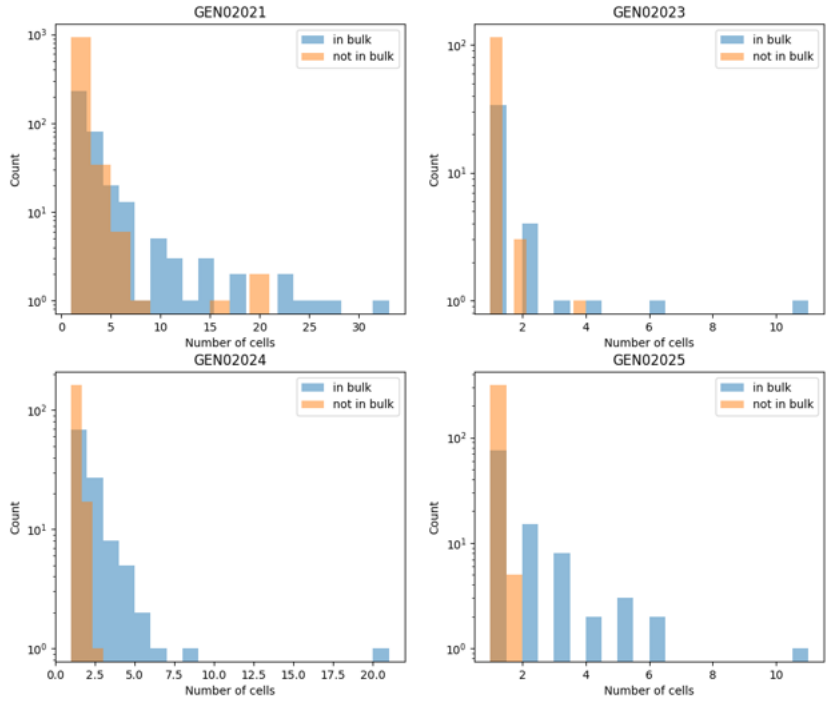


Figure 23. Fraction of single-cell RNAseq calls not passing QC criteria in bulk RNAseq ("Post QC") or not detectable in bulk RNAseq at all ("Poor quality included").

We wondered whether the detectability of variants in bulk RNAseq, linked to allelic frequency of variants in a population of cells, could already be estimated at the level of our single-cells. Indeed, we noticed that variants not detectable in bulk were, on average, found in less than 10 cells for patient GEN02021, less than 4 for GEN02023 and less than 3 for GEN02024 and GEN02025 (**Figure 24a**). Mann-Whitney U test statistics confirmed that there were significant differences between the means of the detectable and non-detectable groups (**Figure 24b**).

a



b

Patient	Mann-U test	p-value	Outcome
GEN02021	259441.5	8.72e-75	statistically greater
GEN02023	2895.5	4.06e-4	statistically greater
GEN02024	13508.0	1.92e-10	statistically greater
GEN02024	21921.5	2.73e-19	statistically greater

Figure 24. a. Comparison of single-cell variants detectable or not in bulk RNAseq. b. Mann-Whitney U test results confirming the differences between the means are significant.

1.12.1.2 Per-cell mutation frequency tissue rankings not consistent between patients

In order to get an overview of the output for different patients and tissues, we calculated the number of calls per cell and grouped them accordingly (**Figure 25**). We found that the majority of cells had less than 50 calls, while, on the other hand, up to 800 calls were made from cells with the most calls. The patient with the highest average number of calls per cell was GEN02021 (250.6 +- 233.5 versus 34.7 +- 60.0 for patient GEN02023, 40.8 +- 59.6

for GEN02024 and 184.9 +- 148.2 for GEN02025), while the tissue with the most calls was Barrett's (210.4 +- 221.6 versus 182.0 +- 212.7 for Oesophagus, 80.9 +- 130.4 for Gastric and 139.0 +- 149.3 for Duodenum).

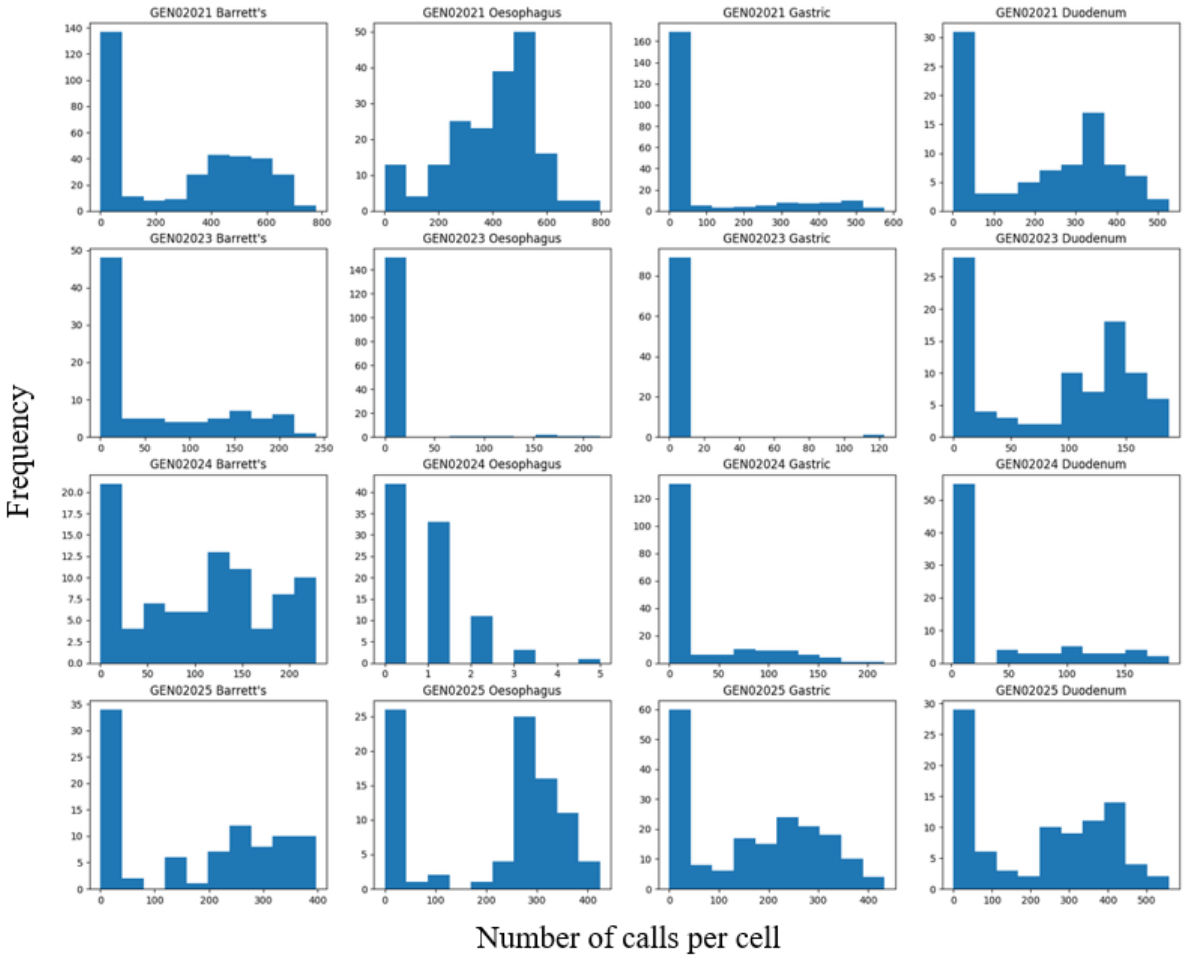


Figure 25. Distributions of the numbers of SNVs called from single-cells, grouped by patient and tissue.

Because coverage and the number of reads per single-cell were diverse, the absolute numbers of calls per cell were not a reliable way of comparing the patients and tissues. In order to correct for that, we calculated mutation frequencies by dividing the numbers of calls by the numbers of positions in which calls could be made (in other words, in regions covered

by at least 3 reads). It was still patient GEN02021 who had the highest frequency of calls (2.3e-06 +- 3.1e-06 versus 8.7e-07 +- 1.4e-06 for patient GEN02023, 6.7e-07 +- 1.2e-06 for GEN02024 and 1.9e-06 +- 2.6e-06 for GEN02025). However, Oesophagus, and not Barrett's, had the highest mutation rate this time (2.1e-06 +- 3.0e-06 versus 2.0e-06 +- 2.8e-06 for Barrett's, 1.2e-06 +- 2.0e-06 for Gastric and 1.5e-06 +- 2.4e-06 for Duodenum).

The next step involved the investigation and subsequent removal of SNPs called from bulk RNAseq. Because SNPs had a significant contribution to the single-cell calls, we expected the average numbers of calls per cell to be much lower once germline variants were eliminated. We found that our method identified only up to 3% of bulk RNAseq SNPs that were present at sufficient quality in at least one single-cell. However, less than 0.1% of those SNPs had suitable neighbours for the linkage method, and therefore, could not in fact be called. Because our variant calling approach was biased towards the quality and not the number of the variants called, we treated a high percentage of SNPs among our calls, rather the fraction of bulk SNPs identified from the single-cells, as a fair validation of our approach. Therefore, we were content to see that the removal of SNPs from our calls resulted in a 98.7% decrease in the number of calls from Barrett's, 99.4% for Oesophagus, 98.5% for Gastric and 99.5% for Duodenum, as it proved the high accuracy of our approach . The average number of calls per Barrett's cell after SNP removal was now 5.4 +- 4.5, 3.3 +- 2.2 for Oesophagus, 4.0 +- 3.2 for Gastric and 2.2 +- 1.6 for Duodenum (single-cells with 0 calls were not taken into account).

We also recalculated the mutation frequencies per cell (**Figure 26**). We found that the removal of SNPs resulted in Barrett's cells having the highest mutation rate (4.9e-08 +- 8.0e-08), followed by Gastric (4.5e-08 +- 1.4e-07), Oesophagus (3.2e-08 +- 7.2e-08) and

Duodenum (1.4e-08 +- 3.1e-08). Interestingly, this ranking was patient-dependent. While Barrett's had the highest mutation rate in most patients, this was not the case for GEN02025. Similarly, while the mutation rate in the Gastric cells was high in comparison to the other tissues in patients GEN02024 and GEN02025, Oesophagus had a higher rate in patient GEN02021 and Duodenum in GEN02023.

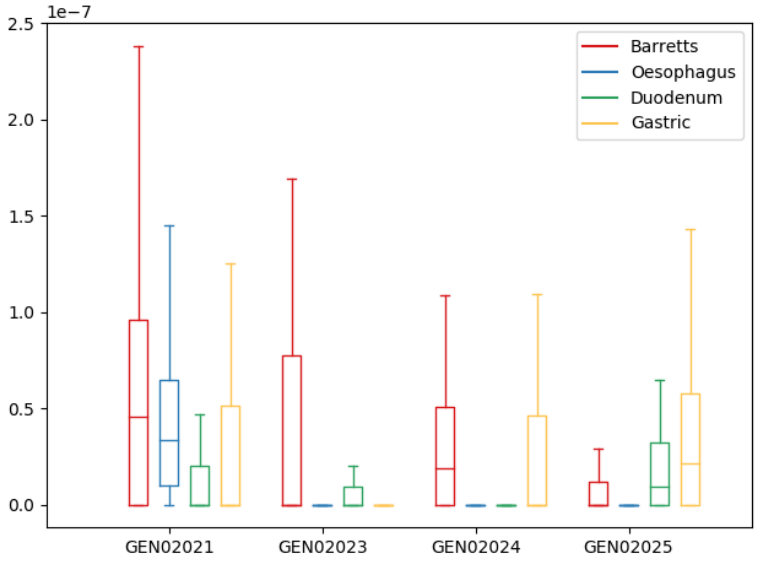


Figure 26. Frequencies of calls identified from single-cells post SNP removal, grouped by patient and tissue.

As already mentioned, the SNPs we removed from the single-cell calls were originally called from bulk RNAseq. However, it was possible that some might have been missed by the caller. In order to maximize the number of SNPs removed from our calls, we decided to eliminate all single-cell calls that had evidence in the bulk. Specifically, we removed all alleles that were present in at least two bulk RNAseq tissues. The number of such positions was not as considerable as we had been expecting, however (1 for GEN02021, none for GEN02023, 5 for GEN02024 and 9 for GEN02025).

1.12.1.3 Mutation counts or frequencies might not be a reliable comparison method for single-cell data

We wondered whether using the single-cell information to compare samples at tissue-level would result in different rates of mutation accumulation per tissue. Specifically, our goal was to compare the tissues in terms of the numbers of their unique mutations by combining single-cell calls from the respective tissue. We anticipated the mutations to be more abundant in clonal and highly heterogeneous tissues of Barrett's. Furthermore, we expected that more non-SNP mutations would be shared between Barrett's cells than cells from Oesophagus, Stomach or Duodenum.

The numbers of calls per tissue are presented in **Table 11.** As previously, Barrett's had the most calls for patients GEN02021 and GEN02023, while Gastric for GEN0204 and GEN02025.

Table 11. Numbers of calls per tissue and patient, constructed from single-cell data

	Barrett's	Oesophagus	Gastric	Duodenum
GEN02021	889	289	276	34
GEN02023	98	41	4	24
GEN02024	69	0	225	35
GEN02025	30	18	276	123

However, we expected those results to be biased due to the fact that some of the calls could have been made in multiple tissues. We found that no mutations were shared by more than 2 tissues, which was expected as those positions were removed from the single-cell calls as potential SNPs. Just under 200 mutations in total were shared between two tissues (always

one of those tissues was Barrett's), and had to be removed for an accurate tissue-specific comparison.

Just as mutation frequencies were used to compare single-cell cohorts, we wanted to ensure the fairness of the per-tissue comparison. In other words, we believed reliable assumptions about the tissue-specific mutations could only be made provided the regions were covered in other tissues. Because single-cells cover a very narrow range of the human transcriptome, we were not sure whether identification of regions covered in all tissues was feasible. However, analysis of the number of healthy tissues in which sufficient coverage existed revealed that this was possible (**Figure 27**).

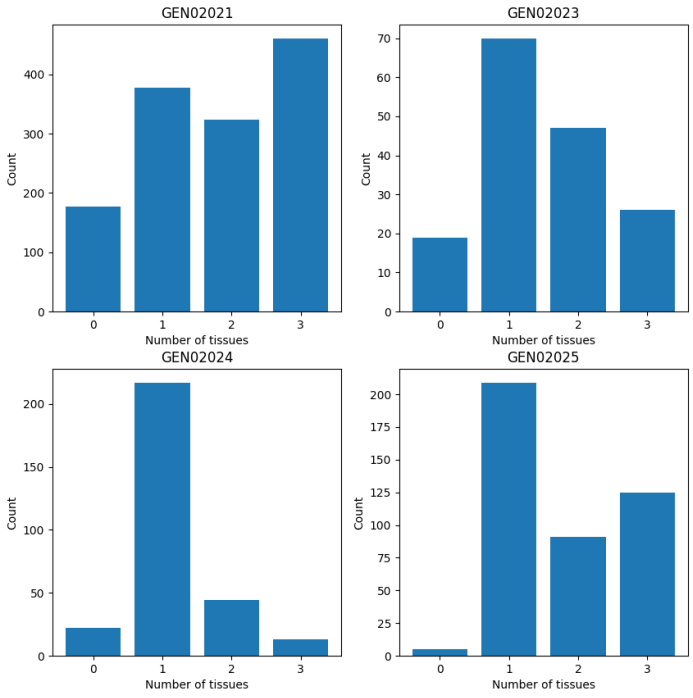


Figure 27. Coverage of each single-cell mutation in other tissues. The plots include regions covered without the actual variants present. Bars with "0" represent coverage solely in the Barrett's tissue.

The final counts of mutations per tissue, present only in the tissue of interest but covered in all other tissues, are presented in **Table 12.** Interestingly, the ranking of the tissues in terms of mutation abundance did not change. It was still Barrett's with the highest number of calls for patients GEN02021 and GEN02023, while more unique calls were identified for Gastric in patients GEN02024 and Duodenum in GEN02025. We investigated the metadata hoping to discover information about the varying disease characteristics that those patients experienced, but the only difference between those two patient groups was sex. We concluded this was not a likely influencer of the results, however, as the coverage of chromosome X was similar across all patients (2.4% of all shared regions in GEN02021, 3.4% in GEN02023, 3.0% in GEN02024 and 2.8% in GEN02025). Furthermore, the ranking of tissues according to the fraction of variants in chromosome X also did not correlate with the results in **Table 12** (highest in Duodenum in GEN02021, Barrett's in GEN02023, no calls in chromosome X in GEN02024 and Gastric in GEN02025).

Table 12. Tissue-specific mutations, restricted to shared regions.

	Barrett's	Oesophagus	Gastric	Duodenum
GEN02021	251	79	12	77
GEN02023	20	1	1	1
GEN02024	2	0	4	7
GEN02025	8	6	40	68

We had two potential explanations for the differences between patients GEN02021 & GEN2023 and GEN02024 & GEN02025. The first one assumed the results were an accurate representation of the biological phenomena and could indicate different

mechanisms occurring during Barrett's progression. Specifically, as had been described in the introduction, they might provide evidence that point mutations are not the only cause of the disease. On the other hand, while raw mutation counts or mutation frequencies are often used to estimate mutational burden within tissues, we believe they should be treated with reservations when working with single-cells. The reason for that is that the transcriptome of single-cells covers a very narrow range of the genome, and might not be a truthful approximation of the general mutational landscape. Furthermore, our linkage method heavily relies on the neighbouring positions differing from the reference in order to collect evidence for a call. Therefore, even if a true mutation is present at a high allelic frequency, it might not be called (unless it had been identified in a different single-cell, then the recalling step would result in calling it as well).

In order to investigate whether the mutation rates were influenced by the way we performed the calling, we recalculated the rates using only "high confidence" variants. Previously defined as positions present in regions with coverage > 50 and allelic frequency >= 0.4. We defined the new mutation rates as the number of "high confidence" variants divided by the total number of positions with coverage of over 50 reads, measured per cell and grouped by patient and tissue type. While the differences between the mutation rates in each tissue from the same patient were not striking, we observed the same patterns as before, namely the highest values for Barrett's in patients GEN02021 and GEN02023, for Duodenum in GEN02024 and Gastric for GEN02025 (**Figure 28**). The results supported the hypothesis of biological differences between the patients, rather than the limitations of our calling in the context of mutation rate reconstruction.

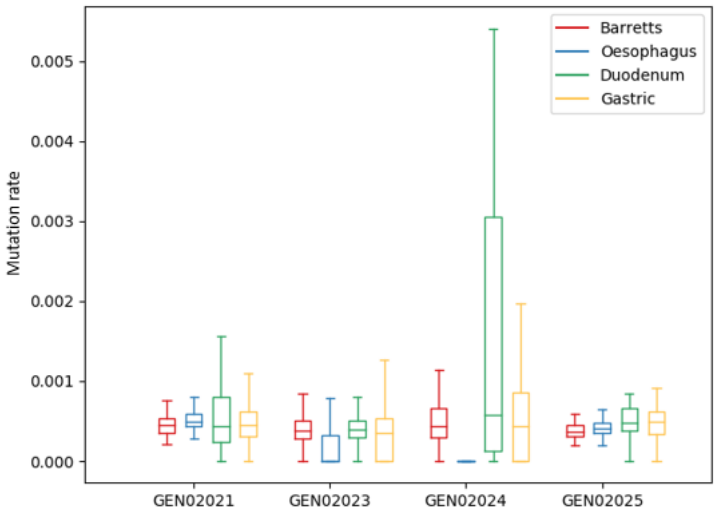


Figure 28. Mutation rates calculated from the "high confidence" variants in regions with coverage of >50 reads, grouped by patient and tissue.

Satisfied with the results so far, we were curious to investigate whether the single-cell calls could reveal any information about Barrett's. Therefore, we decided to progress to more specific questions about the disease. Our approach was to begin with the search for evidence that would support or deny what is currently known. Once such was found, we were hoping to take advantage of the single-cell data to explore it further.

1.12.2 Barrett's-specific analysis

1.12.2.1 No signs of Signature S17 in Barrett's samples

Two predominant mutation signatures have been consistently characterized in Barrett's and oesophageal adenocarcinoma (Galipeau et al., 2018). Those include COSMIC Signatures S1 and S17. Signature S1 is associated with aging and cancer progression, and is linked to spontaneous deamination of 5-methylcytosine accumulating over cell divisions at

regular intervals. Signature S17, on the other hand, is characterized primarily by T > G and T > C substitutions in CTT contexts. It has not only been linked to both oesophageal and gastric cancers, but also identified in surrounding Barrett's tissues. A reconstruction of those signatures from the single-cell calls, especially the unique Signature S17, would be a straightforward way to firstly confirm the genetic grounds of Barrett's in our dataset and, secondly, to prove the ability of our caller to accurately recreate mutational landscapes.

We used calls from **Table 11** to calculate mutation profiles for each tissue and patient (Barrett's mutation profiles in **Figure 29**, the remaining tissues can be found in Supplement 1.15).

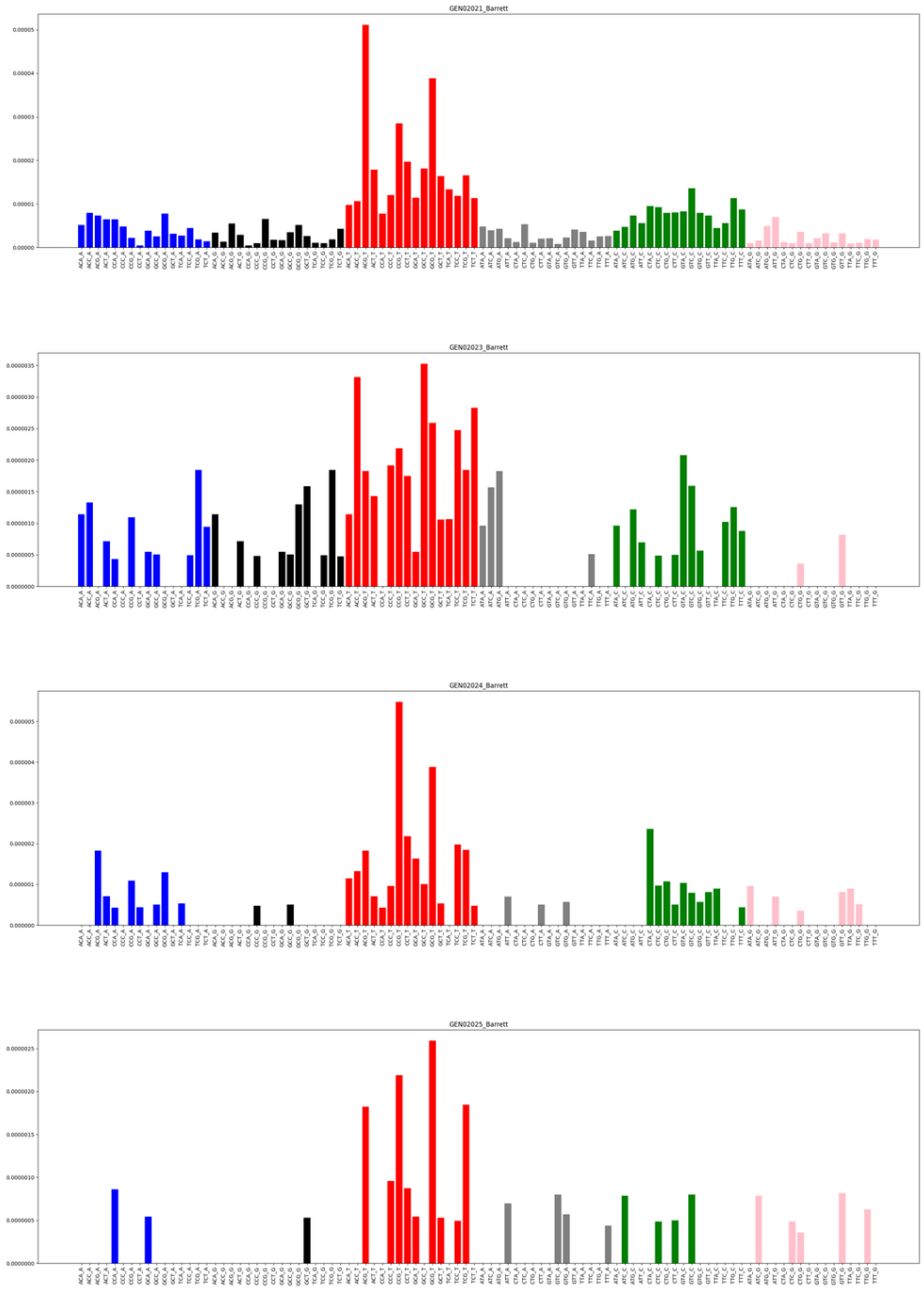


Figure 29. Barrett's mutation profiles reconstructed from single-cell calls.

We did not see any substantial differences between Barrett's and the remaining healthy tissues, neither did we identify any signs of Signature S17. The lack of Signature S17

was further confirmed by the output of the *desconstructSigs* software (**Figure 30**) which, however, did identify Signature 1 in all patients. A wide range of other signatures has also been identified. While they have not been linked to Barrett's, we did not consider this as a failure of our method as it is common to see a combination of various signatures across samples, based on how the signatures are calculated.

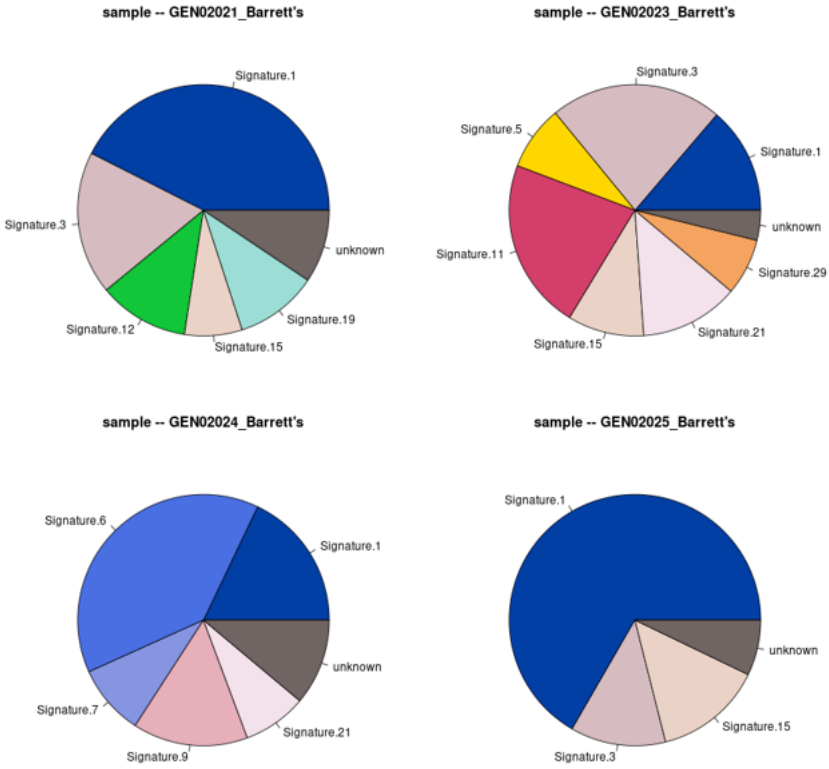


Figure 30. Barrett's mutation signatures reconstructed from single-cell calls.

We wondered whether the Barrett's specific signatures would be identified if only mutations unique to Barrett's were considered. In order to check that, we removed all calls discovered in matching healthy tissues from the Barrett's set and recalculated the signatures.

However, still no evidence of Signature S17 was found (**Figure 31**). Moreover, the removal of calls shared with healthy tissues resulted in the disappearance of Signature S1 in patient GEN02024.

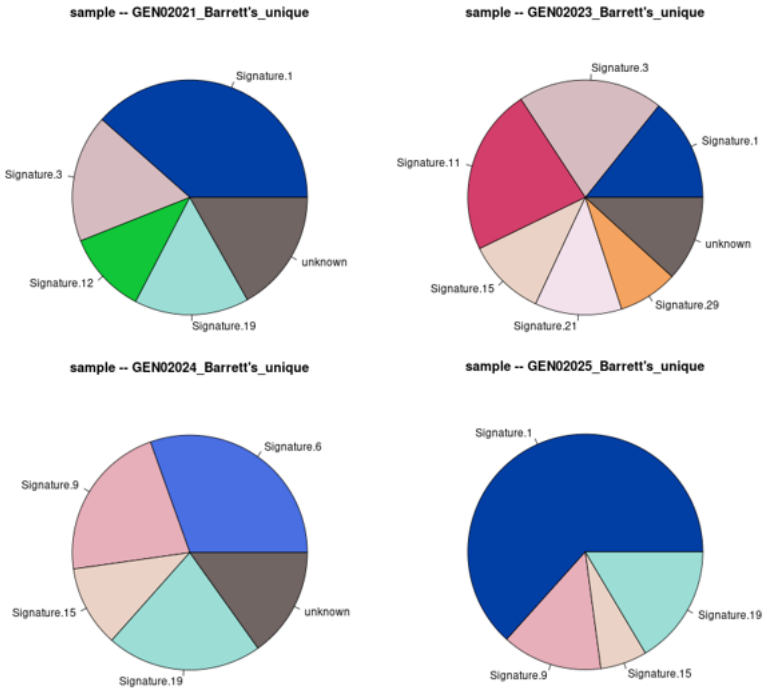


Figure 31. Barrett's mutation signatures reconstructed from Barrett's-specific single-cell calls.

Because we did not have any paired DNA samples to compare our results to the actual mutation profiles of the somatic variants in our data, we were not able to determine whether there was a genuine lack of Barrett's signatures in our data or a bias introduced as a result of the way we performed the calling. Therefore, we decided to progress to more specific aspects of our calls – namely, the individual SNVs.

1.12.2.2 SMARCA4 mutation identified in a Barrett's single-cell

As aforementioned in the Introduction, there have been numerous genes linked to Barrett's and oesophageal adenocarcinoma. We were particularly interested in four genes: TP53, SMAD4, ARID1A and SMARCA4 (Ross-Innes et al., 2015), and wondered whether we could identify them among our single-cell calls.

In order to do this, we annotated the Barrett's-specific variants with the Variant Effect Predictor (VEP) tool. We did not identify any signs of TP53, SMAD4 or ARID1A in our calls. However, there was a hit for SMARCA4 in a Barrett's single-cell for patient GEN02021 (chr19:10986212:T>C, CTC context). Searching for this variant in the remaining single-cells from the same patient resulted in the identification of a second Barrett's single-cell. The mutation had not been called because it did not pass the required quality criteria (only 1 supporting read in a region covered by 2 reads).

Because no single-cells from other tissues contained the SMARCA4 variant, and often did not even cover the region, we were not able to investigate any connections with Barrett's. However, we wondered whether this would be possible at the level of the bulk tissues. Interestingly, we found evidence for this mutation in 7 bulk RNAseq samples (2 for patient GEN02021, 2 for GEN02024 and 3 for GEN02025). The bulk samples from patient GEN02021 belonged to Barrett's and Oesophagus, and the variant appeared at allelic frequencies of 0.08 (1 supporting read, coverage of 13) and 0.03 (1 supporting read, coverage of 29), respectively. The tissues from patient GEN02024, in which the SMARCA4 variant was identified, were Oesophagus (AF=0.05, 1 supporting read, coverage = 19) and Duodenum (AF=0.2, 1 supporting read, coverage = 5). There was also a mutation at the same position in the Barrett's bulk RNAseq from this patient but it was a T>A. Finally, two

samples from patient GEN02025 were from Oesophagus (AF=0.04, 1 supporting read, coverage = 26 and AF=0.03, 1 supporting read, coverage = 29) and one from Duodenum (AF=0.17, 1 supporting read, coverage = 6).

The fact that the SMARCA4 variant was identified in Barrett's and Oesophagus in patient GEN02021 and in none of the Gastric bulk tissues from any patients was definitely interesting, as might indicate Oesophagus as a potential origin of Barrett's. However, there was no support for that in the single-cell data, as no single-cell from Oesophagus with this variant was found. Furthermore, we did not discover any evidence for this variant in single-cells from other tissues and patients. We appreciate that we might not have been lucky enough to have such cells in our dataset, but on the other hand, we could not exclude the possibility of the variant found in the Oesophagus to be a technical error. However, the fact that we identified it in different samples made us suspect that it actually was real.

We were particularly surprised to have discovered it in the bulk samples from the Duodenum, because we had treated it mainly as our healthy control, rather than consider it an actual location of Barrett's progenitor cells. However, Ross-Innes et.al. stated that SMARCA4 was identified even in tissues with a low risk of malignant progress and was therefore unlikely to play a causal role in disease progression. We concluded that this could explain our results if, again, the variant found in the Duodenum was an actual mutation.

The four genes considered were not an extensive list of Barrett's associated mutations, and we considered other genes such as NOTCH1, NOTCH3, FAT1 or PIK3CA (Martincorena et al., 2018). However, the poor overlap of their regions with our single-cells encouraged us to return to the level of individual SNVs. If there were recurrent point mutations within cancer-related genes, we would still identify them. More importantly, this

bottom-up approach would give us an opportunity to identify novel genes that we might miss if we only focused on the genes already known.

1.12.2.3 No evidence for the Barrett's origin in the Oesophagus

Single-cell RNAseq performed by Owen et al., 2018 revealed a profound transcriptional overlap of Barrett's cell populations with oesophageal submucosal glands, but not with gastric or duodenal cells. Because we used the same dataset, we were curious to investigate whether our genomic results would support the transcriptomic discoveries. Combining transcriptomic and genomic aspects of single-cell RNAseq would be an original approach to investigating not only Barrett's, but somatic mutagenesis in clonal diseases in general. If our results did agree with Owen et al., they would provide a solid evidence for the hypothetical origin of Barrett's from the Oesophagus (more specifically, oesophageal submucosal glands).

Owen et al. performed two types of clustering, which we decided to take advantage of in order to compare our results directly. The first type of cell segregation involved the relation to known cell types based on the cellular expression of genes characterized in the past. There were 7 types that the single-cells were assigned to: squamous, non-epithelial, mucus neck, goblet-type, enteroendocrine, enterocyte and Barrett's-type (**Figure 32**). Most cells from Barrett's oesophagus were labelled as Barrett's type, while non-epithelial, goblet-type and enteroendocrine were found as well. The Barrett's type cells were also found in the Gastric and Oesophageal tissues, but not in Duodenum (hence using it mainly as a healthy control). Apart from their connection with Barrett's, goblet-type cells were only found in the Oesophagus. The second type of clustering was based on gene expression, and resulted in

the grouping of single-cells into 11 novel clusters (4 each for Barrett's, Oesophagus and Duodenum and 3 for Gastric). We were focused primarily on Oesophageal clusters O3 and O4, as the detection of TFF3 expression in both of them indicated their relationship with the oesophageal submucosal glands.

Figure 32. Separation of single-cells into cell types and their relationships with the corresponding tissues of origin. Source: Owen et al., 2018

The clustering performed by Owen et al. indicated a similarity of Barrett's tissue to all other tissue types (**Figure 33a**). The cell-type clustering revealed even more complex relationships of Barrett's-type cells to most cell types, namely enterocyte, enteroendocrine, goblet-type, mucus neck and non-epithelial cell types (**Figure 33b**). Interestingly, the oesophageal squamous cells were the only cluster not visibly close to Barrett's. They corresponded to clusters O1 and O2, what indicated that the oesophageal cell types similar to Barrett's were the aforementioned clusters O3 and O4 (oesophageal submucosal glands).

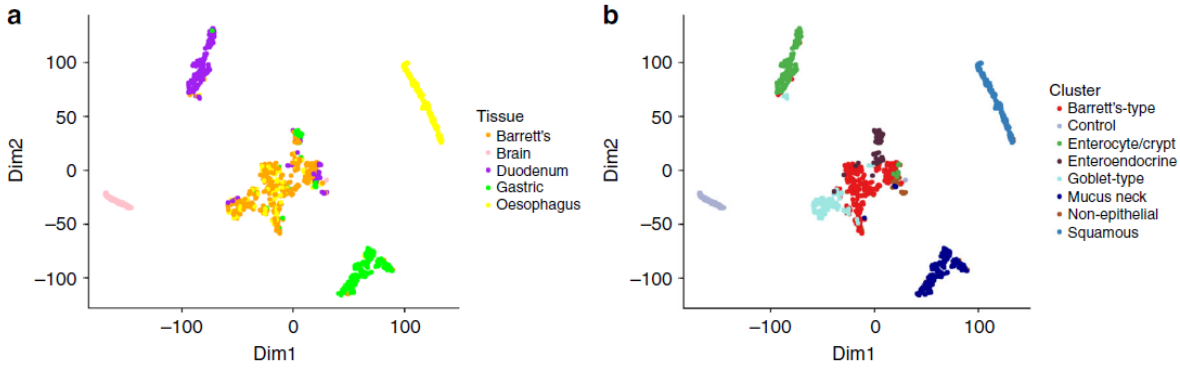


Figure 33. Clustering of single-cells by a. tissue type and b. cell type. Source: Owen et al., 2018

Because we did not use the Brain control tissue in our analysis, we wanted to ensure the cluster relationships were maintained without it. The t-SNE clustering we executed confirmed that (**Figure 34**). Just like in the original clustering, Barrett's-type cells were closely linked to enteroendocrine and goblet cells, while squamous remained distant. Similar relationships with the remaining cell types were also unaltered.

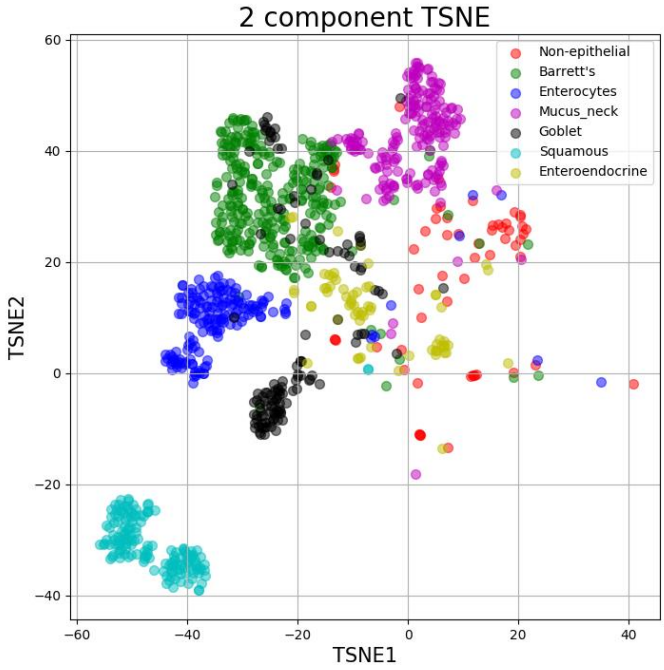


Figure 34. **t-SNE clustering of the single-cells analyzed in this study.**

We began our analysis with the consideration of calls per cell in each cell type. We expected to see higher values for cell types most closely related to Barrett's, which would indicate higher rates of mutagenesis in disease or precursor tissues. On average, there were 4.03 +- 3.40 calls per cell for non-epithelial cells, 4.72 +- 3.68 for Barrett's-type, 2.08 +- 2.05 for enterocytes, 3.63 +- 2.58 for mucus neck, 5.81 +- 5.03 for goblet, 2.01 +- 1.16 for squamous and 8.55 +- 8.61 for enteroendocrine (the exact numbers of mutations per cell type and tissue can be found in Supplement 1,16). Therefore, just as we expected based on the transcriptomic clustering, it was the Barrett's-type, goblet and enteroendocrine cells that had the most calls. The distribution of calls per cell, grouped per cell type, also revealed similar

patterns, with some cells from the three types having substantially more calls than others (**Figure 35**).

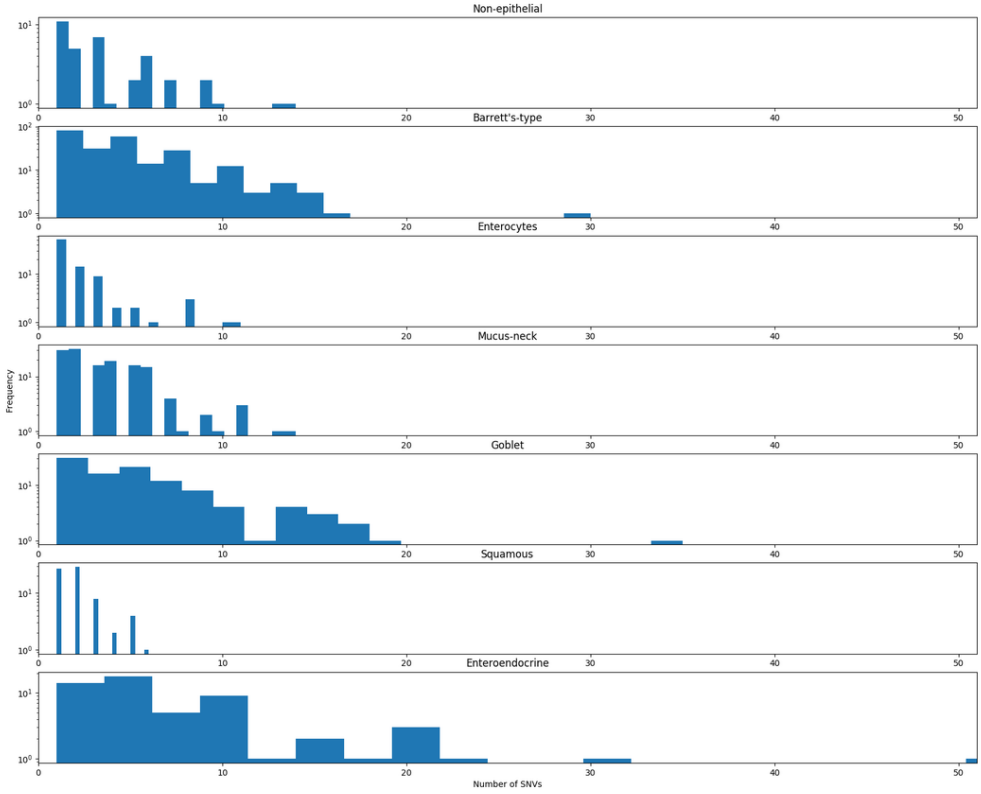


Figure 35. Distributions of the number of calls per cell, grouped by cell type.

Because the number of calls could be the correlated with the cell size, we wanted to confirm this was not the case. We divided the number of calls by the size of the callable region to obtain mutation rates for each cell type and tissue (**Table 11**). Interestingly, it was the non-epithelial and enteroendocrine cells that had the highest mutation rates (with enteroendocrine cells in Barrett's, Oesophagus and Gastric having substantially more mutations than in the Duodenum), and not Barrett's type as we had been expecting. Even

goblet cells, which we however had found to be close to Barrett's-type cells in the clustering, had higher mutation rates than Barrett's-type. On the other hand, mutation rates were lower for mucus neck and squamous cell types. This is something we had been predicting because, based on previous results, we considered them as less likely Barrett's progenitors. We kept in mind that a bias introduced due to limitations of our calling method was possible. However, we still believed that they were a valuable addition to the evidence we were gathering, as we believed this was the fairest way of comparing mutation load in different cell types.

Table 13. Mutation rates in different cell types and tissues.

Cell type	Barrett's	Oesophagus	Gastric	Duodenum
Non-epithelial	1.85	1.33	2.41	1.84
Barrett's-type	1.21	0.99	1.46	0.00
Enterocytes	0.00	0.00	0.00	0.55
Mucus neck	0.87	0.35	1.08	0.00
Goblet	1.67	1.50	0.00	0.55
Squamous	0.00	0.49	0.00	0.00
Enteroendocrine	2.22	1.49	2.19	0.57

While comparing mutation rates was the first step to identifying potentially accelerated mutagenesis in certain tissues and cell types, only direct evidence of connections between them would prove their similarity. We found that three cell types had mutations identified in at least two different tissues: Barrett's-type (28 mutations), goblet (17 mutations) and enteroendocrine cells (10 mutations). Most connections were found between Barrett's & Oesophagus, followed by Barrett's & Gastric, Barrett's & Duodenum and Oesophagus & Duodenum (**Table 14**).

Table 14. Mutations identified across different tissues in cells of the same type.

Cell type	Barrett's-Oesophagus	Barrett's-Duodenum	Barrett's-Gastric	Oesophagus-Duodenum
Barrett's-type	26	0	2	0
Goblet	14	4	0	1
Enteroendocrine	9	2	8	0

The fact that mutations shared between Barrett's & Duodenum and Oesophagus & Duodenum were found made us suspect that SNP variants were still present among our calls. In order to prevent that, we removed variants that were covered only in the cells they were seen in. In other words, if no other cell of the same type with sufficient coverage and a different allele was found in the same tissue, we could not be sure if the suspected mutation was not just present across all cells. This procedure eliminated the majority of calls in different cell and tissue types. Interestingly, the Oesophagus & Duodenum connection was still present, and the only link that disappeared was Barrett's & Gastric for Barrett's-type cells (**Table 15**).

Table 15. Mutations identified across different tissues in cells of the same type after additional SNP removal.

Cell type	Barrett's-Oesophagus	Barrett's-Duodenum	Barrett's-Gastric	Oesophagus-Duodenum
Barrett's-type	9	0	0	0
Goblet	8	3	0	1
Enteroendocrine	1	1	6	0

We were wondering about the distribution of those mutations in different cells. Were they always the same cells? Were they of the same type? If we managed to identify a specific

set of cells that shared multiple mutations with Barrett's, if would be a good indication of their close connection to the disease.

Because most mutations from **Table 15** were identified in patient GEN02021 who had the most samples, we decided to restrict our analysis solely to this individual. We found multiple cells in which at least two mutations were identified. Analysis of their connections to other cells revealed a complex mutation landscape within and across various tissues. What we had been expecting was a set of different Barrett's cell clusters with unique connections to either Gastric or Oesophagus. Instead, we not only found just a single Gastric cell that had multiple mutations shared with Barrett's, but we also witnessed that most cells had various connections with multiple cells of different types. Furthermore, we also did not see a clear separation of Barrett's cells sharing mutations with either Oesophagus or Gastric – in fact, some Barrett's cells (such as GEN02021_Barret_Manual_P4_B10) shared unique mutations with both tissues.

We wondered whether changing perspective to mutation-level would allow us to identify trends in which the mutations spread across the tissues. Assuming that Barrett's originates from a single-cell, identifying mutations present in one type of a cell from a healthy tissue and, simultaneously, in multiple types of Barrett's cells could be an indication of a Barrett's progenitor. In order to investigate that, we considered each of the mutations shared between two tissues separately and checked which cells they were present in (**Figure 36**). We included all cells from patient GEN02021 in the analysis, even if they only contained a single mutation from **Table 15**, and marked all cells that belonged to clusters O3 and O4. Interestingly, we found that all oesophageal cells that shared mutations with Barrett's had originated from oesophageal submucosal glands. Furthermore, there were only two

enteroendocrine Gastric cells which shared mutations with Barrett's. The range of connections was broad. While there were mutations present in only cells of the same type (chr11:1024914:G, two Barrett's-type cells from Barrett's tissue and OSGs), there were also others identified in all cell types (chr3:42177960:C). In addition to that, we identified a few mutations present in one type of a cell from a healthy tissue and in multiple types of Barrett's cells (among others, chr2, chr4 and chr6).

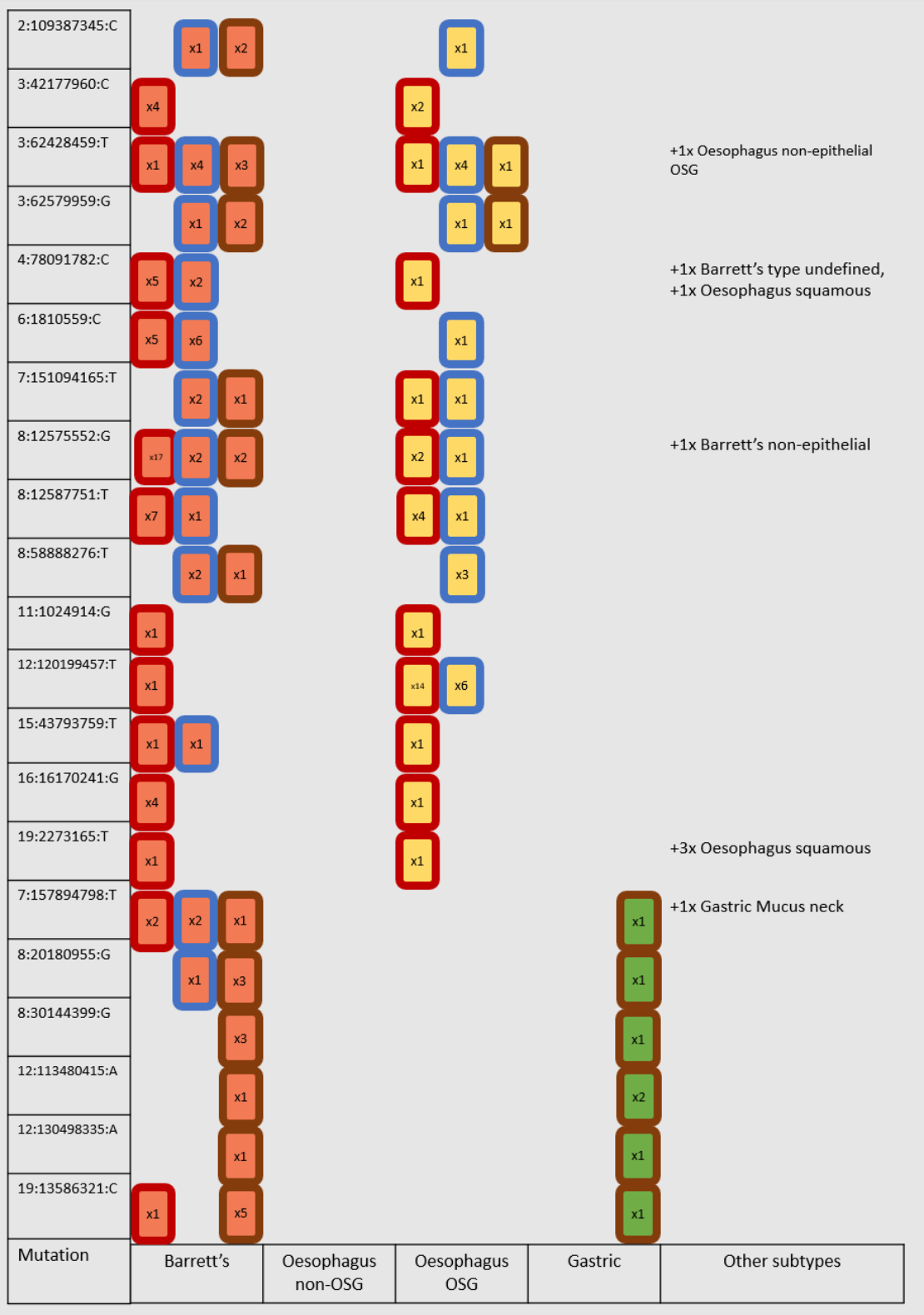


Figure 36. Presence of mutations spanning multiple tissues in different cell types. Frames: red = Barrett's-type, blue = goblet, brown = enteroendocrine.

The fact that majority of Barrett's mutations were shared with single-cells from OSGs could be treated as support of the hypothesis proposed by Owen et al. However, despite the fact that only two Gastric cells were considered, their connection with Barrett's could not be ignored. At this point, we still had not considered coverage of the mutations in different tissue types. As described previously, because we were dealing with RNAseq data, we could not exclude the possibility of the mutations to be absent due to lack of sufficient read support, rather than them not being there in the first place. In order to correct for that, we decided to re-examine every mutation and eliminate it if it was not covered in other healthy tissues. We were not sure whether we should consider different cell types as well. In other words, if a Gastric mutation was present in a Gastric enteroendocrine cell, would it be sufficient if it was only covered in a goblet cell in the Oesophagus? We decided that coverage in single-cells of different type was preferable as a more reliable control measure. Therefore, a mutation was eliminated unless the region was covered in multiple single-cells of different types in another healthy tissue. If it was covered in only one single-cell, it had to be of a different type than the Barrett's and the healthy tissue it was called in. Apart from coverage, the mutation was only considered further if no reads with the allele were found in the other tissues.

We found that only 12 out of 21 mutations (57%) were covered in at least one single-cell from Oesophagus or Gastric tissues (variant-dependent). 67% of them were in a cell of a different type, as required. Finally, only 5 mutations (24%) passed all the initial criteria, as they were not found in a single read in any single-cell from the other tissues. Among those, 3 mutations were Barrett's-Oesophagus-specific and 2 were shared between Barrett's and Gastric.

In order to confirm that the mutations were not found in either of the matched healthy controls (Oesophagus or Gastric) due to lack of coverage or randomness of the single-cell pool, we re-checked their occurrences in the single-cells from Duodenum. Indeed, we found that one Barrett's-Oesophagus connection was present (chr6:1810559:C) and was therefore removed from the list of tissue-specific connections.

Finally, because single-cells only represent a fraction of the whole tissue, we examined the bulk samples as well. We found multiple reads supporting two mutations (chr7: 151094165:T and chr8:58888276:T). Interestingly, both of them were unique to Barrett's and Oesophagus, meaning that the remaining two mutations with no identified presence in the other tissues were shared between Barrett's and Gastric. Using the VEP annotation tool revealed that they occurred in genes RIMBP2 and CACNA1A which, however, are not known to have any strong implications in cancer.

1.12.3 Conclusions and Discussion

Calling variants from single-cell RNAseq is a difficult challenge, and therefore it is not surprising that so few resources that allow that exist. We have found that only a small fraction of our calls was detectable in bulk, and while our quality criteria and calling were strict, we were very limited in terms of the validation of our results. However, the fact that we identified a large share of known SNPs and that multiple mutations were independently detected in various single-cells encouraged us to believe that our calls had reasonable accuracy.

We found two main limitations when analyzing single-cell RNAseq data: coverage and random sampling of cells. We would often find that the mutation was only called in one single-cell, but it was also only covered there. The lack of coverage in the other samples was very problematic, because we would not get any information about whether the mutation was present or not. Because we did not know what we were not able to see, we decided to reverse the perspective and remove ambiguity based on what we could detect. Therefore, we would eliminate Barrett's-Oesophagus-specific mutations if even one read in a single-cell from Gastric or Duodenum was identified. While the read could contain an unfortunate technical error, it was not something we could determine and we would still reject the mutation to reduce the risk of calling a false positive.

The filters applied by our caller are very stringent. It is deliberate, however, as one of our main goals was to minimize the number of false positives. While this action made us confident about the results, we are aware that certain parts of the analysis, especially mutation signatures, can be biased. The main factor influencing signature construction was the exclusion of all A to G mutations, which we decided to do in order to eliminate RNA editing positions. This was not an issue in our case as we calculated the signatures mainly with the aim of finding Signature 17, not heavily influenced by A to G mutations. However, the removal of all A to G mutations should be considered in the future, depending on the research question asked.

Other filters included in our tool could also be altered in order to help guide novel hypotheses in healthy tissues. Preferably, adjustments would be made to reverse transcriptase and PCR error rates, if such are known – in order to make our tool applicable to a wide range of datasets, we assumed the highest values and simultaneously removed substantial amounts

of uncertain calls. Another limitation of our tool was the lack of neighbours to determine the origin of positions different from the reference. Addition of known polymorphisms, on top of those included in the dbSNP database used by our tool, would expand the number of calls identified from data in regions of interest. Our tool is highly modular – and provides a solid base for future developments, such as an expansion to indel calling that we did not manage to add throughout the duration of the project.

The fact that multiple reads were not detectable in single-cells but in multiple bulk samples indicated a substantial amount of heterogeneity within healthy tissues. It made us appreciate the great amounts of single-cells required in order to fully examine the mutation landscapes. While it is difficult to determine the exact number of cells that would be sufficient, it is certain that the value in the order of thousands, and not dozens as in our case, is necessary. Heterogenous diseases with high mutation rates might require even greater numbers. Examining mutation profiles or the number of calls shared between cells would be a reasonable way to estimate the optimal number of cells – while the profiles and overlaps change substantially with every cell added, a saturation point should eventually be reached.

The fact that, ultimately, all Barrett's-Oesophagus-specific mutations were eliminated and only Barrett's-Gastric connections looked likely was peculiar. However, as they were not present within genes known to be implicated in cancer and were covered in very few single-cells from Oesophagus and Duodenum made us doubt that they were implicated in Barrett's, and that the Gastric was the tissue of origin. We strongly believe that more research is needed to confirm that.

5. Conclusions, Discussion and Afterword

The work presented in this Thesis is the result of the second half of a three-year research I conducted as a doctoral student in the Ludwig Institute for Cancer Research. While some time was required to learn the necessary programming and bioinformatics skills, the majority of the first part of my DPhil was spent on developing a different approach to single-cell RNAseq calling which ultimately turned out to not be useful for the overall goals of the project. Originally, our idea was to use machine learning to distinguish between real variants and technical errors. Calling variants from single-cell RNAseq data is a very exciting, but a very difficult, problem. It is therefore not surprising that existing methods are limited to applying strict thresholds to the outputs of tools developed for bulk or heavily rely on paired RNA and DNA samples. Yet, we decided that the idea of developing a single-cell RNAseq caller that would be independent of other tools and bulk tissues was worth taking the risk. We tried to create an optimal set of dataset features that would be fed to the models. Those features included basic information such as coverage, allelic frequency and quality of alignment, but also more complex ones including strand bias, position in the read or frequency of the mutation context that the position of interest was present in. We then experimented with different machine learning models - neural networks which we found to be best suited for the characteristics of our problem, or models such as random forests due to their interpretability. The final result was always similar. Despite achieving good recall, the caller would also identify substantial numbers of other likely false-positive positions (ie

low precision). Due to the "black-box" nature of the machine learning approach, we were not able to explain the calls that were listed by our caller. We tried using the explanations provided by the random forest, however, we found the information such as "the caller was mainly using allelic frequency and base quality" insufficient in order to gain full confidence in our results.

Discovery of things previously unknown requires an extensive validation and the lack of a solid ground truth set of mutations was an issue we struggled with. We constructed the ground truth mutation list based on the intersection of commonly used bulk methods. A consensus approach is commonly applied to calling variants in WGS data, increasing the false-positive rate at the expense of more false negatives. However, any false negatives in our ground truth set would affect the reported accuracy of our single-cell caller.

To overcome the problem of ambiguity among the ground truth list, we experimented with the creation of a clean dataset with simulated mutations and technical errors in known locations. Machine learning was quite straightforward now that we knew the exact origin of every position different from the reference. However, we found that what we learnt from the artificial dataset variant calling could not be directly translated to real data. Namely, there were numerous mutations in the non-artificial dataset that would not resemble any in the artificial dataset in terms of coverage and allelic frequency and vice versa. We concluded that the complexity of the single-cell RNAseq, an understudied area especially in terms of variant calling, could not be captured with the existing models developed based on the understanding of bulk tissues.

Ultimately, we returned to the drawing board. On one hand, we started completely from scratch – we knew that we required a different approach. On the other hand, we were

now more aware of the issues with variant calling from single-cell RNAseq, and we knew what we required from our caller in terms of its interpretability and validation.

We decided that it was the quality, and not the quantity, of variants that we were interested in. In other words, we much preferred calling just a few, but real, mutations per cell instead of thousands of false positives. Therefore, we reversed the way we had been thinking about variant calling before - instead of trying to prove a variant was real, we would only call it if it was unlikely to be a technical error. In order to do this, as explained in the Thesis, we carefully analyzed the characteristics of different types of technical errors in single-cell RNAseq and developed ways to calculate their probability in our data. Our final caller is simple and most likely captures only a fraction of the mutations that are present in single-cells. One the other hand, its simplicity allows for efficient interpretation, and we are now able to clearly determine why each potential mutation candidate had been filtered out. Simultaneously, we achieved our initial goals – creating a method independent of other tools and not requiring any paired bulk samples.

A substantial obstacle in the development and validation of our caller was the quality of the single-cell RNAseq data. While we did take quality into consideration when choosing the datasets, it was still insufficient and ultimately only a few percent of the bulk calls would pass the basic quality thresholds in single-cells. Furthermore, a large fraction of the calls that did fulfill the quality criteria would be removed during the preparation for the linkage method. Because the method relies on relationships between neighbouring reads, it is highly sensitive to coverage and cannot be applied in regions covered by few reads, which form a majority in single-cell RNAseq data. In effect, a great proportion of mutations present in single-cells would never be called. Therefore, we decided to introduce a recalling step. By

re-visiting every cell and searching for mutations called in other single-cells from the same cohort, we were able to significantly increase the number of mutations called per single-cell. This would not only allow us to compare cells more accurately, but also made us aware of the number of SNPs not identified from the bulk tissues and, therefore, still present among our single-cell calls. The fact that the single-cells were of different types and came from various tissues was a crucial factor in the discovery of the latter, as mutations identified in multiple healthy tissue types were unlikely disease-related somatic mutations. Furthermore, the availability of the various types of single-cells allowed us to mitigate the effects of differential gene expression and cell states. The more single-cells we had, the more shared mutations we would find. This would eventually lead to a better validation of the caller (same mutations identified independently in different samples), a more comprehensive removal of germline calls and a more accurate analysis of single-cell relationships.

Single-cell RNAseq has been used mainly in transcriptomic analysis, usually to distinguish between different cell types. Therefore, the quality thresholds that are applied during sample preparation and processing are targeted specifically for this purpose. Unsurprisingly, because using RNAseq to call mutations is still a relatively novel approach, the minima required to successfully identify mutations are not taken into account. Ultimately, very few single-cell RNAseq datasets of sufficient quality exist and the opportunities for method development and validation are limited. We developed and validated our caller on two datasets. The Barrett's dataset was an obvious pick as the method was created in order to gain new insights into the origins of the disease. While cells of multiple types and collected from various tissues were included, the dataset lacked an important component – it did not have any paired DNA samples. As a result, the validation of our calls was very difficult as

we were not able to determine whether the lack of Barrett's signatures was an evidence of the failure of our caller or whether the signal had not been there in the first place. Our solution was to use a different dataset to validate our method. We chose the breast cancer dataset as not only it had paired WES samples included, but also it had been shown to be of high quality and having breast cancer characteristics identified before. However, a downside to this dataset was a different kind of spike-ins than the ones included in our Barrett's dataset, which we discovered only after the completion of the whole processing pipeline to have large amounts of variation from the reference sequences. Because even commonly used bulk callers would identify mutations in the spike-in regions, we could not confidently treat them as technical errors. Therefore, we could not learn much about the error profile of the dataset.

Because we conducted our research on just two single-cell RNAseq datasets which were of high quality, we needed to ensure our method was using strict quality thresholds in order to be applicable to a wider range of datasets in the future. We believe that further analysis of single-cell RNAseq data in the context of variant calling is required. Research into various kinds of single-cell RNAseq datasets would hopefully reveal more differences from the bulk samples, and would allow for the identification of aspects that are conserved or different between single-cell datasets. Discovery of shared features would hopefully give more insights into dataset-specific characteristics. That would allow for the elimination of strict quality thresholds and, ultimately, would lead to the identification of a larger number of mutations.

Understanding relationships between single-cells and, eventually, mechanisms of mutagenesis, requires large numbers of cells (ideally located in close proximity as this would allow for better tracking of the spread of novel mutations). We found that a few dozens of

cells coming from the same tissue (and batch) were far from sufficient, as each consecutive cell merged with our results would shift the interpretation. Specifically, apart from the addition of new mutations, we would identify that some, previously occurring in individual single-cells, would be shared with the new sample. If a sufficient number of single-cells was available, we would expect to see a decrease in the pace at which new shared mutations were found. In other words, at some point there would be enough single-cells to represent a complete mutational profile, and allow for the distinction between larger clones (mutations shared between multiple cells) and rare mutations (occurring in only one or few cells). However, estimation of the exact number of single-cells required is difficult, as it is dependent on the disease. A reconstruction of a full mutation profile in a homogenous tissue would require a smaller number of single-cells. On the other hand, the requirements increase dramatically when analyzing highly clonal diseases such as Barrett's.

In conclusion, I believe we managed to achieve a solid first step in the attempt to identify variants from single-cell RNAseq. Additional time and datasets would be required in order to develop the methodology further. Ideally, the new data would be of high quality, consist of a large number of single-cells from different tissues and paired DNA samples. It would also have high-quality spike-ins which would give better insights into the occurrence and characteristics of technical errors in the single-cell RNAseq data. Successful validation of the method would involve a reconstruction of mutation profiles present in the paired DNA, and identification of other known disease characteristics.

6. Supplement

1.13 Single-cell SNV calls from the breast cancer dataset

Table 16. Number of unique single-cell SNV calls from the breast cancer dataset, grouped by patient and cell type

Patient	Immune cells	Stromal cells	Tumour cells
BC01	0	642	21,581
BC02	0	0	44,181
BC03	22,759	0	26,638
BC04	10,638	1,492	46,577
BC05	0	1,211	46,814
BC06	4,423	1,448	4,691
BC07	34,766	9,978	47,996
BC08	711	5,350	12,872
BC09	17,344	2,180	0

1.14 Differences in WES SNV calls between Mutect2 and Octopus in the breast cancer dataset

1.14.1 Differences between the outputs of two callers not explained by confusion between germline and somatic calls

Because germline and somatic calling was performed independently, we wanted to exclude the possibility of confusion between the two. In other words, we checked that the somatic calls had not been earlier identified as germline variants by either caller, as this would explain the excess of non-shared calls. We found 6 cases of variants called as both germline and somatic, mostly by Octopus (**Table 17**). However, this was insufficient to have an impact on the differences in SNVs called by the two callers (over 400 non-shared positions in the case of patient BC03).

Table 17. Variants called as both germline and somatic.

Variant	Patient	Germline call	Somatic call
chr2:214149062:214149063:A:G	BC01	Octopus	Mutect2, Octopus
chr6:158605151:158605152:G:A	BC01	Octopus	Octopus
chr6:158605155:158605156:G:A	BC01	Octopus	Octopus
chr7:151240005:151240006:T:A	BC01	Haplotypecaller	Octopus
chr19:1457875:1457876:G:A	BC04	Octopus	Octopus
chr7:100963865:100963866:G:A	BC08	Haplotypecaller	Octopus

1.14.2 Insufficient evidence as the main reason for filtering out Octopus-specific calls by Mutect2

Both Mutect2 and Octopus consist of the main calling stage and final filtering. Investigation of calls made solely by Octopus revealed that 44.6% of them passed the variant calling performed by Mutect2 and underwent subsequent filtering. The reasons for removing the potential candidates from the final set of mutations were “weak_evidence” (35.0%), “strand_bias” (29.3%), “orientation” (17.4%), germline (6.6%), “clustered_events” (3.5%) and “haplotype”, “base_qual”, “contamination”, “normal_artifact” (less than 1.0% each). On the other hand, only 92 out of 838 (11.0%) Mutect2-specific calls passed the calling performed by Octopus. They were consequently eliminated as a result of random forest filtering (“RF” filter).

Because Mutect2 would filter out variants called by Octopus partially due to poor quality, we wondered whether it was due to different sample pre-processing. When using the “--bamout” option, Octopus generates realigned BAMs that provide visual evidence for why a call has been made. We found that the internal sample processing performed by Octopus increased the quality of variants not called by Mutect2 in terms of coverage (**Figure 37a**) and allelic frequency (**Figure 37b**). The results suggest that the sample pre-processing (read realignment performed by Octopus) could have an influence on the quality of variants, and subsequently on whether they would be considered during the calling.

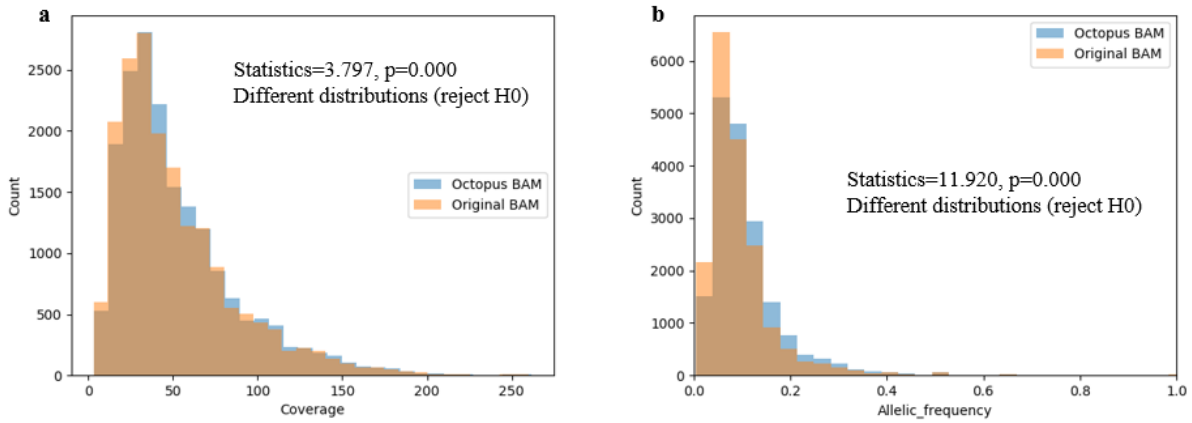


Figure 37. Comparison of characteristics of SNVs called only by Octopus in the original and Octopus-processed BAM files in terms of a. coverage and b. allelic frequency.

1.15 Mutation profiles reconstructed from the breast cancer single-cell SNV calls, grouped by patient and cell type

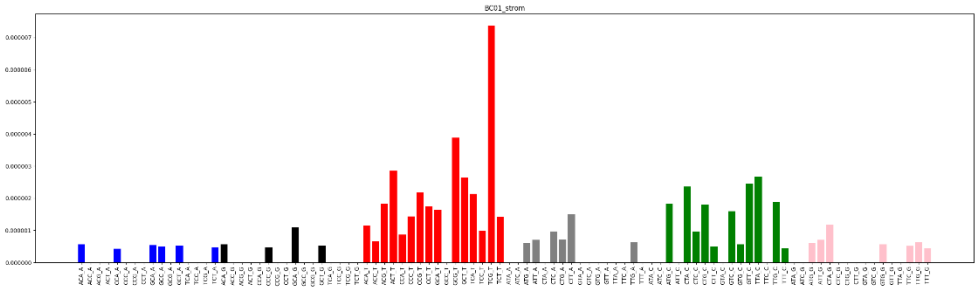


Figure 38. Mutation profile reconstructed from the breast cancer single-cell SNV calls from stromal cells of patient BC01

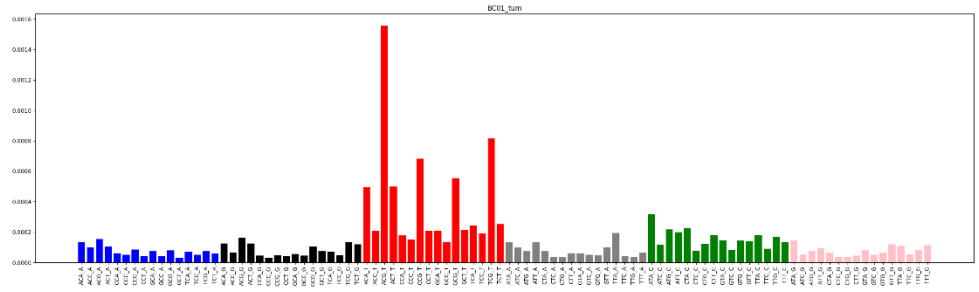


Figure 39. Mutation profile reconstructed from the breast cancer single-cell SNV calls from tumour cells of patient BC01

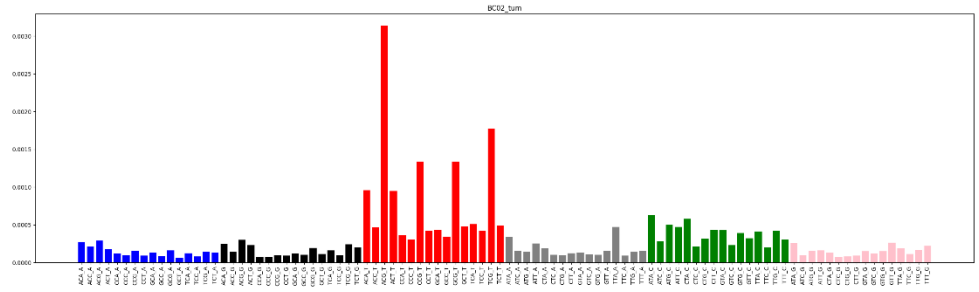


Figure 40. Mutation profile reconstructed from the breast cancer single-cell SNV calls from tumour cells of patient BC02

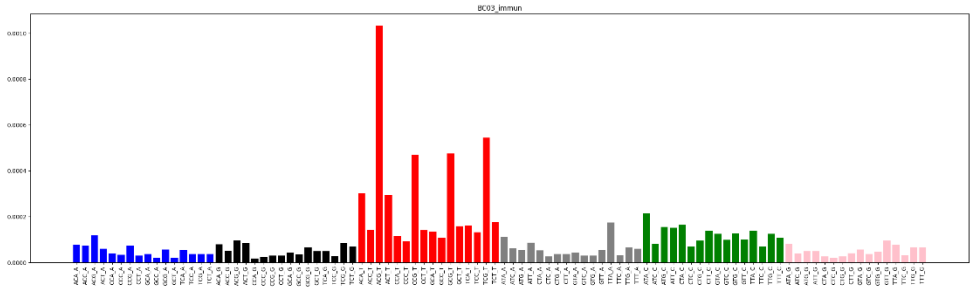


Figure 41. Mutation profile reconstructed from the breast cancer single-cell SNV calls from immune cells of patient BC03

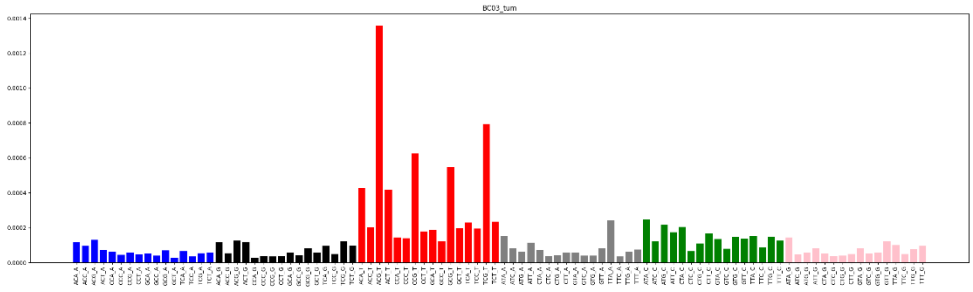


Figure 42. Mutation profile reconstructed from the breast cancer single-cell SNV calls from tumour cells of patient BC03

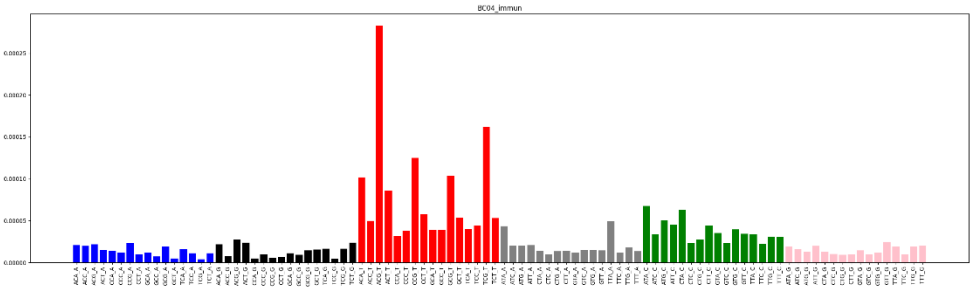


Figure 43. Mutation profile reconstructed from the breast cancer single-cell SNV calls from immune cells of patient BC04

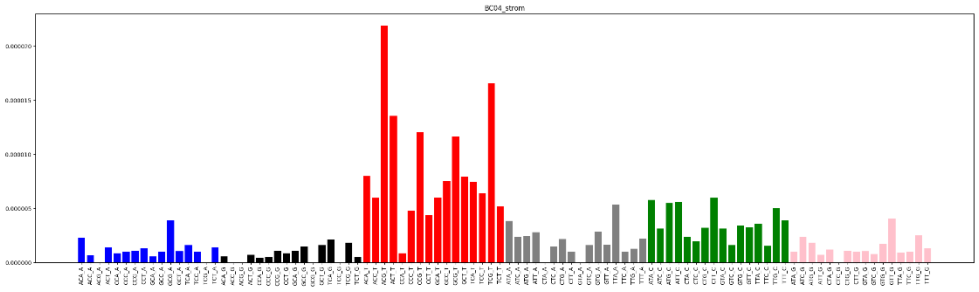


Figure 44. Mutation profile reconstructed from the breast cancer single-cell SNV calls from stromal cells of patient BC04

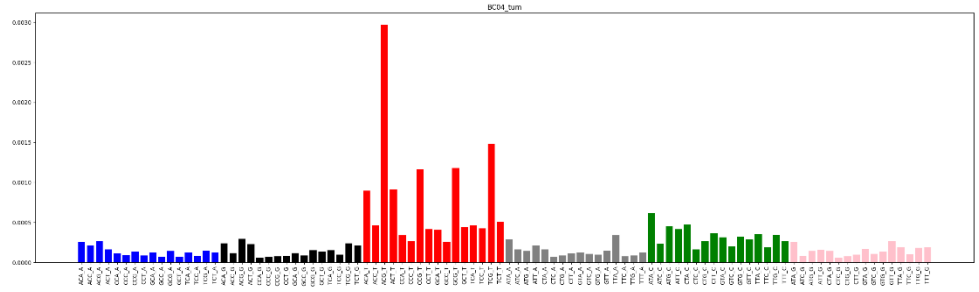


Figure 45. Mutation profile reconstructed from the breast cancer single-cell SNV calls from tumour cells of patient BC04

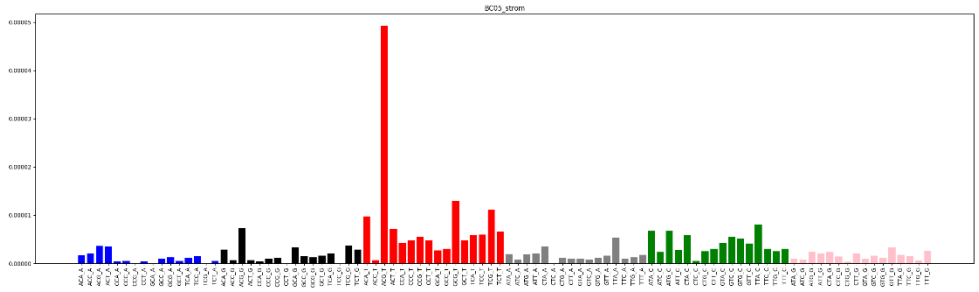


Figure 46. Mutation profile reconstructed from the breast cancer single-cell SNV calls from stromal cells of patient BC05

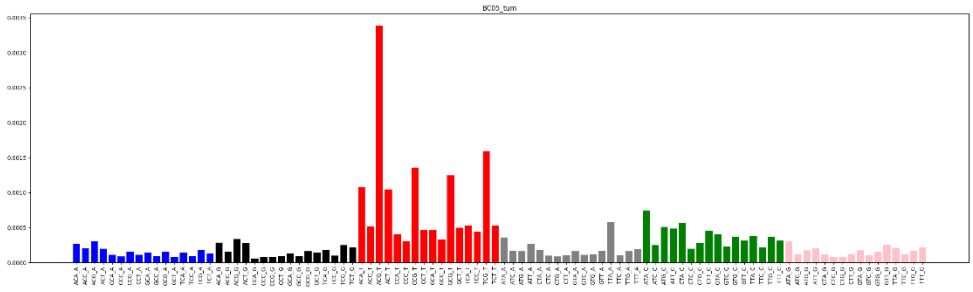


Figure 47. Mutation profile reconstructed from the breast cancer single-cell SNV calls from tumour cells of patient BC05

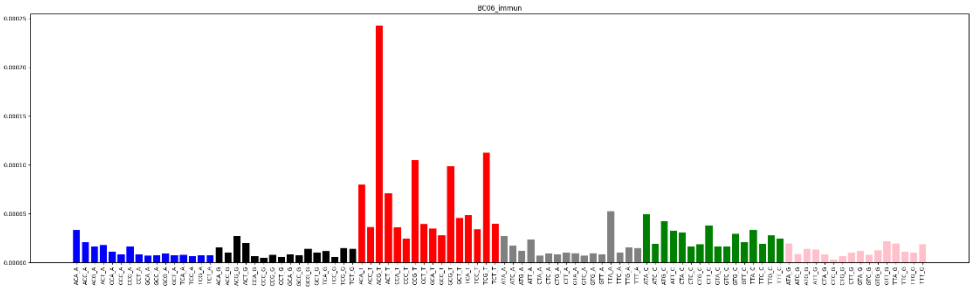


Figure 48. Mutation profile reconstructed from the breast cancer single-cell SNV calls from immune cells of patient BC06

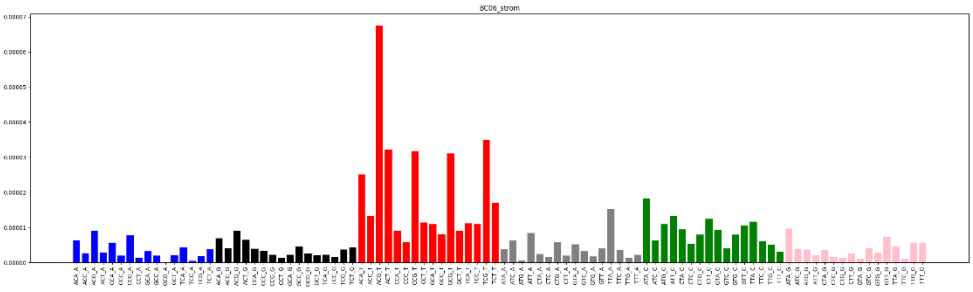


Figure 49. Mutation profile reconstructed from the breast cancer single-cell SNV calls from stromal cells of patient BC06

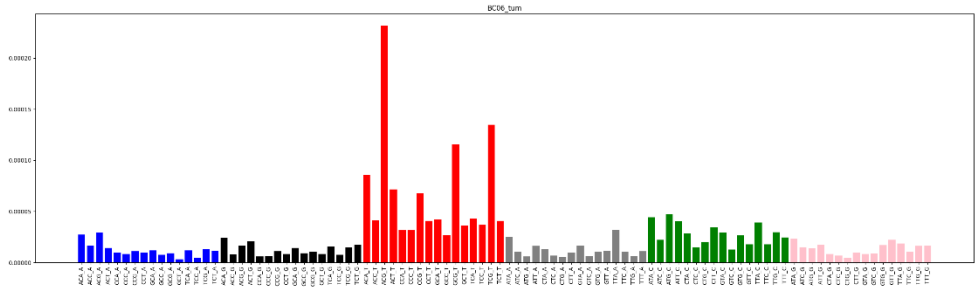


Figure 50. Mutation profile reconstructed from the breast cancer single-cell SNV calls from tumour cells of patient BC06

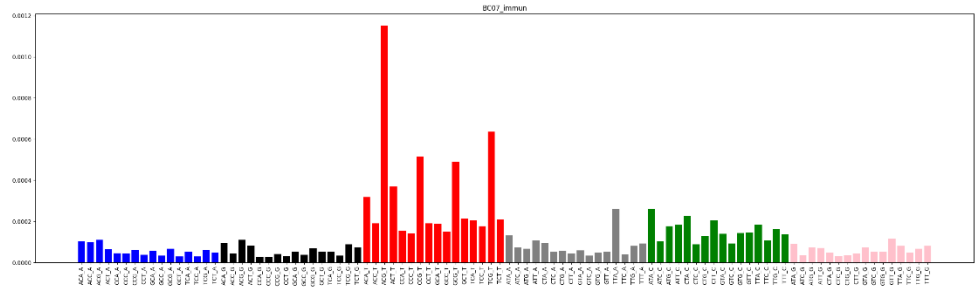


Figure 51. Mutation profile reconstructed from the breast cancer single-cell SNV calls from immune cells of patient BC07

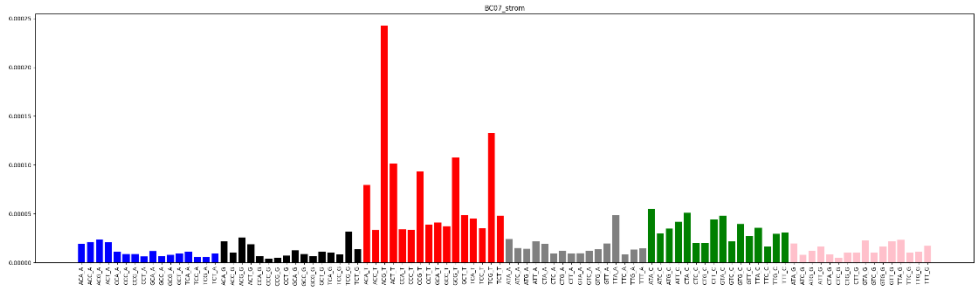


Figure 52. Mutation profile reconstructed from the breast cancer single-cell SNV calls from stromal cells of patient BC07

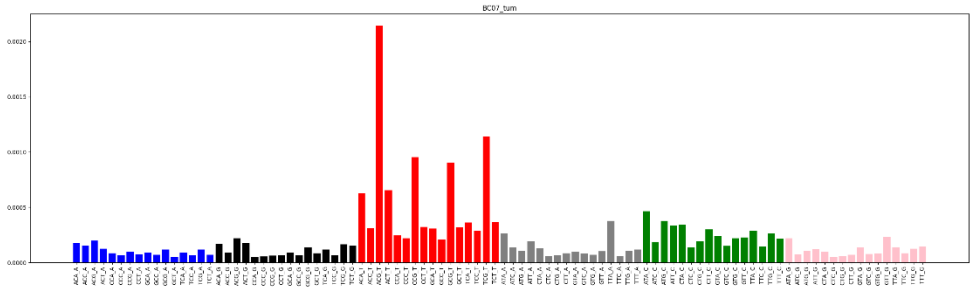


Figure 53. Mutation profile reconstructed from the breast cancer single-cell SNV calls from tumour cells of patient BC07

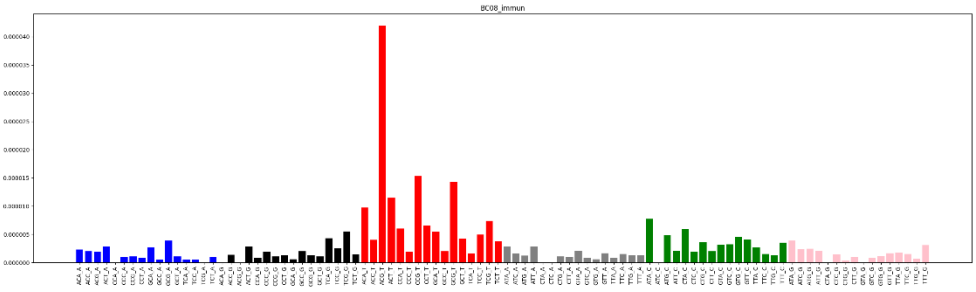


Figure 54. Mutation profile reconstructed from the breast cancer single-cell SNV calls from immune cells of patient BC08

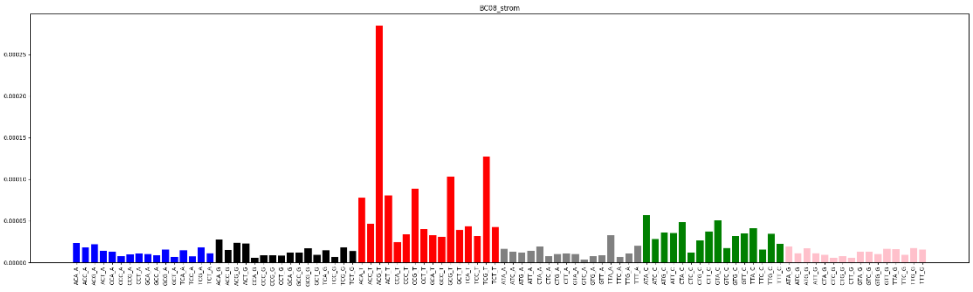


Figure 55. Mutation profile reconstructed from the breast cancer single-cell SNV calls from stromal cells of patient BC08

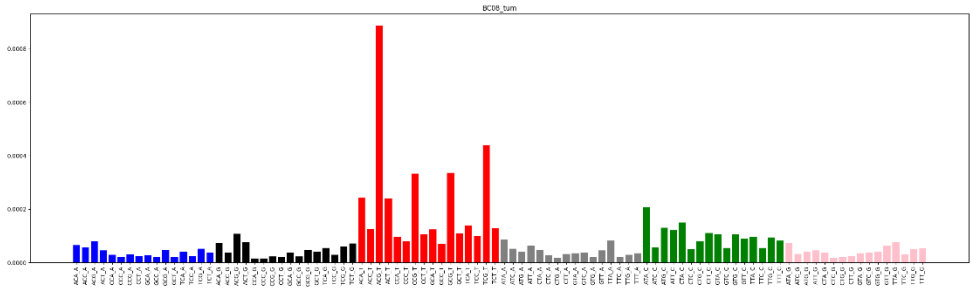


Figure 56. Mutation profile reconstructed from the breast cancer single-cell SNV calls from tumour cells of patient BC08

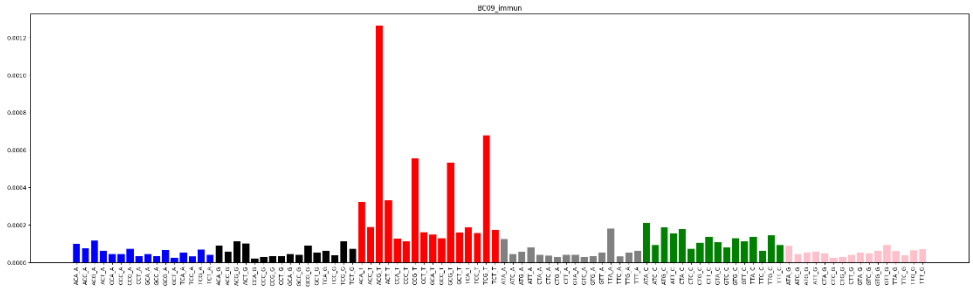


Figure 57. Mutation profile reconstructed from the breast cancer single-cell SNV calls from immune cells of patient BC09

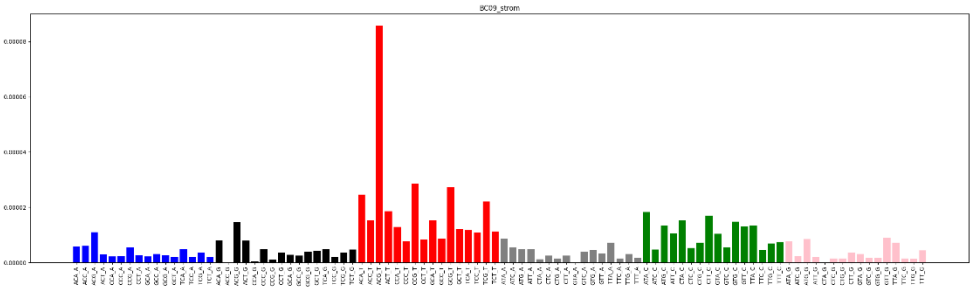


Figure 58. Mutation profile reconstructed from the breast cancer single-cell SNV calls from stromal cells of patient BC09

1.16 Single-cell SNV calls from the Barrett's dataset

Table 18. Number of unique single-cell SNV calls from the Barrett's dataset, grouped by tissue and cell type

Cell type	Total	Barrett's	Oesophagus	Gastric	Duodenum
Non-epithelial	149	46	19	38	46
Barrett's-type	721	598	130	21	0
Enterocytes	146	0	0	0	146
Mucus neck	440	3	2	435	0
Goblet	454	355	101	0	16
Squamous	101	0	101	0	0
Enteroendocrine	370	211	20	150	8

Table 19. Number of unique single-cell SNV calls from the Barrett's dataset, grouped by tissue and cell type, in patient GEN02021

Cell type	Barrett's	Oesophagus	Gastric
Non-epithelial	35	8	15
Barrett's-type	497	130	21
Enterocytes	0	0	0
Mucus neck	3	2	115
Goblet	284	89	0
Squamous	0	84	0
Enteroendocrine	196	20	37

Table 20. Number of unique single-cell SNV calls from the Barrett's dataset, grouped by tissue and cell type, in patient GEN02023

Cell type	Barrett's	Oesophagus	Gastric
Non-epithelial	2	9	0
Barrett's-type	50	1	0
Enterocytes	0	0	0
Mucus neck	0	0	2
Goblet	35	13	0
Squamous	0	1	0
Enteroendocrine	12	0	0

Table 21. Number of unique single-cell SNV calls from the Barrett's dataset, grouped by tissue and cell type, in patient GEN02024

Cell type	Barrett's	Oesophagus	Gastric
Non-epithelial	7	0	14
Barrett's-type	26	0	0
Enterocytes	0	0	0
Mucus neck	0	0	84
Goblet	28	0	0
Squamous	0	0	0
Enteroendocrine	1	0	101

Table 22. Number of unique single-cell SNV calls from the Barrett's dataset, grouped by tissue and cell type, in patient GEN02025

Cell type	Barrett's	Oesophagus	Gastric
Non-epithelial	1	2	9
Barrett's-type	25	1	0
Enterocytes	0	0	0
Mucus neck	0	0	233
Goblet	8	1	0
Squamous	0	16	0
Enteroendocrine	4	0	11

Table 23. Number of unique single-cell SNV calls from the Barrett's dataset, grouped by tissue subtype and patient

Tissue subtype	GEN02021	GEN02023	GEN02024	GEN02025
B1	487	64	26	21
B2	261	7	27	8
B3	69	15	16	6
B4	217	12	1	5
G1	12	0	0	136
G2	151	2	97	124
G3	26	0	101	11
O1	84	0	0	0
O2	0	1	0	16
O3	131	1	0	1
O4	108	22	0	3

1.17 Overlap of Barrett's and OSG variants

Table 24. Single-cell SNVs identified in Barrett's-type and OSG cells

Chromosome	Position [0-offset]	Position [1-offset]	Allele
chr1	19282709	19282710	A
chr2	109387344	109387345	A
chr2	175419149	175419150	C
chr2	238239951	238239952	G
chr3	136819348	136819349	C
chr3	178518786	178518787	G
chr3	187669659	187669660	G
chr3	187670122	187670123	A
chr3	42177959	42177960	G
chr3	42178632	42178633	C
chr3	62428458	62428459	T
chr3	62508932	62508933	T
chr3	62509003	62509004	T
chr3	62579958	62579959	G
chr3	62718205	62718206	G
chr4	139866540	139866541	T
chr4	16043742	16043743	T
chr4	56188685	56188686	G
chr4	91303681	91303682	A
chr4	9601276	9601277	C
chr4	9602144	9602145	T
chr6	1810558	1810559	G
chr6	325960	325961	C
chr6	326133	326134	C
chr6	39067219	39067220	A
chr6	39067355	39067356	A
chr7	101020572	101020573	T
chr7	151094164	151094165	C
chr7	157545671	157545672	T
chr7	32206443	32206444	T
chr7	75813536	75813537	T
chr7	8239535	8239536	G
chr8	124033990	124033991	A
chr8	12575551	12575552	A
chr8	12587750	12587751	G
chr8	41697462	41697463	T
chr8	72632092	72632093	G
chr8	72654225	72654226	A
chr8	72680405	72680406	T
chr8	72693444	72693445	G
chr8	72711407	72711408	C
chr8	72859987	72859988	T
chr8	72860984	72860985	A
chr9	121876654	121876655	T
chr9	135738596	135738597	G

chr11	1024913	1024914	T
chr12	120199456	120199457	G
chr12	130476189	130476190	T
chr15	43793758	43793759	A
chr16	16170240	16170241	T
chr16	16170558	16170559	G
chr16	976397	976398	G
chrX	79362614	79362615	C

1.18 Mutation profiles reconstructed from Barrett's single-cell SNV calls, grouped by patient and tissue type

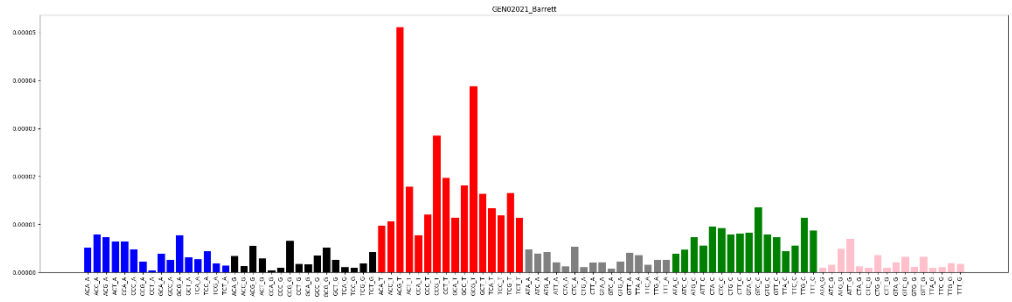


Figure 59. Mutation profile reconstructed from the Barrett's single-cell SNV calls from Barrett's tissue of patient GEN02021

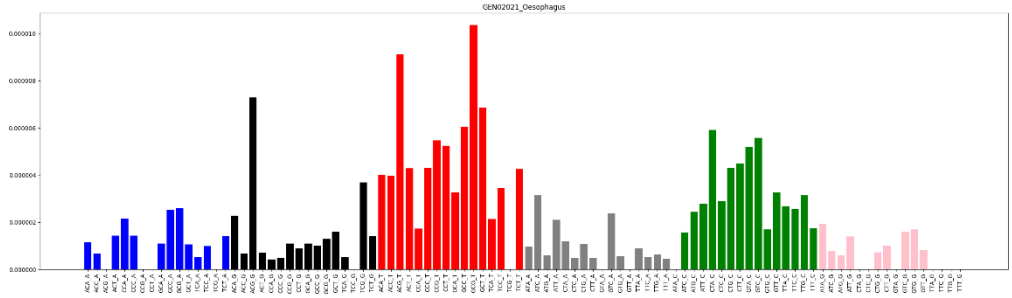


Figure 60. Mutation profile reconstructed from the Barrett's single-cell SNV calls from Oesophagus tissue of patient GEN02021

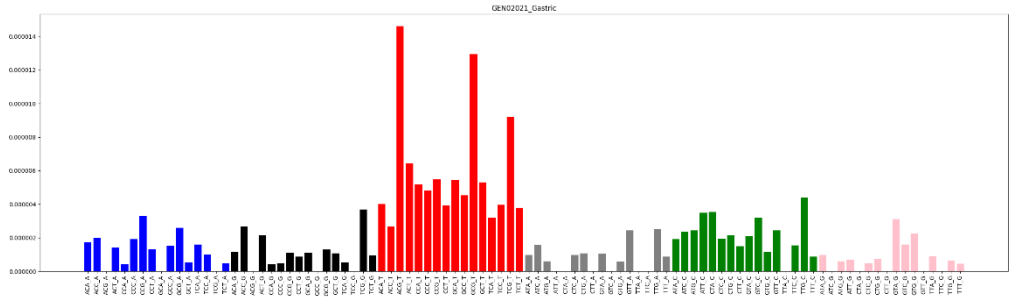


Figure 61. Mutation profile reconstructed from the Barrett's single-cell SNV calls from Gastric tissue of patient GEN02021

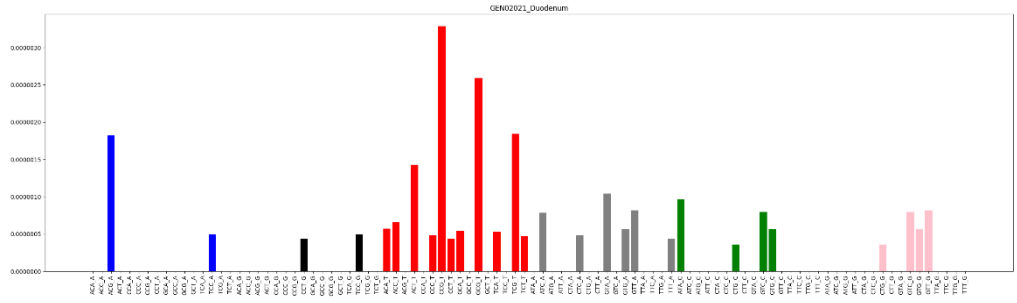


Figure 62. Mutation profile reconstructed from the Barrett's single-cell SNV calls from Duodenum tissue of patient GEN02021

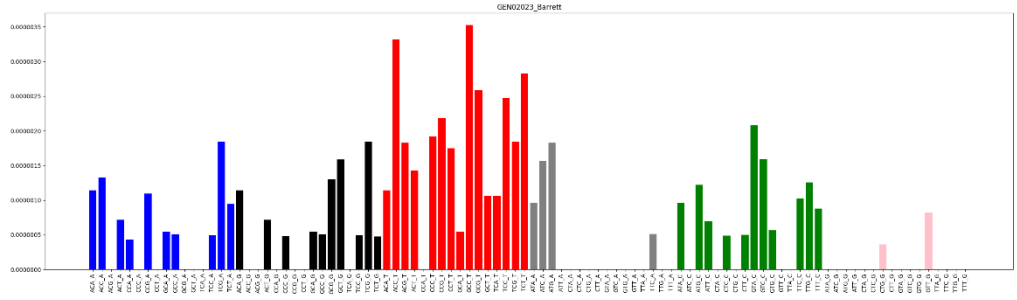


Figure 63. Mutation profile reconstructed from the Barrett's single-cell SNV calls from Barrett's tissue of patient GEN02023

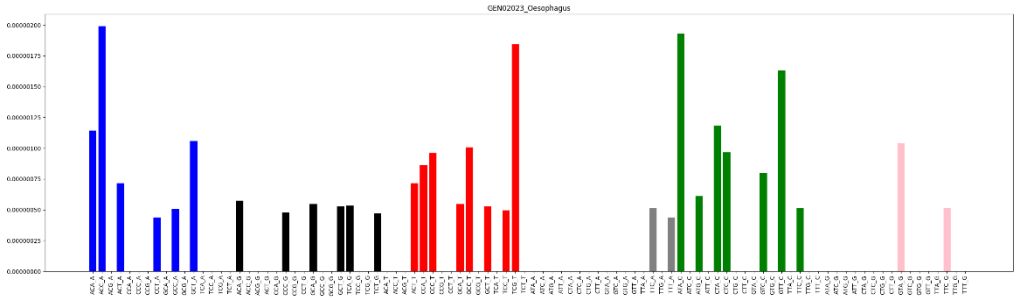


Figure 64. Mutation profile reconstructed from the Barrett's single-cell SNV calls from Oesophagus tissue of patient GEN02023

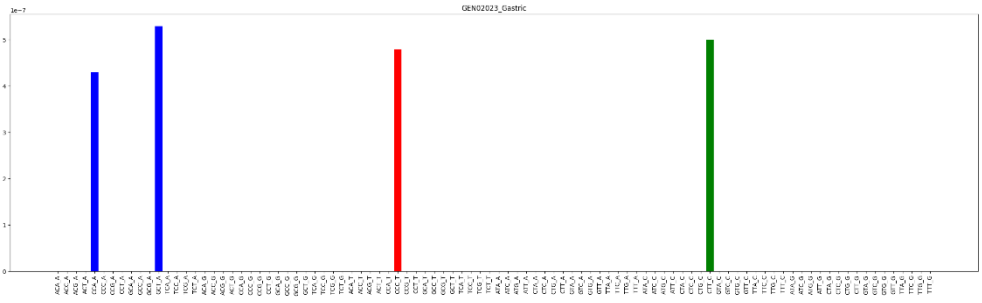


Figure 65. Mutation profile reconstructed from the Barrett's single-cell SNV calls from Gastric tissue of patient GEN02023

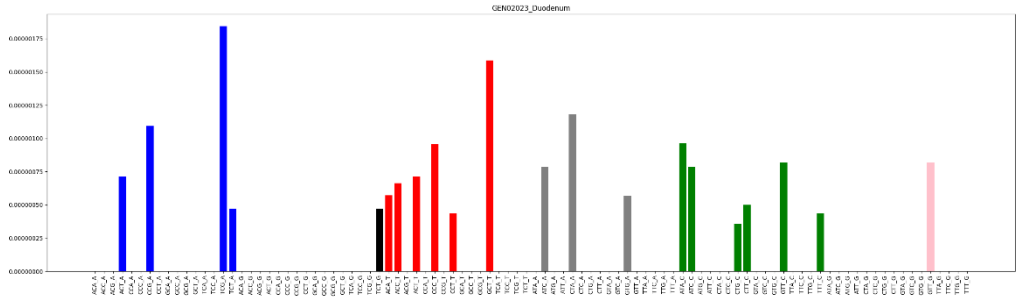


Figure 66. Mutation profile reconstructed from the Barrett's single-cell SNV calls from Duodenum tissue of patient GEN02023

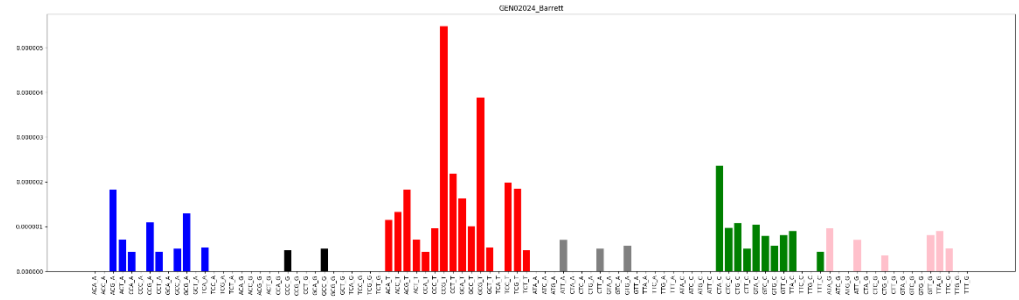


Figure 67. Mutation profile reconstructed from the Barrett's single-cell SNV calls from Barrett's tissue of patient GEN02024

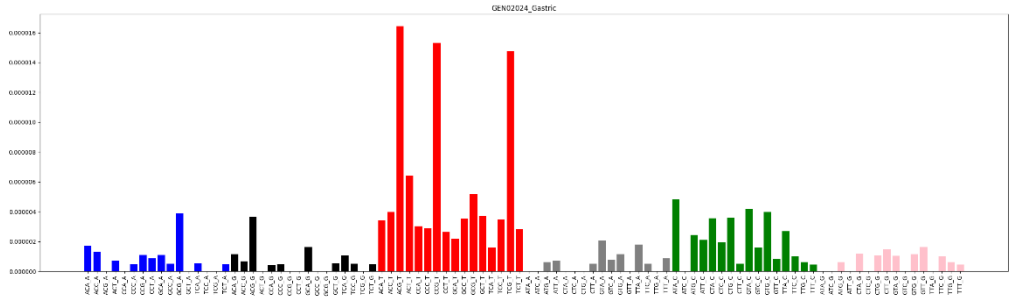


Figure 68. Mutation profile reconstructed from the Barrett's single-cell SNV calls from Gastric tissue of patient GEN02024

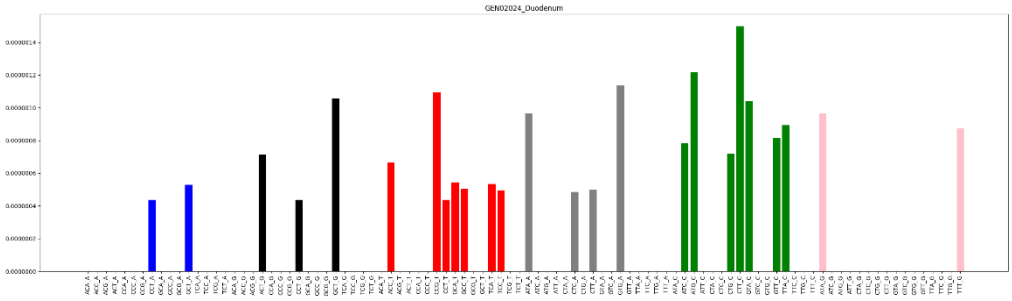


Figure 69. Mutation profile reconstructed from the Barrett's single-cell SNV calls from Duodenum tissue of patient GEN02024

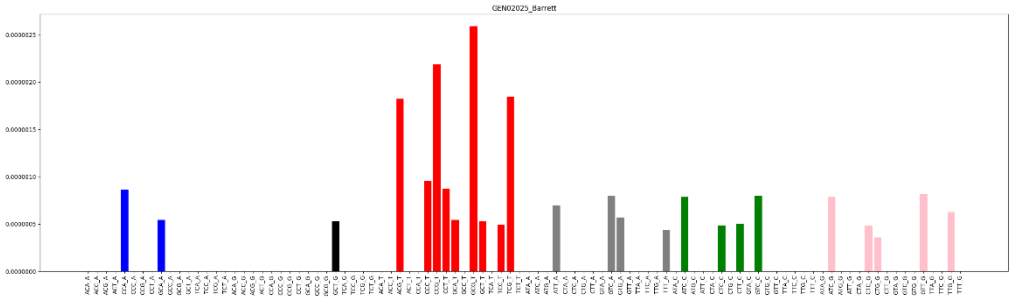


Figure 70. Mutation profile reconstructed from the Barrett's single-cell SNV calls from Barrett's tissue of patient GEN02025

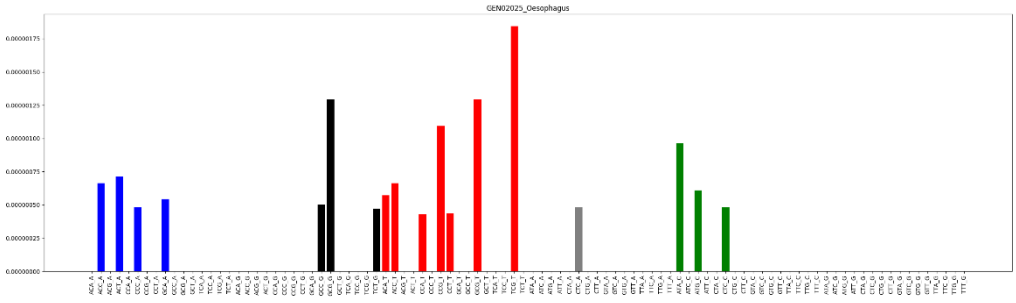


Figure 71. Mutation profile reconstructed from the Barrett's single-cell SNV calls from Oesophagus tissue of patient GEN02025

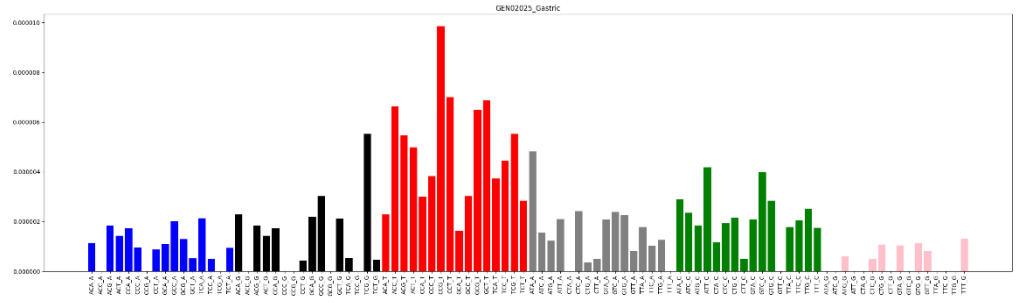


Figure 72. Mutation profile reconstructed from the Barrett's single-cell SNV calls from Gastric tissue of patient GEN02025

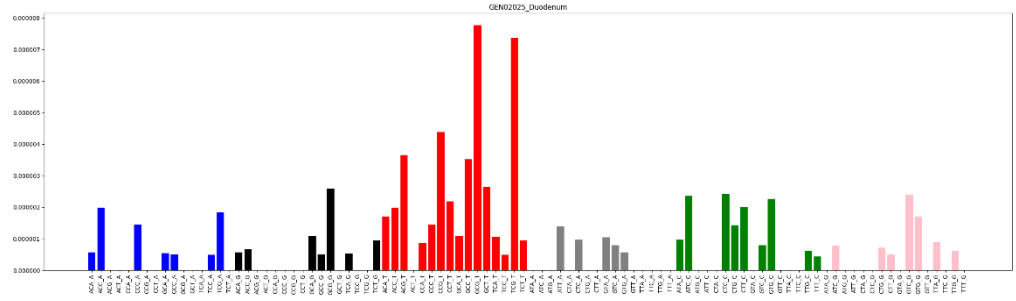


Figure 73. Mutation profile reconstructed from the Barrett's single-cell SNV calls from Duodenum tissue of patient GEN02025

1.19 SNVs shared between different cell types in the Barrett's dataset

We found a number of SNVs that were shared between different cell types. Apart from comparing all cells, we grouped them by tissue type.

Table 25. SNV shared between different cell types in the Barrett's dataset

	Non-epithelial	Barrett's-type	Enterocytes	Mucus neck	Goblet	Squamous	Enteroendocrine
Non-epithelial	-	-	-	-	-	-	-
Barrett's-type	30	-	-	-	-	-	-
Enterocytes	4	13	-	-	-	-	-
Mucus neck	7	48	0	-	-	-	-
Goblet	19	108	4	26	-	-	-
Squamous	2	15	0	0	7	-	-
Enteroendocrine	9	56	6	19	60	3	-

Table 26. SNV shared between different cell types in the Barrett's dataset in Barrett's tissue

Barrett's	Non-epithelial	Barrett's-type	Enterocytes	Mucus neck	Goblet	Squamous	Enteroendocrine
Non-epithelial	-	-	-	-	-	-	-
Barrett's-type	15	-	-	-	-	-	-
Enterocytes	0	0	-	-	-	-	-
Mucus neck	0	0	0	-	-	-	-
Goblet	12	77	0	1	-	-	-
Squamous	0	0	0	0	0	-	-
Enteroendocrine	2	43	0	0	40	0	-

Table 27. SNV shared between different cell types in the Barrett's dataset in Oesophagus tissue

Oesophagus	Non-epithelial	Barrett's-type	Enterocytes	Mucus neck	Goblet	Squamous	Enteroendocrine
Non-epithelial	-	-	-	-	-	-	-
Barrett's-type	2	-	-	-	-	-	-
Enterocytes	0	0	-	-	-	-	-
Mucus neck	0	2	0	-	-	-	-
Goblet	1	27	0	2	-	-	-
Squamous	1	4	0	0	0	-	-
Enteroendocrine	1	4	0	0	8	1	-

Table 28. SNV shared between different cell types in the Barrett's dataset in Gastric tissue

Gastric	Non-epithelial	Barrett's-type	Enterocytes	Mucus neck	Goblet	Squamous	Enteroendocrine
Non-epithelial	-	-	-	-	-	-	-
Barrett's-type	0	-	-	-	-	-	-
Enterocytes	0	0	-	-	-	-	-
Mucus neck	2	2	0	-	-	-	-
Goblet	0	0	0	0	-	-	-
Squamous	0	0	0	0	0	-	-
Enteroendocrine	1	0	0	10	0	0	-

1.20 Other software used

- Python 2.7.5 (Python documentation)
- Samtools 1.9 (Li et al., 2009)
- Bcftools 1.9 (Danecek et al., 2021)
- Vcftools 0.1.14 (Danecek et al., 2011)
- Bedtools 0.26.0 (Quinlan and Hall, 2010)
- FastQC 0.11.5 (LaMar, 2015)

7. Bibliography

3.11.0 Documentation [WWW Document], n.d. URL https://docs.python.org/3/ (accessed 11.18.22).

Agrawal, N., Jiao, Y., Bettegowda, C., Hutfless, S.M., Wang, Y., David, S., Cheng, Y., Twaddell, W.S., Latt, N.L., Shin, E.J., Wang, L.-D., Wang, L., Yang, W., Velculescu, V.E., Vogelstein, B., Papadopoulos, N., Kinzler, K.W., Meltzer, S.J., 2012. Comparative Genomic Analysis of Esophageal Adenocarcinoma and Squamous Cell Carcinoma. Cancer Discov. 2, 899–905. https://doi.org/10.1158/2159-8290.CD-12-0189

Alser, M., Rotman, J., Deshpande, D., Taraszka, K., Shi, H., Baykal, P.I., Yang, H.T., Xue, V., Knyazev, S., Singer, B.D., Balliu, B., Koslicki, D., Skums, P., Zelikovsky, A., Alkan, C., Mutlu, O., Mangul, S., 2021. Technology dictates algorithms: recent developments in read alignment. Genome Biol. 22, 249. https://doi.org/10.1186/s13059-021-02443-7

Bae, T., Tomasini, L., Mariani, J., Zhou, B., Roychowdhury, T., Franjic, D., Pletikos, M., Pattni, R., Chen, B.-J., Venturini, E., Riley-Gillis, B., Sestan, N., Urban, A.E., Abyzov, A., Vaccarino, F.M., 2018. Different mutational rates and mechanisms in human cells at pregastrulation and neurogenesis. Science 359, 550–555. https://doi.org/10.1126/science.aan8690

Baer, C.F., Miyamoto, M.M., Denver, D.R., 2007. Mutation rate variation in multicellular eukaryotes: causes and consequences. Nat. Rev. Genet. 8, 619–631. https://doi.org/10.1038/nrg2158

Balázs, Z., Tombácz, D., Csabai, Z., Moldován, N., Snyder, M., Boldogkői, Z., 2019. Template-switching artifacts resemble alternative polyadenylation. BMC Genomics 20, 824. https://doi.org/10.1186/s12864-019-6199-7

Barrett, N.R., 1950. Chronic peptic ulcerz of the œophagus and 'œsophagitis.' BJS Br. J. Surg. 38, 175–182. https://doi.org/10.1002/bjs.18003815005

Bremner, C.G., Lynch, V.P., Ellis, F.H., 1970. Barrett's esophagus: congenital or acquired? An experimental study of esophageal mucosal regeneration in the dog. Surgery 68, 209–216.

Brouard, J.-S., Schenkel, F., Marete, A., Bissonnette, N., 2019. The GATK joint genotyping workflow is appropriate for calling variants in RNA-seq experiments. J. Anim. Sci. Biotechnol. 10, 44. https://doi.org/10.1186/s40104-019-0359-0

Brunner, S.F., Roberts, N.D., Wylie, L.A., Moore, L., Aitken, S.J., Davies, S.E., Sanders, M.A., Ellis, P., Alder, C., Hooks, Y., Abascal, F., Stratton, M.R., Martincorena, I., Hoare, M., Campbell, P.J., 2019. Somatic mutations and clonal dynamics in healthy and cirrhotic human liver. Nature 574, 538–542. https://doi.org/10.1038/s41586-019-1670-9

Callari, M., Sammut, S.-J., De Mattos-Arruda, L., Bruna, A., Rueda, O.M., Chin, S.-F., Caldas, C., 2017. Intersect-then-combine approach: improving the performance of somatic variant calling in whole exome sequencing data using multiple aligners and callers. Genome Med. 9, 35. https://doi.org/10.1186/s13073-017-0425-1

Chen, G., Ning, B., Shi, T., 2019. Single-Cell RNA-Seq Technologies and Related Computational Data Analysis. Front. Genet. 10. https://doi.org/10.3389/fgene.2019.00317

Chen, Z., Yuan, Y., Chen, X., Chen, J., Lin, S., Li, X., Du, H., 2020. Systematic comparison of somatic variant calling performance among different sequencing depth and mutation frequency. Sci. Rep. 10, 3501. https://doi.org/10.1038/s41598-020-60559-5

Chung, W., Eum, H.H., Lee, H.-O., Lee, K.-M., Lee, H.-B., Kim, K.-T., Ryu, H.S., Kim, S., Lee, J.E., Park, Y.H., Kan, Z., Han, W., Park, W.-Y., 2017. Single-cell RNA-seq enables comprehensive tumour and immune cell profiling in primary breast cancer. Nat. Commun. 8, 1–12. https://doi.org/10.1038/ncomms15081

Cibulskis, K., Lawrence, M.S., Carter, S.L., Sivachenko, A., Jaffe, D., Sougnez, C., Gabriel, S., Meyerson, M., Lander, E.S., Getz, G., 2013. Sensitive detection of somatic point mutations in impure and

heterogeneous cancer samples. Nat. Biotechnol. 31, 213–219. https://doi.org/10.1038/nbt.2514

Coad, R.A., Woodman, A.C., Warner, P.J., Barr, H., Wright, N.A., Shepherd, N.A., 2005. On the histogenesis of Barrett's oesophagus and its associated squamous islands: a three-dimensional study of their morphological relationship with native oesophageal gland ducts. J. Pathol. 206, 388–394. https://doi.org/10.1002/path.1804

Collins, A.R., Cadet, J., Möller, L., Poulsen, H.E., Viña, J., 2004. Are we sure we know how to measure 8-oxo-7,8-dihydroguanine in DNA from human cells? Arch. Biochem. Biophys. 423, 57–65. https://doi.org/10.1016/j.abb.2003.12.022

Conrad, D.F., Keebler, J.E.M., DePristo, M.A., Lindsay, S.J., Zhang, Y., Cassals, F., Idaghdour, Y., Hartl, C.L., Torroja, C., Garimella, K.V., Zilversmit, M., Cartwright, R., Rouleau, G., Daly, M., Stone, E.A., Hurles, M.E., Awadalla, P., n.d. Variation in genome-wide mutation rates within and between human families. Nat. Genet. 43, 712. https://doi.org/10.1038/ng.862

Cooke, D.P., Wedge, D.C., Lunter, G., 2018. A unified haplotype-based method for accurate and comprehensive variant calling. bioRxiv 456103. https://doi.org/10.1101/456103

Corbett, S., Courtiol, A., Lummaa, V., Moorad, J., Stearns, S., 2018. The transition to modernity and chronic disease: mismatch and natural selection. Nat. Rev. Genet. 19, 419–430. https://doi.org/10.1038/s41576-018-0012-3

Cornish, A., Guda, C., 2015. A Comparison of Variant Calling Pipelines Using Genome in a Bottle as a Reference. BioMed Res. Int. 2015, 456479. https://doi.org/10.1155/2015/456479

Cornish, A., Roychoudhury, S., Sarma, K., Pramanik, S., Bhakat, K., Dudley, A., Mishra, N.K., Guda, C., 2020. Red Panda: A novel method for detecting variants in single-cell RNA sequencing (preprint). Bioinformatics. https://doi.org/10.1101/2020.01.08.898874

Crick, F., 1970. Central dogma of molecular biology. Nature 227, 561–563. https://doi.org/10.1038/227561a0

Dabney, J., Meyer, M., 2012. Length and GC-biases during sequencing library amplification: A comparison of various polymerase-buffer systems with ancient and modern DNA sequencing libraries. BioTechniques 52, 87–94. https://doi.org/10.2144/000113809

Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T., Sherry, S.T., McVean, G., Durbin, R., 1000 Genomes Project Analysis Group, 2011. The variant call format and VCFtools. Bioinformatics 27, 2156–2158. https://doi.org/10.1093/bioinformatics/btr330

Danecek, P., Bonfield, J.K., Liddle, J., Marshall, J., Ohan, V., Pollard, M.O., Whitwham, A., Keane, T., McCarthy, S.A., Davies, R.M., Li, H., 2021. Twelve years of SAMtools and BCFtools. GigaScience 10, giab008. https://doi.org/10.1093/gigascience/giab008

De Cario, R., Kura, A., Suraci, S., Magi, A., Volta, A., Marcucci, R., Gori, A.M., Pepe, G., Giusti, B., Sticchi, E., 2020. Sanger Validation of High-Throughput Sequencing in Genetic Diagnosis: Still the Best Practice? Front. Genet. 11.

De, S., 2011. Somatic mosaicism in healthy human tissues. Trends Genet. 27, 217–223. https://doi.org/10.1016/j.tig.2011.03.002

Dou, Y., Gold, H.D., Luquette, L.J., Park, P.J., 2018. Detecting Somatic Mutations in Normal Cells. Trends Genet. 34, 545–557. https://doi.org/10.1016/j.tig.2018.04.003

Fang, L.T., Afshar, P.T., Chhibber, A., Mohiyuddin, M., Fan, Y., Mu, J.C., Gibeling, G., Barr, S., Asadi, N.B., Gerstein, M.B., Koboldt, D.C., Wang, W., Wong, W.H., Lam, H.Y.K., 2015. An ensemble approach to accurately detect somatic mutations using SomaticSeq. Genome Biol. 16, 197. https://doi.org/10.1186/s13059-015-0758-2

Fangal, V.D., n.d. CTAT Mutations: A Machine Learning Based RNA-seq Variant Calling Pipeline Incorporating 53.

Galipeau, P.C., Oman, K.M., Paulson, T.G., Sanchez, C.A., Zhang, Q., Marty, J.A., Delrow, J.J., Kuhner, M.K., Vaughan, T.L., Reid, B.J., Li, X., 2018. NSAID use and somatic exomic mutations in Barrett's esophagus. Genome Med. 10, 17. https://doi.org/10.1186/s13073-018-0520-y

García-Nieto, P.E., Morrison, A.J., Fraser, H.B., 2019. The somatic mutation landscape of the human body. Genome Biol. 20, 298. https://doi.org/10.1186/s13059-019-1919-5

Garrison, E., Marth, G., 2012. Haplotype-based variant detection from short-read sequencing. https://doi.org/10.48550/arXiv.1207.3907

Geboes, K., Hoorens, A., 2021. The cell of origin for Barrett's esophagus. Science 373, 737–738. https://doi.org/10.1126/science.abj9797

Glickman, J.N., Yang, A., Shahsafaei, A., McKeon, F., Odze, R.D., 2001. Expression of p53-related protein p63 in the gastrointestinal tract and in esophageal metaplastic and neoplastic disorders. Hum. Pathol. 32, 1157–1165. https://doi.org/10.1053/hupa.2001.28951

Gout, J.-F., Li, W., Fritsch, C., Li, A., Haroon, S., Singh, L., Hua, D., Fazelinia, H., Smith, Z., Seeholzer, S., Thomas, K., Lynch, M., Vermulst, M., 2017. The landscape of transcription errors in eukaryotic cells. Sci. Adv. 3, e1701484. https://doi.org/10.1126/sciadv.1701484

Haggitt, R.C., 1994. Barrett's esophagus, dysplasia, and adenocarcinoma. Hum. Pathol. 25, 982–993. https://doi.org/10.1016/0046-8177(94)90057-4

Hanahan, D., Weinberg, R.A., 2000. The Hallmarks of Cancer. Cell 100, 57–70. https://doi.org/10.1016/S0092-8674(00)81683-9

Haque, A., Engel, J., Teichmann, S.A., Lönnberg, T., 2017. A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. Genome Med. 9, 75. https://doi.org/10.1186/s13073-017-0467-4

Hiltemann, S., Jenster, G., Trapman, J., Spek, P. van der, Stubbs, A., 2015. Discriminating somatic and germline mutations in tumor DNA samples without matching normals. Genome Res. 25, 1382–1390. https://doi.org/10.1101/gr.183053.114

Hoeijmakers, J.H.J., 2001. Genome maintenance mechanisms for preventing cancer. Nature 411, 366–374. https://doi.org/10.1038/35077232

Hu, P., Zhang, W., Xin, H., Deng, G., 2016. Single Cell Isolation and Analysis. Front. Cell Dev. Biol. 4, 116. https://doi.org/10.3389/fcell.2016.00116

Hu, Y., Williams, V.A., Gellersen, O., Jones, C., Watson, T.J., Peters, J.H., 2007. The Pathogenesis of Barrett's Esophagus: Secondary Bile Acids Upregulate Intestinal Differentiation Factor CDX2 Expression in Esophageal Cells. J. Gastrointest. Surg. 11, 827–834. https://doi.org/10.1007/s11605-007-0174-3

Hutchinson, L., Stenstrom, B., Chen, D., Piperdi, B., Levey, S., Lyle, S., Wang, T.C., Houghton, J., 2011. Human Barrett's Adenocarcinoma of the Esophagus, Associated Myofibroblasts, and Endothelium Can Arise from Bone Marrow-Derived Cells After Allogeneic Stem Cell Transplant. Stem Cells Dev. 20, 11–17. https://doi.org/10.1089/scd.2010.0139

Jin, R.U., Mills, J.C., 2018. Are Gastric and Esophageal Metaplasia Relatives? The Case for Barrett's Stemming from SPEM. Dig. Dis. Sci. 63, 2028–2041. https://doi.org/10.1007/s10620-018-5150-0

Kim, J., Eberwine, J., 2010. RNA: state memory and mediator of cellular phenotype. Trends Cell Biol. 20, 311–318. https://doi.org/10.1016/j.tcb.2010.03.003

Kivioja, T., Vähärautio, A., Karlsson, K., Bonke, M., Enge, M., Linnarsson, S., Taipale, J., 2011. Counting absolute numbers of molecules using unique molecular identifiers. Nat. Methods 9, 72–74. https://doi.org/10.1038/nmeth.1778

Koboldt, D.C., 2020. Best practices for variant calling in clinical sequencing. Genome Med. 12, 91. https://doi.org/10.1186/s13073-020-00791-w

Koboldt, D.C., Zhang, Q., Larson, D.E., Shen, D., McLellan, M.D., Lin, L., Miller, C.A., Mardis, E.R., Ding, L., Wilson, R.K., 2012. VarScan 2: somatic mutation and copy number alteration discovery in

cancer by exome sequencing. Genome Res. 22, 568–576. https://doi.org/10.1101/gr.129684.111

Krøigård, A.B., Thomassen, M., Lænkholm, A.-V., Kruse, T.A., Larsen, M.J., 2016. Evaluation of Nine Somatic Variant Callers for Detection of Somatic Mutations in Exome and Targeted Deep Sequencing Data. PLOS ONE 11, e0151664. https://doi.org/10.1371/journal.pone.0151664

Kukurba, K.R., Montgomery, S.B., 2015. RNA Sequencing and Analysis. Cold Spring Harb. Protoc. 2015, 951–969. https://doi.org/10.1101/pdb.top084970

Lahens, N.F., Kavakli, I.H., Zhang, R., Hayer, K., Black, M.B., Dueck, H., Pizarro, A., Kim, J., Irizarry, R., Thomas, R.S., Grant, G.R., Hogenesch, J.B., 2014. IVT-seq reveals extreme bias in RNA sequencing. Genome Biol. 15, R86. https://doi.org/10.1186/gb-2014-15-6-r86

Lähnemann, D., Köster, J., Szczurek, E., McCarthy, D.J., Hicks, S.C., Robinson, M.D., Vallejos, C.A., Campbell, K.R., Beerenwinkel, N., Mahfouz, A., Pinello, L., Skums, P., Stamatakis, A., Attolini, C.S.-O., Aparicio, S., Baaijens, J., Balvert, M., Barbanson, B. de, Cappuccio, A., Corleone, G., Dutilh, B.E., Florescu, M., Guryev, V., Holmer, R., Jahn, K., Lobo, T.J., Keizer, E.M., Khatri, I., Kielbasa, S.M., Korbel, J.O., Kozlov, A.M., Kuo, T.-H., Lelieveldt, B.P.F., Mandoiu, I.I., Marioni, J.C., Marschall, T., Mölder, F., Niknejad, A., Raczkowski, L., Reinders, M., Ridder, J. de, Saliba, A.-E., Somarakis, A., Stegle, O., Theis, F.J., Yang, H., Zelikovsky, A., McHardy, A.C., Raphael, B.J., Shah, S.P., Schönhuth, A., 2020. Eleven grand challenges in single-cell data science. Genome Biol. 21, 31. https://doi.org/10.1186/s13059-020-1926-6

Lai, Z., Markovets, A., Ahdesmaki, M., Chapman, B., Hofmann, O., McEwen, R., Johnson, J., Dougherty, B., Barrett, J.C., Dry, J.R., 2016. VarDict: a novel and versatile variant caller for next-generation sequencing in cancer research. Nucleic Acids Res. 44, e108. https://doi.org/10.1093/nar/gkw227

LaMar, D., 2015. FastQC. https://qubeshub.org/resources/fastqc

Leedham, S.J., Preston, S.L., McDonald, S. a. C., Elia, G., Bhandari, P., Poller, D., Harrison, R., Novelli, M.R., Jankowski, J.A., Wright, N.A., 2008. Individual crypt genetic heterogeneity and the origin of metaplastic glandular epithelium in human Barrett's oesophagus. Gut 57, 1041–1048. https://doi.org/10.1136/gut.2007.143339

Lee-Six, H., Obro, N.F., Shepherd, M.S., Grossmann, S., Dawson, K., Belmonte, M., Osborne, R.J., Huntly, B.J.P., Martincorena, I., Anderson, E., O'Neill, L., Stratton, M.R., Laurenti, E., Green, A.R., Kent, D.G., Campbell, P.J., 2018. Population dynamics of normal human blood inferred from somatic mutations. Nature 561, 473-+. https://doi.org/10.1038/s41586-018-0497-0

Lee-Six, H., Olafsson, S., Ellis, P., Osborne, R.J., Sanders, M.A., Moore, L., Georgakopoulos, N., Torrente, F., Noorani, A., Goddard, M., Robinson, P., Coorens, T.H.H., O'Neill, L., Alder, C., Wang, J., Fitzgerald, R.C., Zilbauer, M., Coleman, N., Saeb-Parsy, K., Martincorena, I., Campbell, P.J., Stratton, M.R., 2019. The landscape of somatic mutation in normal colorectal epithelial cells. Nature 574, 532–537. https://doi.org/10.1038/s41586-019-1672-7

Li, H., 2011. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics 27, 2987–2993. https://doi.org/10.1093/bioinformatics/btr509

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., 1000 Genome Project Data Processing Subgroup, 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25, 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Li, R., Di, L., Li, J., Fan, W., Liu, Y., Guo, W., Liu, W., Liu, L., Li, Q., Chen, L., Chen, Y., Miao, C., Liu, H., Wang, Y., Ma, Y., Xu, D., Lin, D., Huang, Y., Wang, J., Bai, F., Wu, C., 2021. A body map of somatic mutagenesis in morphologically normal human tissues. Nature 597, 398–403. https://doi.org/10.1038/s41586-021-03836-1

Li, S., Mason, C.E., 2014. The Pivotal Regulatory Landscape of RNA Modifications. Annu. Rev. Genomics Hum. Genet. 15, 127–150. https://doi.org/10.1146/annurev-genom-090413-025405

Li, W., Lynch, M., 2020. Universally high transcript error rates in bacteria. eLife 9, e54898. https://doi.org/10.7554/eLife.54898

Lindahl, T., 1993. Instability and decay of the primary structure of DNA. Nature 362, 709–715. https://doi.org/10.1038/362709a0

Lindahl, T., Wood, R.D., 1999. Quality Control by DNA Repair. Science 286, 1897–1905. https://doi.org/10.1126/science.286.5446.1897

Liu, F., Zhang, Y., Zhang, L., Li, Z., Fang, Q., Gao, R., Zhang, Z., 2019. Systematic comparative analysis of single-nucleotide variant detection methods from single-cell RNA sequencing data. Genome Biol. 20, 242. https://doi.org/10.1186/s13059-019-1863-4

Lo Giudice, C., Silvestris, D.A., Roth, S.H., Eisenberg, E., Pesole, G., Gallo, A., Picardi, E., 2020. Quantifying RNA Editing in Deep Transcriptome Datasets. Front. Genet. 11.

Lodato, M.A., Rodin, R.E., Bohrson, C.L., Coulter, M.E., Barton, A.R., Kwon, M., Sherman, M.A., Vitzthum, C.M., Luquette, L.J., Yandava, C.N., Yang, P., Chittenden, T.W., Hatem, N.E., Ryu, S.C., Woodworth, M.B., Park, P.J., Walsh, C.A., 2018. Aging and neurodegeneration are associated with increased mutations in single human neurons. Science 359, 555–559. https://doi.org/10.1126/science.aao4426

Lörinc, E., Mellblom, L., Öberg, S., 2015. The immunophenotypic relationship between the submucosal gland unit, columnar metaplasia and squamous islands in the columnar-lined oesophagus. Histopathology 67, 792–798. https://doi.org/10.1111/his.12719

Lörinc, E., Öberg, S., 2012. Submucosal glands in the columnar-lined oesophagus: evidence of an association with metaplasia and neosquamous epithelium. Histopathology 61, 53–58. https://doi.org/10.1111/j.1365-2559.2012.04180.x

Ma, X., Shao, Y., Tian, L., Flasch, D.A., Mulder, H.L., Edmonson, M.N., Liu, Y., Chen, X., Newman, S., Nakitandwe, J., Li, Y., Li, B., Shen, S., Wang, Z., Shurtleff, S., Robison, L.L., Levy, S., Easton, J., Zhang, J., 2019. Analysis of error profiles in deep next-generation sequencing data. Genome Biol. 20, 50. https://doi.org/10.1186/s13059-019-1659-6

Martincorena, I., Campbell, P.J., 2015. Somatic mutation in cancer and normal cells. Science 349, 1483–1489. https://doi.org/10.1126/science.aab4082

Martincorena, I., Fowler, J.C., Wabik, A., Lawson, A.R.J., Abascal, F., Hall, M.W.J., Cagan, A., Murai, K., Mahbubani, K., Stratton, M.R., Fitzgerald, R.C., Handford, P.A., Campbell, P.J., Saeb-Parsy, K., Jones, P.H., 2018. Somatic mutant clones colonize the human esophagus with age. Science 362, 911–917. https://doi.org/10.1126/science.aau3879

Martincorena, I., Roshan, A., Gerstung, M., Ellis, P., Van Loo, P., McLaren, S., Wedge, D.C., Fullam, A., Alexandrov, L.B., Tubio, J.M., Stebbings, L., Menzies, A., Widaa, S., Stratton, M.R., Jones, P.H., Campbell, P.J., 2015. High burden and pervasive positive selection of somatic mutations in normal human skin. Science 348, 880–886. https://doi.org/10.1126/science.aaa6806

Mattick, J.S., Makunin, I.V., 2006. Non-coding RNA. Hum. Mol. Genet. 15, R17–R29. https://doi.org/10.1093/hmg/ddl046

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., DePristo, M.A., 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 20, 1297–1303. https://doi.org/10.1101/gr.107524.110

McLaren, W., Gil, L., Hunt, S.E., Riat, H.S., Ritchie, G.R.S., Thormann, A., Flicek, P., Cunningham, F., 2016. The Ensembl Variant Effect Predictor. Genome Biol. 17, 122. https://doi.org/10.1186/s13059-016-0974-4

Method of the Year 2013 | Nature Methods [WWW Document], n.d. URL https://www.nature.com/articles/nmeth.2801 (accessed 6.20.22).

Moore, L., Leongamornlert, D., Coorens, T.H.H., Sanders, M.A., Ellis, P., Dentro, S.C., Dawson, K.J., Butler, T., Rahbari, R., Mitchell, T.J., Maura, F., Nangalia, J., Tarpey, P.S., Brunner, S.F., Lee-Six, H., Hooks, Y., Moody, S., Mahbubani, K.T., Jimenez-Linan, M., Brosens, J.J., Iacobuzio-Donahue,

C.A., Martincorena, I., Saeb-Parsy, K., Campbell, P.J., Stratton, M.R., 2020. The mutational landscape of normal human endometrial epithelium. Nature 580, 640–646. https://doi.org/10.1038/s41586-020-2214-z

Mustjoki, S., Young, N.S., 2021. Somatic Mutations in "Benign" Disease. N. Engl. J. Med. 384, 2039–2052. https://doi.org/10.1056/NEJMra2101920

Nik-Zainal, S., Morganella, S., 2017. Mutational Signatures in Breast Cancer: The Problem at the DNA Level. Clin. Cancer Res. Off. J. Am. Assoc. Cancer Res. 23, 2617–2629. https://doi.org/10.1158/1078-0432.CCR-16-2810

Nowicki-Osuch, K., Zhuang, L., Jammula, S., Bleaney, C.W., Mahbubani, K.T., Devonshire, G., Katz-Summercorn, A., Eling, N., Wilbrey-Clark, A., Madissoon, E., Gamble, J., Di Pietro, M., O'Donovan, M., Meyer, K.B., Saeb-Parsy, K., Sharrocks, A.D., Teichmann, S.A., Marioni, J.C., Fitzgerald, R.C., 2021. Molecular phenotyping reveals the identity of Barrett's esophagus and its malignant transition. Science 373, 760–767. https://doi.org/10.1126/science.abd1449

Ouatu–Lascar, R., Fitzgerald, R.C., Triadafilopoulos, G., 1999. Differentiation and proliferation in Barrett's esophagus and the effects of acid suppression. Gastroenterology 117, 327–335. https://doi.org/10.1053/gast.1999.0029900327

Owen, R.P., n.d. Single cell RNA-sequencing in the upper gastrointestinal tract 274.

Owen, R.P., White, M.J., Severson, D.T., Braden, B., Bailey, A., Goldin, R., Wang, L.M., Ruiz-Puig, C., Maynard, N.D., Green, A., Piazza, P., Buck, D., Middleton, M.R., Ponting, C.P., Schuster-Böckler, B., Lu, X., 2018. Single cell RNA-seq reveals profound transcriptional similarity between Barrett's oesophagus and oesophageal submucosal glands. Nat. Commun. 9, 4261. https://doi.org/10.1038/s41467-018-06796-9

Paulson, T.G., Galipeau, P.C., Oman, K.M., Sanchez, C.A., Kuhner, M.K., Smith, L.P., Hadi, K., Shah, M., Arora, K., Shelton, J., Johnson, M., Corvelo, A., Maley, C.C., Yao, X., Sanghvi, R., Venturini, E., Emde, A.-K., Hubert, B., Imielinski, M., Robine, N., Reid, B.J., Li, X., 2022. Somatic whole genome dynamics of precancer in Barrett's esophagus reveals features associated with disease progression. Nat. Commun. 13, 2300. https://doi.org/10.1038/s41467-022-29767-7

Pfeiffer, F., Gröber, C., Blank, M., Händler, K., Beyer, M., Schultze, J.L., Mayer, G., 2018. Systematic evaluation of error rates and causes in short samples in next-generation sequencing. Sci. Rep. 8, 10950. https://doi.org/10.1038/s41598-018-29325-6

Phan, V., Gao, S., Tran, Q., Vo, N.S., 2015. How genome complexity can explain the difficulty of aligning reads to genomes. BMC Bioinformatics 16, S3. https://doi.org/10.1186/1471-2105-16-S17-S3

Pienaar, E., Theron, M., Nelson, M., Viljoen, H., 2006. A QUANTITATIVE MODEL OF ERROR ACCUMULATION DURING PCR AMPLIFICATION. Comput. Biol. Chem. 30, 102–111. https://doi.org/10.1016/j.compbiolchem.2005.11.002

Piskol, R., Ramaswami, G., Li, J.B., 2013. Reliable Identification of Genomic Variants from RNA-Seq Data. Am. J. Hum. Genet. 93, 641–651. https://doi.org/10.1016/j.ajhg.2013.08.008

Pohl, H., Sirovich, B., Welch, H.G., 2010. Esophageal Adenocarcinoma Incidence: Are We Reaching the Peak? Cancer Epidemiol. Biomarkers Prev. 19, 1468–1470. https://doi.org/10.1158/1055-9965.EPI-10-0012

Potapov, V., Ong, J.L., 2017. Examining Sources of Error in PCR by Single-Molecule Sequencing. PLoS ONE 12, e0169774. https://doi.org/10.1371/journal.pone.0169774

Qiu, P., 2020. Embracing the dropouts in single-cell RNA-seq analysis. Nat. Commun. 11, 1169. https://doi.org/10.1038/s41467-020-14976-9

Quante, M., Bhagat, G., Abrams, J.A., Marache, F., Good, P., Lee, M.D., Lee, Y., Friedman, R., Asfaha, S., Dubeykovskaya, Z., Mahmood, U., Figueiredo, J.-L., Kitajewski, J., Shawber, C., Lightdale, C.J., Rustgi, A.K., Wang, T.C., 2012. Bile Acid and Inflammation Activate Gastric Cardia Stem Cells in a Mouse Model of Barrett-Like Metaplasia. Cancer Cell 21, 36–51. https://doi.org/10.1016/j.ccr.2011.12.004

Quinlan, A.R., Hall, I.M., 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26, 841–842. https://doi.org/10.1093/bioinformatics/btq033

Quinn, E.M., Cormican, P., Kenny, E.M., Hill, M., Anney, R., Gill, M., Corvin, A.P., Morris, D.W., 2013. Development of Strategies for SNP Detection in RNA-Seq Data: Application to Lymphoblastoid Cell Lines and Evaluation Using 1000 Genomes Data. PLOS ONE 8, e58815. https://doi.org/10.1371/journal.pone.0058815

Rahbari, R., Wuster, A., Lindsay, S.J., Hardwick, R.J., Alexandrov, L.B., Al Turki, S., Dominiczak, A., Morris, A., Porteous, D., Smith, B., Stratton, M.R., Hurles, M.E., 2016. Timing, rates and spectra of human germline mutation. Nat. Genet. 48, 126–133. https://doi.org/10.1038/ng.3469

Rhee, H., Wang, D.H., 2018. Cellular Origins of Barrett's Esophagus: the Search Continues. Curr. Gastroenterol. Rep. 20, 51. https://doi.org/10.1007/s11894-018-0657-2

Rimmer, A., Phan, H., Mathieson, I., Iqbal, Z., Twigg, S.R.F., Wilkie, A.O.M., McVean, G., Lunter, G., 2014. Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. Nat. Genet. 46, 912–918. https://doi.org/10.1038/ng.3036

Rosenthal, R., McGranahan, N., Herrero, J., Taylor, B.S., Swanton, C., 2016. deconstructSigs: delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution. Genome Biol. 17, 31. https://doi.org/10.1186/s13059-016-0893-4

Ross-Innes, C.S., Becq, J., Warren, A., Cheetham, R.K., Northen, H., O'Donovan, M., Malhotra, S., di Pietro, M., Ivakhno, S., He, M., Weaver, J.M.J., Lynch, A.G., Kingsbury, Z., Ross, M., Humphray, S., Bentley, D., Fitzgerald, R.C., 2015. Whole-genome sequencing provides new insights into the clonal architecture of Barrett's esophagus and esophageal adenocarcinoma. Nat. Genet. 47, 1038–1046. https://doi.org/10.1038/ng.3357

RT-PCR: One-Step vs. Two-Step | Thermo Fisher Scientific - UK [WWW Document], n.d. URL https://www.thermofisher.com/uk/en/home/brands/thermo-scientific/molecular-biology/molecular-biology-learning-center/molecular-biology-resource-library/spotlight-articles/onestep-vs-twostep-rtpcr.html (accessed 5.24.22).

Sahlin, K., Mäkinen, V., 2021. Accurate spliced alignment of long RNA sequencing reads. Bioinformatics 37, 4643–4651. https://doi.org/10.1093/bioinformatics/btab540

Sarosi, G., Brown, G., Jaiswal, K., Feagins, L.A., Lee, E., Crook, T.W., Souza, R.F., Zou, Y.S., Shay, J.W., Spechler, S.J., 2008. Bone marrow progenitor cells contribute to esophageal regeneration and metaplasia in a rat model of Barrett's esophagus. Dis. Esophagus 21, 43–50. https://doi.org/10.1111/j.1442-2050.2007.00744.x

Saunders, C.T., Wong, W.S.W., Swamy, S., Becq, J., Murray, L.J., Cheetham, R.K., 2012. Strelka: accurate somatic small-variant calling from sequenced tumor–normal sample pairs. Bioinformatics 28, 1811–1817. https://doi.org/10.1093/bioinformatics/bts271

Shaheen, N.J., Sharma, P., Overholt, B.F., Wolfsen, H.C., Sampliner, R.E., Wang, K.K., Galanko, J.A., Bronner, M.P., Goldblum, J.R., Bennett, A.E., Jobe, B.A., Eisen, G.M., Fennerty, M.B., Hunter, J.G., Fleischer, D.E., Sharma, V.K., Hawes, R.H., Hoffman, B.J., Rothstein, R.I., Gordon, S.R., Mashimo, H., Chang, K.J., Muthusamy, V.R., Edmundowicz, S.A., Spechler, S.J., Siddiqui, A.A., Souza, R.F., Infantolino, A., Falk, G.W., Kimmey, M.B., Madanick, R.D., Chak, A., Lightdale, C.J., 2009. Radiofrequency Ablation in Barrett's Esophagus with Dysplasia. N. Engl. J. Med. 360, 2277–2288. https://doi.org/10.1056/NEJMoa0808145

Souza, R.F., 2016. From Reflux Esophagitis to Esophageal Adenocarcinoma. Dig. Dis. 34, 483–490. https://doi.org/10.1159/000445225

Stegle, O., Teichmann, S.A., Marioni, J.C., 2015. Computational and analytical challenges in single-cell transcriptomics. Nat. Rev. Genet. 16, 133–145. https://doi.org/10.1038/nrg3833

Stoler, N., Nekrutenko, A., 2021. Sequencing error profiles of Illumina sequencing instruments. NAR Genomics Bioinforma. 3, lqab019. https://doi.org/10.1093/nargab/lqab019

Stratton, M.R., Campbell, P.J., Futreal, P.A., 2009. The cancer genome. Nature 458, 719–724. https://doi.org/10.1038/nature07943

Sturtevant, A.H., 1937. Essays on Evolution. I. On the Effects of Selection on Mutation Rate. Q. Rev. Biol. 12, 464–467. https://doi.org/10.1086/394543

Sun, Y., Buhler, J., 2006. Choosing the best heuristic for seeded alignment of DNA sequences. BMC Bioinformatics 7, 133. https://doi.org/10.1186/1471-2105-7-133

Tang, J., Fewings, E., Chang, D., Zeng, H., Liu, S., Jorapur, A., Belote, R.L., McNeal, A.S., Tan, T.M., Yeh, I., Arron, S.T., Judson-Torres, R.L., Bastian, B.C., Shain, A.H., 2020. The genomic landscapes of individual melanocytes from human skin. Nature 586, 600–605. https://doi.org/10.1038/s41586-020-2785-8

Tang, X., Baheti, S., Shameer, K., Thompson, K.J., Wills, Q., Niu, N., Holcomb, I.N., Boutet, S.C., Ramakrishnan, R., Kachergus, J.M., Kocher, J.-P.A., Weinshilboum, R.M., Wang, L., Thompson, E.A., Kalari, K.R., 2014. The eSNV-detect: a computational system to identify expressed single nucleotide variants from transcriptome sequencing data. Nucleic Acids Res. 42, e172. https://doi.org/10.1093/nar/gku1005

Tang, X., Huang, Y., Lei, J., Luo, H., Zhu, X., 2019. The single-cell sequencing: new developments and medical applications. Cell Biosci. 9, 53. https://doi.org/10.1186/s13578-019-0314-y

Tate, J.G., Bamford, S., Jubb, H.C., Sondka, Z., Beare, D.M., Bindal, N., Boutselakis, H., Cole, C.G., Creatore, C., Dawson, E., Fish, P., Harsha, B., Hathaway, C., Jupe, S.C., Kok, C.Y., Noble, K., Ponting, L., Ramshaw, C.C., Rye, C.E., Speedy, H.E., Stefancsik, R., Thompson, S.L., Wang, S., Ward, S., Campbell, P.J., Forbes, S.A., 2019. COSMIC: the Catalogue Of Somatic Mutations In Cancer. Nucleic Acids Res. 47, D941–D947. https://doi.org/10.1093/nar/gky1015

The 1000 Genomes Project Consortium, Corresponding authors, Auton, A., Abecasis, G.R., Steering committee, Altshuler, D.M., Durbin, R.M., Abecasis, G.R., Bentley, D.R., Chakravarti, A., Clark, A.G., Donnelly, P., Eichler, E.E., Flicek, P., Gabriel, S.B., Gibbs, R.A., Green, E.D., Hurles, M.E., Knoppers, B.M., Korbel, J.O., Lander, E.S., Lee, C., Lehrach, H., Mardis, E.R., Marth, G.T., McVean, G.A., Nickerson, D.A., Schmidt, J.P., Sherry, S.T., Wang, J., Wilson, R.K., Production group, Baylor College of Medicine, Gibbs, R.A., Boerwinkle, E., Doddapaneni, H., Han, Y., Korchina, V., Kovar, C., Lee, S., Muzny, D., Reid, J.G., Zhu, Y., BGI-Shenzhen, Wang, J., Chang, Y., Feng, Q., Fang, X., Guo, X., Jian, M., Jiang, H., Jin, X., Lan, T., Li, G., Li, J., Li, Yingrui, Liu, S., Liu, Xiao, Lu, Y., Ma, X., Tang, M., Wang, B., Wang, G., Wu, H., Wu, R., Xu, X., Yin, Y., Zhang, D., Zhang, W., Zhao, J., Zhao, M., Zheng, X., Broad Institute of MIT and Harvard, Lander, E.S., Altshuler, D.M., Gabriel, S.B., Gupta, N., Coriell Institute for Medical Research, Gharani, N., Toji, L.H., Gerry, N.P., Resch, A.M., European Molecular Biology Laboratory, European Bioinformatics Institute, Flicek, P., Barker, J., Clarke, L., Gil, L., Hunt, S.E., Kelman, G., Kulesha, E., Leinonen, R., McLaren, W.M., Radhakrishnan, R., Roa, A., Smirnov, D., Smith, R.E., Streeter, I., Thormann, A., Toneva, I., Vaughan, B., Zheng-Bradley, X., Illumina, Bentley, D.R., Grocock, R., Humphray, S., James, T., Kingsbury, Z., Max Planck Institute for Molecular Genetics, Lehrach, H., Sudbrak, R., Albrecht, M.W., Amstislavskiy, V.S., Borodina, T.A., Lienhard, M., Mertes, F., Sultan, M., Timmermann, B., Yaspo, M.-L., McDonnell Genome Institute at Washington University, Mardis, E.R., Wilson, R.K., Fulton, L., Fulton, R., US National Institutes of Health, Sherry, S.T., Ananiev, V., Belaia, Z., Beloslyudtsev, D., Bouk, N., Chen, C., Church, D., Cohen, R., Cook, C., Garner, J., Hefferon, T., Kimelman, M., Liu, C., Lopez, J., Meric, P., O'Sullivan, C., Ostapchuk, Y., Phan, L., Ponomarov, S., Schneider, V., Shekhtman, E., Sirotkin, K., Slotta, D., Zhang, H., University of Oxford, McVean, G.A., Wellcome Trust Sanger Institute, Durbin, R.M., Balasubramaniam, S., Burton, J., Danecek, P., Keane, T.M., Kolb-Kokocinski, A., McCarthy, S., Stalker, J., Quail, M., Analysis group, Affymetrix, Schmidt, J.P., Davies, C.J., Gollub, J., Webster, T., Wong, B., Zhan, Y., Albert Einstein College of Medicine, Auton, A., Campbell, C.L., Kong, Y., Marcketta, A., Baylor College of Medicine, Gibbs, R.A., Yu, F., Antunes, L., Bainbridge, M., Muzny, D., Sabo, A., Huang, Z., BGI-Shenzhen, Wang, J., Coin, L.J.M., Fang, L., Guo, X., Jin, X., Li, G., Li, Q., Li, Yingrui, Li, Z., Lin, H., Liu, B., Luo, R., Shao, H., Xie, Y., Ye, C., Yu, C., Zhang, F., Zheng, H., Zhu, H., Bilkent University, Alkan, C., Dal, E., Kahveci, F., Boston

College, Marth, G.T., Garrison, E.P., Kural, D., Lee, W.-P., Fung Leong, W., Stromberg, M., Ward, A.N., Wu, J., Zhang, M., Broad Institute of MIT and Harvard, Daly, M.J., DePristo, M.A., Handsaker, R.E., Altshuler, D.M., Banks, E., Bhatia, G., del Angel, G., Gabriel, S.B., Genovese, G., Gupta, N., Li, H., Kashin, S., Lander, E.S., McCarroll, S.A., Nemesh, J.C., Poplin, R.E., Cold Spring Harbor Laboratory, Yoon, S.C., Lihm, J., Makarov, V., Cornell University, Clark, A.G., Gottipati, S., Keinan, A., Rodriguez-Flores, J.L., European Molecular Biology Laboratory, Korbel, J.O., Rausch, T., Fritz, M.H., Stütz, A.M., European Molecular Biology Laboratory, European Bioinformatics Institute, Flicek, P., Beal, K., Clarke, L., Datta, A., Herrero, J., McLaren, W.M., Ritchie, G.R.S., Smith, R.E., Zerbino, D., Zheng-Bradley, X., Harvard University, Sabeti, P.C., Shlyakhter, I., Schaffner, S.F., Vitti, J., Human Gene Mutation Database, Cooper, D.N., Ball, E.V., Stenson, P.D., Illumina, Bentley, D.R., Barnes, B., Bauer, M., Keira Cheetham, R., Cox, A., Eberle, M., Humphray, S., Kahn, S., Murray, L., Peden, J., Shaw, R., Icahn School of Medicine at Mount Sinai, Kenny, E.E., Louisiana State University, Batzer, M.A., Konkel, M.K., Walker, J.A., Massachusetts General Hospital, MacArthur, D.G., Lek, M., Max Planck Institute for Molecular Genetics, Sudbrak, R., Amstislavskiy, V.S., Herwig, R., McDonnell Genome Institute at Washington University, Mardis, E.R., Ding, L., Koboldt, D.C., Larson, D., Ye, Kai, McGill University, Gravel, S., National Eye Institute, NIH, Swaroop, A., Chew, E., New York Genome Center, Lappalainen, T., Erlich, Y., Gymrek, M., Frederick Willems, T., Ontario Institute for Cancer Research, Simpson, J.T., Pennsylvania State University, Shriver, M.D., Rutgers Cancer Institute of New Jersey, Rosenfeld, J.A., Stanford University, Bustamante, C.D., Montgomery, S.B., De La Vega, F.M., Byrnes, J.K., Carroll, A.W., DeGorter, M.K., Lacroute, P., Maples, B.K., Martin, A.R., Moreno-Estrada, A., Shringarpure, S.S., Zakharia, F., Tel-Aviv University, Halperin, E., Baran, Y., The Jackson Laboratory for Genomic Medicine, Lee, C., Cerveira, E., Hwang, J., Malhotra, A., Plewczynski, D., Radew, K., Romanovitch, M., Zhang, C., Thermo Fisher Scientific, Hyland, F.C.L., Translational Genomics Research Institute, Craig, D.W., Christoforides, A., Homer, N., Izatt, T., Kurdoglu, A.A., Sinari, S.A., Squire, K., US National Institutes of Health, Sherry, S.T., Xiao, C., University of California, San Diego, Sebat, J., Antaki, D., Gujral, M., Noor, A., Ye, Kenny, University of California, San Francisco, Burchard, E.G., Hernandez, R.D., Gignoux, C.R., University of California, Santa Cruz, Haussler, D., Katzman, S.J., James Kent, W., University of Chicago, Howie, B., University College London, Ruiz-Linares, A., University of Geneva, Dermitzakis, E.T., University of Maryland School of Medicine, Devine, S.E., University of Michigan, Abecasis, G.R., Min Kang, H., Kidd, J.M., Blackwell, T., Caron, S., Chen, W., Emery, S., Fritsche, L., Fuchsberger, C., Jun, G., Li, B., Lyons, R., Scheller, C., Sidore, C., Song, S., Sliwerska, E., Taliun, D., Tan, A., Welch, R., Kate Wing, M., Zhan, X., University of Montréal, Awadalla, P., Hodgkinson, A., University of North Carolina at Chapel Hill, Li, Yun, University of North Carolina at Charlotte, Shi, X., Quitadamo, A., University of Oxford, Lunter, G., McVean, G.A., Marchini, J.L., Myers, S., Churchhouse, C., Delaneau, O., Gupta-Hinch, A., Kretzschmar, W., Iqbal, Z., Mathieson, I., Menelaou, A., Rimmer, A., Xifara, D.K., University of Puerto Rico, Oleksyk, T.K., University of Texas Health Sciences Center at Houston, Fu, Yunxin, Liu, Xiaoming, Xiong, M., University of Utah, Jorde, L., Witherspoon, D., Xing, J., University of Washington, Eichler, E.E., Browning, B.L., Browning, S.R., Hormozdiari, F., Sudmant, P.H., Weill Cornell Medical College, Khurana, E., Wellcome Trust Sanger Institute, Durbin, R.M., Hurles, M.E., Tyler-Smith, C., Albers, C.A., Ayub, Q., Balasubramaniam, S., Chen, Y., Colonna, V., Danecek, P., Jostins, L., Keane, T.M., McCarthy, S., Walter, K., Xue, Y., Yale University, Gerstein, M.B., Abyzov, A., Balasubramanian, S., Chen, J., Clarke, D., Fu, Yao, Harmanci, A.O., Jin, M., Lee, D., Liu, J., Jasmine Mu, X., Zhang, J., Zhang, Yan, Structural variation group, BGI-Shenzhen, Li, Yingrui, Luo, R., Zhu, H., Bilkent University, Alkan, C., Dal, E., Kahveci, F., Boston College, Marth, G.T., Garrison, E.P., Kural, D., Lee, W.-P., Ward, A.N., Wu, J., Zhang, M., Broad Institute of MIT and Harvard, McCarroll, S.A., Handsaker, R.E., Altshuler, D.M., Banks, E., del Angel, G., Genovese, G., Hartl, C., Li, H., Kashin, S., Nemesh, J.C., Shakir, K., Cold Spring Harbor Laboratory, Yoon,

S.C., Lihm, J., Makarov, V., Cornell University, Degenhardt, J., European Molecular Biology Laboratory, Korbel, J.O., Fritz, M.H., Meiers, S., Raeder, B., Rausch, T., Stütz, A.M., European Molecular Biology Laboratory, European Bioinformatics Institute, Flicek, P., Paolo Casale, F., Clarke, L., Smith, R.E., Stegle, O., Zheng-Bradley, X., Illumina, Bentley, D.R., Barnes, B., Keira Cheetham, R., Eberle, M., Humphray, S., Kahn, S., Murray, L., Shaw, R., Leiden University Medical Center, Lameijer, E.-W., Louisiana State University, Batzer, M.A., Konkel, M.K., Walker, J.A., McDonnell Genome Institute at Washington University, Ding, L., Hall, I., Ye, Kai, Stanford University, Lacroute, P., The Jackson Laboratory for Genomic Medicine, Lee, C., Cerveira, E., Malhotra, A., Hwang, J., Plewczynski, D., Radew, K., Romanovitch, M., Zhang, C., Translational Genomics Research Institute, Craig, D.W., Homer, N., US National Institutes of Health, Church, D., Xiao, C., University of California, San Diego, Sebat, J., Antaki, D., Bafna, V., Michaelson, J., Ye, Kenny, University of Maryland School of Medicine, Devine, S.E., Gardner, E.J., University of Michigan, Abecasis, G.R., Kidd, J.M., Mills, R.E., Dayama, G., Emery, S., Jun, G., University of North Carolina at Charlotte, Shi, X., Quitadamo, A., University of Oxford, Lunter, G., McVean, G.A., University of Texas MD Anderson Cancer Center, Chen, K., Fan, X., Chong, Z., Chen, T., University of Utah, Witherspoon, D., Xing, J., University of Washington, Eichler, E.E., Chaisson, M.J., Hormozdiari, F., Huddleston, J., Malig, M., Nelson, B.J., Sudmant, P.H., Vanderbilt University School of Medicine, Parrish, N.F., Weill Cornell Medical College, Khurana, E., Wellcome Trust Sanger Institute, Hurles, M.E., Blackburne, B., Lindsay, S.J., Ning, Z., Walter, K., Zhang, Yujun, Yale University, Gerstein, M.B., Abyzov, A., Chen, J., Clarke, D., Lam, H., Jasmine Mu, X., Sisu, C., Zhang, J., Zhang, Yan, Exome group, Baylor College of Medicine, Gibbs, R.A., Yu, F., Bainbridge, M., Challis, D., Evani, U.S., Kovar, C., Lu, J., Muzny, D., Nagaswamy, U., Reid, J.G., Sabo, A., Yu, J., BGI-Shenzhen, Guo, X., Li, W., Li, Yingrui, Wu, R., Boston College, Marth, G.T., Garrison, E.P., Fung Leong, W., Ward, A.N., Broad Institute of MIT and Harvard, del Angel, G., DePristo, M.A., Gabriel, S.B., Gupta, N., Hartl, C., Poplin, R.E., Cornell University, Clark, A.G., Rodriguez-Flores, J.L., European Molecular Biology Laboratory, European Bioinformatics Institute, Flicek, P., Clarke, L., Smith, R.E., Zheng-Bradley, X., Massachusetts General Hospital, MacArthur, D.G., McDonnell Genome Institute at Washington University, Mardis, E.R., Fulton, R., Koboldt, D.C., McGill University, Gravel, S., Stanford University, Bustamante, C.D., Translational Genomics Research Institute, Craig, D.W., Christoforides, A., Homer, N., Izatt, T., US National Institutes of Health, Sherry, S.T., Xiao, C., University of Geneva, Dermitzakis, E.T., University of Michigan, Abecasis, G.R., Min Kang, H., University of Oxford, McVean, G.A., Yale University, Gerstein, M.B., Balasubramanian, S., Habegger, L., Functional interpretation group, Cornell University, Yu, H., European Molecular Biology Laboratory, European Bioinformatics Institute, Flicek, P., Clarke, L., Cunningham, F., Dunham, I., Zerbino, D., Zheng-Bradley, X., Harvard University, Lage, K., Berg Jespersen, J., Horn, H., Stanford University, Montgomery, S.B., DeGorter, M.K., Weill Cornell Medical College, Khurana, E., Wellcome Trust Sanger Institute, Tyler-Smith, C., Chen, Y., Colonna, V., Xue, Y., Yale University, Gerstein, M.B., Balasubramanian, S., Fu, Yao, Kim, D., Chromosome Y group, Albert Einstein College of Medicine, Auton, A., Marcketta, A., American Museum of Natural History, Desalle, R., Narechania, A., Arizona State University, Wilson Sayres, M.A., Boston College, Garrison, E.P., Broad Institute of MIT and Harvard, Handsaker, R.E., Kashin, S., McCarroll, S.A., Cornell University, Rodriguez-Flores, J.L., European Molecular Biology Laboratory, European Bioinformatics Institute, Flicek, P., Clarke, L., Zheng-Bradley, X., New York Genome Center, Erlich, Y., Gymrek, M., Frederick Willems, T., Stanford University, Bustamante, C.D., Mendez, F.L., David Poznik, G., Underhill, P.A., The Jackson Laboratory for Genomic Medicine, Lee, C., Cerveira, E., Malhotra, A., Romanovitch, M., Zhang, C., University of Michigan, Abecasis, G.R., University of Queensland, Coin, L., Shao, H., Virginia Bioinformatics Institute, Mittelman, D., Wellcome Trust Sanger Institute, Tyler-Smith, C., Ayub, Q., Banerjee, R., Cerezo, M., Chen, Y., Fitzgerald, T.W., Louzada, S., Massaia, A., McCarthy, S., Ritchie, G.R., Xue, Y., Yang, F., Data

coordination center group, Baylor College of Medicine, Gibbs, R.A., Kovar, C., Kalra, D., Hale, W., Muzny, D., Reid, J.G., BGI-Shenzhen, Wang, J., Dan, X., Guo, X., Li, G., Li, Yingrui, Ye, C., Zheng, X., Broad Institute of MIT and Harvard, Altshuler, D.M., European Molecular Biology Laboratory, European Bioinformatics Institute, Flicek, P., Clarke, L., Zheng-Bradley, X., Illumina, Bentley, D.R., Cox, A., Humphray, S., Kahn, S., Max Planck Institute for Molecular Genetics, Sudbrak, R., Albrecht, M.W., Lienhard, M., McDonnell Genome Institute at Washington University, Larson, D., Translational Genomics Research Institute, Craig, D.W., Izatt, T., Kurdoglu, A.A., US National Institutes of Health, Sherry, S.T., Xiao, C., University of California, Santa Cruz, Haussler, D., University of Michigan, Abecasis, G.R., University of Oxford, McVean, G.A., Wellcome Trust Sanger Institute, Durbin, R.M., Balasubramaniam, S., Keane, T.M., McCarthy, S., Stalker, J., Samples and ELSI group, Chakravarti, A., Knoppers, B.M., Abecasis, G.R., Barnes, K.C., Beiswanger, C., Burchard, E.G., Bustamante, C.D., Cai, H., Cao, H., Durbin, R.M., Gerry, N.P., Gharani, N., Gibbs, R.A., Gignoux, C.R., Gravel, S., Henn, B., Jones, D., Jorde, L., Kaye, J.S., Keinan, A., Kent, A., Kerasidou, A., Li, Yingrui, Mathias, R., McVean, G.A., Moreno-Estrada, A., Ossorio, P.N., Parker, M., Resch, A.M., Rotimi, C.N., Royal, C.D., Sandoval, K., Su, Y., Sudbrak, R., Tian, Z., Tishkoff, S., Toji, L.H., Tyler-Smith, C., Via, M., Wang, Y., Yang, H., Yang, L., Zhu, J., Sample collection, British from England and Scotland (GBR), Bodmer, W., Colombians in Medellín, Colombia (CLM), Bedoya, G., Ruiz-Linares, A., Han Chinese South (CHS), Cai, Z., Gao, Y., Chu, J., Finnish in Finland (FIN), Peltonen, L., Iberian Populations in Spain (IBS), Garcia-Montero, A., Orfao, A., Puerto Ricans in Puerto Rico (PUR), Dutil, J., Martinez-Cruzado, J.C., Oleksyk, T.K., African Caribbean in Barbados (ACB), Barnes, K.C., Mathias, R.A., Hennis, A., Watson, H., McKenzie, C., Bengali in Bangladesh (BEB), Qadri, F., LaRocque, R., Sabeti, P.C., Chinese Dai in Xishuangbanna, China (CDX), Zhu, J., Deng, X., Esan in Nigeria (ESN), Sabeti, P.C., Asogun, D., Folarin, O., Happi, C., Omoniwa, O., Stremlau, M., Tariyal, R., Gambian in Western Division – Mandinka (GWD), Jallow, M., Sisay Joof, F., Corrah, T., Rockett, K., Kwiatkowski, D., Indian Telugu in the UK (ITU) and Sri Lankan Tamil in the UK (STU), Kooner, J., Kinh in Ho Chi Minh City, Vietnam (KHV), Tịnh Hiê`n, T., Dunstan, S.J., Thuy Hang, N., Mende in Sierra Leone (MSL), Fonnie, R., Garry, R., Kanneh, L., Moses, L., Sabeti, P.C., Schieffelin, J., Grant, D.S., Peruvian in Lima, Peru (PEL), Gallo, C., Poletti, G., Punjabi in Lahore, Pakistan (PJL), Saleheen, D., Rasheed, A., Scientific management, Brooks, L.D., Felsenfeld, A.L., McEwen, J.E., Vaydylevich, Y., Green, E.D., Duncanson, A., Dunn, M., Schloss, J.A., Wang, J., Yang, H., Writing group, Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Min Kang, H., Korbel, J.O., Marchini, J.L., McCarthy, S., McVean, G.A., Abecasis, G.R., 2015. A global reference for human genetic variation. Nature 526, 68–74. https://doi.org/10.1038/nature15393

van Nieuwenhove, Y., Destordeur, H., Willems, G., 2001. Spatial distribution and cell kinetics of the glands in the human esophageal mucosa. Eur. J. Morphol. 39, 163–168. https://doi.org/10.1076/ejom.39.3.163.4674

Van Nieuwenhove, Y., Willems, G., 1998. Gastroesophageal reflux triggers proliferative activity of the submucosal glands in the canine esophagus. Dis. Esophagus 11, 89–93. https://doi.org/10.1093/dote/11.2.89

von Furstenberg, R.J., Li, J., Stolarchuk, C., Feder, R., Campbell, A., Kruger, L., Gonzalez, L.M., Blikslager, A.T., Cardona, D.M., McCall, S.J., Henning, S.J., Garman, K.S., 2017. Porcine Esophageal Submucosal Gland Culture Model Shows Capacity for Proliferation and Differentiation. Cell. Mol. Gastroenterol. Hepatol. 4, 385–404. https://doi.org/10.1016/j.jcmgh.2017.07.005

Vu, T.N., Nguyen, H.-N., Calza, S., Kalari, K.R., Wang, L., Pawitan, Y., 2019. Cell-level somatic mutation detection from single-cell RNA sequencing. Bioinformatics. https://doi.org/10.1093/bioinformatics/btz288

Wang, Q., Jia, P., Li, F., Chen, H., Ji, H., Hucks, D., Dahlman, K.B., Pao, W., Zhao, Z., 2013. Detecting somatic point mutations in cancer genome sequencing data: a comparison of mutation callers. Genome Med. 5, 91. https://doi.org/10.1186/gm495

Wang, Z., Gerstein, M., Snyder, M., 2009. RNA-Seq: a revolutionary tool for transcriptomics. Nat. Rev. Genet. 10, 57–63. https://doi.org/10.1038/nrg2484

Watson, C.J., Papula, A.L., Poon, G.Y.P., Wong, W.H., Young, A.L., Druley, T.E., Fisher, D.S., Blundell, J.R., 2020. The evolutionary dynamics and fitness landscape of clonal hematopoiesis. Science 367, 1449–1454. https://doi.org/10.1126/science.aay9333

Watson, J.D., Crick, F.H., 1953. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. Nature 171, 737–738. https://doi.org/10.1038/171737a0

Weaver, J.M.J., Ross-Innes, C.S., Shannon, N., Lynch, A.G., Forshew, T., Barbera, M., Murtaza, M., Ong, C.-A.J., Lao-Sirieix, P., Dunning, M.J., Smith, L., Smith, M.L., Anderson, C.L., Carvalho, B., O'Donovan, M., Underwood, T.J., May, A.P., Grehan, N., Hardwick, R., Davies, J., Oloumi, A., Aparicio, S., Caldas, C., Eldridge, M.D., Edwards, P.A.W., Rosenfeld, N., Tavaré, S., Fitzgerald, R.C., 2014. Ordering of mutations in preinvasive disease stages of esophageal carcinogenesis. Nat. Genet. 46, 837–843. https://doi.org/10.1038/ng.3013

Werner, B., Case, J., Williams, M.J., Chkhaidze, K., Temko, D., Fernandez-Mateos, J., Cresswell, G.D., Nichol, D., Cross, W., Spiteri, I., Huang, W., Tomlinson, I.P.M., Barnes, C.P., Graham, T.A., Sottoriva, A., 2020. Measuring single cell divisions in human tissues from multi-region sequencing data. Nat. Commun. 11, 1035. https://doi.org/10.1038/s41467-020-14844-6

Wroblewski, L.E., Peek, R.M., Wilson, K.T., 2010. Helicobacter pylori and Gastric Cancer: Factors That Modulate Disease Risk. Clin. Microbiol. Rev. 23, 713–739. https://doi.org/10.1128/CMR.00011-10

Xian, W., Duleba, M., Zhang, Y., Yamamoto, Y., Ho, K.Y., Crum, C., McKeon, F., 2019. The Cellular Origin of Barrett's Esophagus and Its Stem Cells, in: Birbrair, A. (Ed.), Stem Cells Heterogeneity - Novel Concepts, Advances in Experimental Medicine and Biology. Springer International Publishing, Cham, pp. 55–69. https://doi.org/10.1007/978-3-030-11096-3_5

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

Yamamoto, Y., Wang, X., Bertrand, D., Kern, F., Zhang, T., Duleba, M., Srivastava, S., Khor, C.C., Hu, Y., Wilson, L.H., Blaszyk, H., Rolshud, D., Teh, M., Liu, J., Howitt, B.E., Vincent, M., Crum, C.P., Nagarajan, N., Ho, K.Y., McKeon, F., Xian, W., 2016. Mutational spectrum of Barrett's stem cells suggests paths to initiation of a precancerous lesion. Nat. Commun. 7, 10380. https://doi.org/10.1038/ncomms10380

Yizhak, K., Aguet, F., Kim, J., Hess, J., Kubler, K., Grimsby, J., Frazer, R., Zhang, H., Haradhvala, N.J., Rosebrock, D., Livitz, D., Li, X., Landkof, E.-A., Shoresh, N., Stewart, C., Segre, A., Branton, P.A., Polak, P., Ardlie, K., Getz, G., 2018. A comprehensive analysis of RNA sequences reveals macroscopic somatic clonal expansion across normal tissues. bioRxiv. https://doi.org/10.1101/416339

Yokoyama, A., Kakiuchi, N., Yoshizato, T., Nannya, Y., Suzuki, H., Takeuchi, Y., Shiozawa, Y., Sato, Yusuke, Aoki, K., Kim, S.K., Fujii, Y., Yoshida, K., Kataoka, K., Nakagawa, M.M., Inoue, Y., Hirano, T., Shiraishi, Y., Chiba, K., Tanaka, H., Sanada, M., Nishikawa, Y., Amanuma, Y., Ohashi, S., Aoyama, I., Horimatsu, T., Miyamoto, S., Tsunoda, S., Sakai, Y., Narahara, M., Brown, J.B., Sato, Yoshitaka, Sawada, G., Mimori, K., Minamiguchi, S., Haga, H., Seno, H., Miyano, S., Makishima, H., Muto, M., Ogawa, S., 2019. Age-related remodelling of oesophageal epithelia by mutated cancer drivers. Nature 565, 312–317. https://doi.org/10.1038/s41586-018-0811-x

Yoshida, K., Gowers, K.H.C., Lee-Six, H., Chandrasekharan, D.P., Coorens, T., Maughan, E.F., Beal, K., Menzies, A., Millar, F.R., Anderson, E., Clarke, S.E., Pennycuick, A., Thakrar, R.M., Butler, C.R., Kakiuchi, N., Hirano, T., Hynds, R.E., Stratton, M.R., Martincorena, I., Janes, S.M., Campbell, P.J., 2020. Tobacco smoking and somatic mutations in human bronchial epithelium. Nature 578, 266–272. https://doi.org/10.1038/s41586-020-1961-1

Zagari, R.M., Fuccio, L., Wallander, M.-A., Johansson, S., Fiocca, R., Casanova, S., Farahmand, B.Y., Winchester, C.C., Roda, E., Bazzoli, F., 2008. Gastro-oesophageal reflux symptoms, oesophagitis and Barrett's oesophagus in the general population: the Loiano-Monghidoro study. Gut 57, 1354–1359. https://doi.org/10.1136/gut.2007.145177

Zhang, L., Vijg, J., 2018. Somatic Mutagenesis in Mammals and Its Implications for Human Disease and Aging. Annu. Rev. Genet. 52, 397–419. https://doi.org/10.1146/annurev-genet-120417-031501