Benjamin Gernot Spiegl, BSc

Clustering Algorithm for High Resolution Somatic Variant Calling Using Unique Molecular Identifiers

MASTER'S THESIS

to achieve the university degree of

Diplom-Ingenieur

Master's degree programme: Biomedical Engineering

submitted to

Graz University of Technology

Supervisor

Dr. Christoph Wilhelm Sensen

Institute of Computational Biotechnology

Dr. Ellen Heitzer
Diagnostics & Research Institute of Human Genetics, Medical University of Graz

Graz, April 2019

AFFIDAVIT

I declare that I have authored this thesis independently, that I have not used other than the declared sources/resources, and that I have explicitly indicated all material which has been quoted either literally or by content from the sources used. The text document uploaded to TUGRAZonline is identical to the present master's thesis.

15th April 2019
Date

Benjamin Grimpel
Signature

Abstract

Introduction & Background: The detection of rare variants is of utmost importance in a variety of clinical scenarios, in particular for the analysis of cell-free circulating tumour DNA (ctDNA) from blood. As tumour-derived cell-free DNA is often underrepresented and massively diluted by DNA from normal cells, high resolution approaches are needed for the detection of somatic mutations, which are often present at frequencies below 1%. Therefore, high-resolution methods for mutation detection are needed which are able to distinguish mutations from background noise. To this end, unique molecular identifiers (UMIs) are used with Next Generation Sequencing (NGS) approaches in order to tag each template molecule in library preparation for a subsequent correction of PCR and sequencing errors. However, the bioinformatic downstream analyses are not well established and there is no pipeline available which enables a generalized analysis of various UMI-based approaches from different vendors. Therefore, it is unclear which combination of UMI protocol and variant calling software is most sensitive for detecting rare variants.

Methods: In this thesis, aspects critical to error suppression of various commercially available UMI-based NGS approaches were explored. A variant calling pipeline was developed and optimized. To this end, a well-characterized reference DNA was used as a ground truth for variant validation. Visualization of alignment clusters was used to assess the validity of a heuristic clustering procedure, which was developed for the analysis.

Results: Including UMI information in data pre-processing for variant calling only following general rules of thumb would have led to severe errors for certain UMI tagging protocols. A similarity-based clustering approach was developed which performed almost as good as a variant calling analysis conducted by QIAGEN, Venlo, NL. The performance was consistently high regardless of the utilized UMI tagging protocol or variant caller. SmCounter, a UMI-aware variant caller, was found to have a 10-fold lower detection limit than the well-known Mutect variant caller in tumour only mode. Somatic mutations with variant allele fractions of 0.125% were detected from 100 ng DNA samples. It was shown that ground truth variants were less affected by increasingly permissive clustering than the ambiguous background portion of variant calls.

Significance: These findings provide information on how to avoid pitfalls in processing UMI tagged paired-end sequencing data with respect to the applied tagging and targeting protocols. It is not always wise to apply rules of thumb for processing UMI-tagged sequencing data. Applying these rules would fail to correct for artefacts like premature strand synthesis termination. The clustering approach presented here also corrects for these artefacts and ensures a better reduction of alignments to the original library state.

Kurzfassung

Einführung & Hintergrund: Die Detektion von seltenen genetischen Varianten ist in vielen klinischen Anwendungen von großer Wichtigkeit, speziell bei Analysen von zellfreier Tumor-DNS (ctDNA) aus Blut. Da ctDNA durch DNS aus normalen Zellen stark verdünnt ist, werden hochauflösende Methoden zur Detektion benötigt, um Mutationen von Hintergrundrauschen unterscheiden zu können. Zu diesem Zweck werden eindeutige molekulare Identifikatoren (UMIs) in Kombination mit Next Generation Sequencing (NGS) Lösungen verwendet, um jedes DNS-Molekül während der Erstellung der Sequenzierbibliothek für eine spätere Korrektur von PCR-Fehlern und Sequenzierfehlern zu markieren. Die nachfolgende bioinformatische Analyse ist allerdings wenig etabliert. Weiters steht keine Software zur Verfügung, welche die generalisierte Analyse von UMI-basierten Lösungen verschiedener Anbieter ermöglichen würde. Ebenso ist unklar welche Kombination aus UMI-Protokoll und Variantenbestimmungssoftware (VBS) am empfindlichsten wäre um seltene Varianten zu detektieren.

Methoden: In dieser Arbeit wurden Aspekte der Fehlerunterdrückung durch kommerziell verfügbaren UMI-basierten NGS-Lösungen untersucht. Es wurde eine Softwarepipeline entwickelt und optimiert. Zu diesem Zweck wurde Referenz-DNS herangezogen um einen Satz an wahren Varianten zu erhalten. Visualisierungen von Clustern aus gemappten Sequenzreads wurden verwendet, um den entwickelten heuristischen Gruppierungsvorgang zu validieren.

Resultate: Die Verarbeitung der UMI-Information gemäß Faustregeln würde für manche Markierungsprotokolle zu Fehlern führen. Es wurde eine ähnlichkeitsbasierte Herangehensweise zur Gruppierung entwickelt, die vergleichbare Resultate in der Variantenbestimmung erzielte wie eine Analyse von QIAGEN, Venlo, NL. Die Leistungsfähigkeit dieser Gruppierung war unabhängig vom verwendeten UMI-Protokoll und der verwendeten VBS. SmCounter erreichte ein zehnmal niedrigeres Detektionslimit als die bekannte Mutect VBS im Nur-Tumor-Modus. Somatische Mutationen mit einer Variantenallelfrequenz von 0.125% konnten aus 100 ng DNS-Proben detektiert werden. Es wurde gezeigt, dass bekannte wahre Varianten weniger durch tolerantere Gruppierung beeinflusst werden als Hintergrundvarianten.

Signifikanz: Die hier präsentierten Erkenntnisse bieten wichtige Informationen, um Verarbeitungsfehler von UMI-markierten, gepaarten Sequenzread-Daten unter Berücksichtigung der angewandten Markierungsprotokolle zu vermeiden. Würden in der Praxis nur Faustregeln zur Datenverarbeitung angewendet, so würden Artefakte wie vorzeitiger Syntheseabbruch nicht korrigiert. Der vorgestellte Gruppierungsalgorithmus korrigiert auch diese und bietet eine bessere Rückrechnung der Sequenzierdaten auf den Ursprungszustand der Sequenzierbibliothek.

Acknowledgments

First, I would like to offer my special thanks to Prof. Christoph Sensen for supervising this thesis, providing me with the computational power I needed for carrying out my analyses, and for his rich feedback concerning scientific writing.

This work would also not have been possible without the support from the Diagnostics and Research Institute of Human Genetics at the Medical University of Graz. I was tremendously fortunate to be recognized by Prof. Michael Speicher and invited to do this thesis.

Advice given by Prof. Ellen Heitzer and Peter Ulz, PhD, who always lent me an ear despite being buried in tons of work, has been a great help.

I am also very grateful for being permitted to use data which was created by many hard-working hands at the Medical University. I am particularly grateful for the validations provided by Sabrina Weber, MSc, which enabled me to benchmark my results against a manufacturer analysis procedure.

The constructive discourse I had with colleagues from the University of Technology as well as the Medical University during the time of coding and writing was greatly appreciated. I want to thank Max Malek, MSc, Dipl.-Ing. Stefan Grabuschnig, Verena Rupp, PhD, and Isaac Lazzeri, MSc.

Last, I wish to thank the many people who supported me with their encouragement and including my family, girlfriend, and friends. To all dear people that passed away when I was occupied with work: I wish there had been more time. You will not be forgotten.

Contents

List of Abbreviations

Abbreviation	Meaning	Abbreviation	Meaning
A	Adenine	IGV	Integrative Genomics Viewer
API	Application programming interface	Indels	Insertions and/or deletions
BAM	Binary alignment/map	LOESS	Locally estimated scatterplot smoothing
BED	Browser extensible data	MIT	Massachusetts Institute of Technology
BLAST	Basic Local Alignment Search Tool	MOC	Multiple observations-supported variant call
BSD	Berkeley Software Distribution	MUG	Medical University of Graz
BWA	Burrows-Wheeler aligner	NCBI	National Centre for Biotechnology Information
C	Cytosine	NGS	Next generation sequencing
cfDNA	Circulating cell-free DNA	OS	Operating system
CIGAR	Compact idiosyncratic gapped alignment report	PCR	Polymerase chain reaction
CNA	Copy number alteration	PSF	Python Software Foundation
ddPCR	Digital droplet PCR	RE	Runtime environment
DNA	Deoxyribonucleic acid	SAM	Sequence alignment/map
dNTP	Deoxy nucleoside triphosphate	SD	Standard deviation
DOC	Dual observation-supported variant call	SNLR	Signal-to-noise level relation
D & R	Diagnostic & Research	SNV	Single nucleotide variant
EFP	Estimated false positives filter	SOC	Single observation-supported variant call
FTP	File transfer protocol	T	Thymine
G	Guanine	TUG	University of Technology, Graz
GATK	Genome analysis toolkit	UCSC	University of California, Santa Cruz
GPL	GNU General Public License	UMI	Unique molecular identifier
GTV	Ground truth variant	VAF	Variant allele frequency
HSNLR	Half-signal-to-noise level relation	VBS	Variantenbestimmungssoftware
ICBT	Institute for Computational Biotechnology	VCF	Variant call format
IDE	Integrated development environment	WT	Wildtype

1. Introduction

The Human Genome

Assembling the complete sequence of the human genome was pursued by the human genome project. This goal was officially achieved in 2003 [1, p. 54]. Since then, obtaining deoxyribonucleic acid (DNA) sequence information from patients has gained tremendous importance for diagnostics and treatment of various diseases. The cost of sequencing an entire human genome has decreased from $100 million in 2001 (4) to around $1,500 in 2017 [2] and might decrease further in the future. Studies in the field of 'omics' technologies combined with systems biology approaches thrive to describe the aggregated information of an organism with respect to its DNA methylation status, its genes, encoded RNA transcripts, translated proteins, as well as other sources of molecular information such as lipids and metabolites. Insights from omics, ontologies and molecular pathways greatly furthered our understanding of the human biology from a molecular to a system-wide scope. For example, a successful application of omics technologies was reported in a study that investigated drug induced cell stress [3, p. 11].

In medicine, omics technologies have helped in moving the focus from symptom-based treatment to patient-centred approaches. The massive accumulation of comprehensive knowledge about the human organisms allows for accurate interpretation of easily accessible molecular profiles and metabolite levels amongst other biomarkers such as small variations of the DNA sequence.

International efforts successfully established comprehensive databases to catalogue variants in special contexts like OMIM [4] (genetic disorders), to aggregate data of certain types like gnomAD [5, p. 16ff] (exon and genome sequencing data), or various tumours, which aid in the interpretation of disease-associated genetic alteration such as single nucleotide variants (SNVs), short deletions or sequence insertions (indels), somatic copy number alterations (CNAs), and gene fusions.

Genetic Variation and Variant Detection

As a result of the 1000 Genomes Project, the average healthy human individual was estimated to carry around 180 sites with protein truncating variants, up to 12,000 variants altering peptide sequence, up to 2,500 structural variants, and more than half a million variants overlapping with known regulatory regions [6, p. 1f]. The number of bases affected by structural variation was estimated to be around 18.4 Mbp [7, p. 5]. Therefore, it is of utmost importance to distinguish harmless mutations from those that have the potential to act in tumorigenesis or cause genetic disorders. Nevertheless, healthy individuals might also carry disease-associated variants which

predispose to certain diseases. It was estimated that approximately 60 missense variants that severely damage protein structure and approximately 100 loss-of-function variants are carried by each healthy individual [8, p. 9].

The identification of genetic variants is based on complex bioinformatics algorithms and the default analysis pipeline for detecting mutations can be simplified as a sequence of the following steps:

1) sampling and purifying DNA material from an affected individual
2) determining the base sequence of sample DNA material through sequencing
3) applying quality checks and pre-processing to obtained sequence reads
4) mapping reads to the organism's reference genome involving sequence alignment
5) alignment processing (*e.g.* marking duplicates)
6) application of a variant calling algorithm to:
    a) determine sequence deviations from the reference genome or
    b) region copy number aberrations relative to other sample regions.

Cancer

Cancer comprises a large group of multifactorial, complex diseases that involve abnormal cell growth. During a lifetime, cells accumulate a variety of mutations and if these mutations affect so called cancer genes, tumours might arise. There are two types of genes involved in tumorigenesis: tumour suppressor genes, which control the cell cycle or initiate apoptosis, and (proto-)oncogenes, which normally promote cell growth, proliferation or inhibition of apoptosis. According to the two-hit hypothesis, at least two mutational events must occur in a tumour suppressor gene in a cell for it to start malignant tumour growth. In contrast, for the activation of an oncogene, a single mutation is sufficient to confer a growth advantage. In late stage cancers, tumour cells may also migrate to tissues distant to the primary tumour site in a process called metastasis formation.

An example of a tumour suppressor gene commonly mutated in tumours is TP53. In healthy individuals, the expression of the unstable p53 protein is stabilized upon DNA damage which leads to senescence allowing for repair mechanisms to act. In case that the DNA repair mechanism failed, stabilized p53 protein induces apoptosis [9, p. 1f].

Precision Oncology and Liquid Biopsies

Although treatment advances have led to improved response rates and survival for a variety of tumour types, early detection and longitudinal monitoring of patients is the only means to accomplish the best possible treatment outcome.

Most cancers would be curable if detected early [10, p. 29]. Currently, screening routines capable of detecting cancer in early stages are only established for breast, cervical, colorectal, and prostate cancer. In some cases, the benefit is under debate because of overdiagnosis and treatment side effects [11, p. 4f] [12, p. 7ff]. Aside of a wider repertoire of screening tests, also the availability of a cost-effective monitoring of tumour response to treatment would help in quick adaptation of medication. In this context, the use of so-called liquid biopsies, which include the analysis of circulating tumour components such as cells or cell-free DNA, is a promising tool for early cancer detection or early detection of recurrence, and identification of actionable targets or resistance mechanisms [13, pp. 1-3]. Therefore, biomarkers accessible from liquid biopsies such as blood or urine samples might become an alternative to invasive solid tissue biopsies in diagnosis and tumour classification. Moreover, in some cases, solid biopsies were shown to accelerate migration of tumour cells into neighbouring tissues and distant metastasis formation [14, p. 8]

Liquid biopsies were shown to be of diagnostic value in non-invasive therapy monitoring and detection of cancer-causing mutations 1 up to 2 years in advance of diagnosis [15, p. 5]. In a successful therapy monitoring example, response of a metastatic prostate cancer to treatment with two androgen receptor axis-targeting drugs could be detected by keeping track of region copy numbers assessed from shallow sequenced plasma-Seq samples [16, p. 6f].

Liquid Biopsy Challenges

Circulating tumour DNA (ctDNA) is released into the circulation through several mechanisms [17, p. 2] [18, p. 3] [19, p. 5f] [20, p. 3ff]. ctDNA can be used as a biomarker in clinically informative genomic profiling. This is done by extracting information present in DNA fragments like SNVs and indel variants of cancer-associated genes as well as detection of fusion genes. The amount of KRAS alleles circulating tumour DNA was found to correlate with tumour burden and, therefore, increase over the four tumour stages. Furthermore, the amount of detectable ctDNA depends on the type of cancer [21, pp. 2f, 5]. These circumstances highlight the necessity of detecting low tumour variant allele frequencies (VAFs) from liquid biopsy samples in early cancer detection.

A further limitation of ctDNA is the relatively low amount of material released from early stage tumours and certain types of cancers. The amount of ctDNA seems to depend on tumour

size [22, p. 6]. For example, the number of circulating KRAS mutated alleles was found to correlate with tumour burden and, therefore, increases with advanced tumour stages [23, p. 4f]. To observe a mutant DNA frequency of 0.1%, a spherical tumour would need to reach a volume of 10 cm^3 [24, p. 5].

This aggravates the use of ctDNA of early cancer detection [25, p. 4] [26, p. 4]. The screening for single or few mutations alone does not reach sufficient sensitivity. However, combinations of biomarkers were shown to increase sensitivity of detecting pancreatic adenocarcinoma [27, p. 4] and ovarian cancer [28, p. 5] compared to tests relying on a single type of biomarkers. The detection of tumour-specific methylation patterns [29, p. 8f] was also described to be useful in early diagnosis of disseminated breast cancer up to one year in advance of conventional diagnosis [30, p. 7f]. Also, ovarian cancer could be correctly diagnosed up to two years in advance [31, p. 12]. Methylation patterns can be used in tissue deconvolution to identify the tissue of origin in cases of cancers with unknown primary tumour site [32, p. 9].

Furthermore, the amount of detectable ctDNA depends on the type of cancer and not all tumour types are equally suited for ctDNA-based analyses [21, pp. 2f, 5].

UMI-Based Approaches

The above-mentioned liquid biopsy challenges highlight the necessity of detecting low tumour VAFs. Using NGS-approaches, the VAFs in many ctDNA samples is in the range of background noise. To be able to distinguish DNA replication errors from true variants, methods involving tagging of the original unamplified sequencing library with so-called unique molecular identifiers (UMIs) were developed. Other terms describing molecular identifiers may be encountered in literature: molecular tag, molecular barcode, or simply barcode. Targeted sequencing panels can be combined with these products to reduce sequencing to regions of interest which increases cost and time efficiency.

Currently, tests relying on the detection of low VAF variants without making use of noise filters or suppressing measures are unlikely to reach clinical validity, even if an effective molecular tagging strategy is combined with sensitive variant calling. It was only recently that the required quality of variant calling results was obtained by applying deep learning methods for sequence specific noise suppression and removal. The corresponding computational error suppression approach was presented by Newman *et al.* [33, p. 4].

Several other successful applications of UMI protocols for detection of variants with VAFs below 1% from liquid biopsy samples were recently reported [34, p. 1]. The focus of these

publications was often put on the validity of the protocol and not on the validity of UMI tagged data processing. In many cases, an ideal data situation was outlined and used to justify simplified bioinformatic approaches. While - in some cases - this might be valid [35, p. 2f], detailed descriptions and validations of bioinformatic approaches like alignment grouping and filtering were missing in others [36, p. 3]. Furthermore, investigations of clustering strategies including parameter optimization were not available at beginning of this thesis.

Aims and Expected Results

It was the main goal of this thesis to shed light into the ambiguous area of processing UMI tagged circulating cell-free DNA (cfDNA) sequencing data and to develop recommendations based on variant calling validation. Also, the suitability of different combinations of UMI tagging protocols and variant callers was to be validated based on their variant calling performance. Finally, a piece of software implementing the developed recommendations was to be created.

It was expected to observe non-ideal UMI group size distributions in empirical data that greatly differ from a perfectly constant group size across all UMI groups. Polymerase errors and sequencing errors were expected to distort the group size distribution by creating new, erroneous UMI sequences. In addition to single base substitutions, insertions and deletions should be identifiable to cause UMI sequence alterations. Thus, the error rate of alignments governed by a UMI group was expected to exhibit some variability.

Unravelling the nature of UMI-altering artefacts and increasing the specificity of alignment reduction should yield an improved tumour-only variant calling performance. This measure is mainly defined by ground truth variant (GTV) recall and by the number of background calls. The variant calling optimization can be analysed after the error correction step and after the complete library reduction by using two types of variant callers:

- a barcode-aware variant caller
- a variant caller that requires *in silico* reduction of the amplified sequencing library data

A comparison of these variant callers was expected to show an advantage of using the more recent barcode-aware variant caller.

Molecular noise in liquid biopsy sequencing experiments using Illumina platforms mainly consists of sequencing errors and base substitution errors occurring during amplification of the sequencing library. The latter were shown to depend on the library preparation method [37, p. 7ff]. Rare formations of chimeric sequences may occur prior to sequencing in amplification by polymerase

chain reaction (PCR). In case of targeted sequencing approaches, sequence-dependent errors may accumulate at certain reference positions which can lead to false positive variant calls.

In theory, base calling errors along with molecular noise could be eliminated by creating multiple copies of template molecules from the original (unamplified) state of the biopsy sample. These copies had to share a unique label which was added during the unamplified state of the sequencing library.

The goal of the UMI tagging approach combined with subsequent bioinformatic analysis was to obtain a better picture of the original sequencing library state compared to standard procedures which do not correct for polymerase errors or base calling errors. Before attempts were made to use UMI information for noise suppression, it was common practice to carry out deduplication. This procedure omits all copies of a template molecule but one. Therefore, deduplication is more of an error avoiding strategy rather than an active error correction method. Polymerase errors and base calling errors are indistinguishable from true variants that occur at low frequency and are therefore the major complication in computing the original state of the sequencing library.

There are two prerequisites for variant calling in the low variant frequency domain:

- full correction of base calling errors and molecular noise
- preservation of the original mutant allele to normal allele ratio

In practice, this is not the case which results in false positive and missed variant calls as well as VAF distortions. False positive calls may result in an overestimation of GTV recall. Therefore, the extent of false positive calls must be corrected or at least estimated. Recall and allelic frequency accuracy as the most important performance measures amongst others collectively form the performance of variant calling approaches on UMI tagged liquid biopsy data.

Control data sets allow for an estimation of the level of false positive GTV calls which is innate to the whole analysis procedure from drawing the sample to variant calling. Thus, an uncertainty measure for the conducted analysis based on a dilution series was introduced. Ideally, a control dataset would yield no GTV calls. Such an analysis can then be viewed as unbiased in a sense that all observed GTV calls are due to an efficient wet lab approach that was paired with performant *in silico* suppression or even removal of noise. In the opposite case, the number of artefacts that may have been caused by sample contamination, uncorrected molecular noise, and/or base calling errors can be estimated.

2. Materials and Methods

In this thesis, different UMI tagging procedures were used for library preparation. The individual variant calling performance resulting from their use was assessed. To ensure uniform bioinformatic analysis, a variant calling pipeline was developed and optimized. Pipeline optimization across samples was carried out to erase any occurring bias introduced by *in silico* analysis. Reference DNA material was used to obtain a priori knowledge of variants contained in each sample. These variants formed a ground-truth set which is used in variant calling validation. The use of a set of ground-truth variants and an optimized pipeline rendered results comparable. Variant calling was tested for two callers: Mutect and smCounter.

At first, a minimalistic exploratory analysis was conducted to identify problems of UMI integration and exploitation for noise reduction in an analysis pipeline for somatic variant calling. This pipeline was derived from the Broad Institute's best practices for somatic variant calling using their genome analysis toolkit (GATK) software. Best practices were developed by the GATK team [38]. The GATK contains the Mutect variant caller.

Findings from the exploratory analysis were incorporated in the development process of a variant calling pipeline. This software was used in reanalysis to assess performance measures for data created by an Illumina NextSeq sequencing machine and to choose the best suited UMI tagging procedure. Pipeline parameters were optimized based on variant calling performance. The variant caller best suited for low VAF variant calling was chosen based on these performance results. These were then benchmarked against results of a data analysis service offered by QIAGEN, Venlo, NL.

2.1. Reference DNA Material

To assess the analytical sensitivity and specificity of a test, usually reference materials are used as ground truth. In the context of next generation sequencing and the detection of rare variants, mostly cell line DNA or synthetic DNA oligonucleotides with defined variant allele frequencies are used. Here, standard material from two different vendors, *i.e.* Horizon and SeraCare, was used. These materials include clinically-relevant single nucleotide variants (SNVs) and indels (for details see table 2.1 and 2.2).

The Tru-Q HDxTM number 4 ('TruQ4') standard is based on three colon-carcinoma cell lines: RKO (ATCC®, Virginia, USA: CRL-2577™), SW48 (ATCC®: CCL-231™), and HTC 116 (ATCC®: CCL-247™). SW48 and RKO are cell lines from female individuals, while HTC 116 stems from a male sample. The cfDNA of an anonymous person was used as a wildtype control

for experiments including the TruQ4 reference material. This individual was diagnosed healthy after being screened for various types of cancer and, thus, this cfDNA was regarded as ctDNA-free.

Table 2.1: details of reference materials used in sequencing experiments.

Reference Material Name	Manufacturer	Number of Variants	VAF Range	VAF Deviation[1]	WT[2] Control
Tru-Q HDx™ number 4	Horizon Discovery Group plc, Cambridge, UK	14	5% - 30%	1% - 3%	No
Seraseq™ ctDNA Reference Material v2	SeraCare Life Sciences, Inc., Milford, MA, USA	42	0.125% - 2%	-	Yes

1 The VAF deviations of the Horizon Discovery reference are given in absolute VAF and depend on the VAF of each variant (extremes shown). SeraCare publishes the VAFs measured with digital droplet PCR (ddPCR) for every batch.
2 WT, wild type

In contrast, the SeraCare reference standard was manufactured from a single reference cell line GM24385 (Coriell Institute for Medical Research, Camden, USA), a male B lymphocyte sample, along with synthetic DNA constructs. The Seraseq™ ctDNA Reference Material v2 ('Seraseq' reference material) is available at several VAF levels, *i.e.* all contained variants share the same VAF. The VAF levels used in the Seraseq dilution series are as follows: 2%, 1%, 0.5%, 0.25%, 0.125%, and 0% (WT) control.

For the validation of the next generation sequencing (NGS) assays, two DNA mixtures were created from the TruQ4 reference material and the wildtype control DNA: 25% TruQ4 plus 75% wildtype DNA, and 50% each. Both mixtures along with 100% TruQ4 reference material and 100% wildtype control form the samples of the NEBNext [39] TruQ4 dilution series (percentages tumour/wildtype: 100/0, 50/50, 25/75, 0/100). The smMIP [40] TruQ4 dilution series, in contrast, did not include a 100% wild type control sample.

Quality control of GTV VAF of DNA reference materials was carried out per batch by the corresponding manufacturer utilizing ddPCR. As a result, a pool of GTVs at certain allelic frequencies with limited variability could be used for validation of variant calls.

Table 2.2: variants detectable by utilized variant callers covered by reference materials.

Genes	DNA Reference Material		Chromosome	Variant Type[1]
	Tru-Q HDx™ number 4	Seraseq™ ctDNA reference material v2		
MPL	-	Covered	1	SNV[2]
NRAS	Covered	Covered	1	SNV S, T
IDH1	-	Covered	2	SNV
PIK3CA	Covered	Covered	3	INS S, SNV S, T
CTNNB1	-	Covered	3	SNV
FGFR3	-	Covered	4	SNV
PDGFRA	Covered	Covered	4	INS S, SNV S, T
KIT	Covered	Covered	4	SNV S, T
APC	-	Covered	5	INS, SNV
NPM1	-	Covered	5	INS
EGFR	Covered	Covered	7	DEL S, INS S, SNV S, T
BRAF	Covered	Covered	7	SNV S, T
JAK2	-	Covered	9	SNV
GNAQ	-	Covered	9	SNV
ABL1	Covered	-	9	SNV
RET	-	Covered	10	SNV
PTEN	-	Covered	10	DEL, INS
ATM	-	Covered	11	DEL
KRAS	Covered	Covered	12	SNV S, T
FLT3	-	Covered	13	SNV
AKT1	-	Covered	14	SNV
IDH2	Covered	-	15	SNV
TP53	-	Covered	17	DEL, SNV
ERBB2	-	Covered	17	INS
SMAD4	-	Covered	18	INS
GNA11	-	Covered	19	SNV
GNAS	-	Covered	20	SNV

[1] In case both reference materials cover a gene listed in this table, the respective variant types for each reference material are labelled by superscript letters: 'S' for Seraseq and 'T' for TruQ4. Gene fusions were omitted.

[2] SNV, single nucleotide variant

Reference DNA was manufactured from engineered cell lines and technical constructs. Mechanical shearing ensured obtaining the desired fragment length distribution. After DNA mixing, the mix was stabilized for shipment and subsequent use. Additional test methods for fragmentation size

and quantification of the product from Horizon Discovery were D1000 DNA ScreenTape assay and Qubit dsDNA BR Assay (post-fragmentation) respectively [41].

2.2. Sample Preparation

All data used in this thesis were generated at the D&R Institute of Human Genetics at the Medical University Graz (MUG). A total of 4 different NGS assays including one hybrid capture based protocol and 4 amplicon-based protocols were tested.

Wet lab work was performed by employees of the D & R Institute of Human Genetics. The generalized procedure of sample preparation was as follows:

1. cfDNA extraction from reference DNA samples according to the plasma-Seq protocol [42, p. 9] (no shearing or lysis of exosomes)
2. capture (extra step required for ThruPLEX® protocol [43]) and UMI tagging of isolated cfDNA molecules including 1 to 3 rounds of PCR according to the chosen molecular tagging procedure
3. pre-sequencing amplification step with several PCR cycles also depending on the tagging procedure of choice to achieve high amplification of tagged templates (see table 2.3)
4. sequencing with one of the platforms described in subsection 'Sequencing'

To describe the state of order of the sequencing library prior to UMI tagging, the term 'original state' will be used. This term is intended to describe a state of order where all molecules are considered flawless regarding the base sequence as found within the liquid biopsy sample (here: DNA reference material) and, thus, represents the true or 'original' information contained in the sample. Information entropy [44, p. 14ff] of the sequencing library is increased with every cycle of amplification due to DNA polymerase errors which mutate copies of original template molecules during synthesis. The attempt to compute the original state from the disordered amplified state, which would be equivalent to solving an inverse problem, will be termed 'alignment reduction' to distinguish it from deduplication.

2.3. Sequencing Panels and Tagging

Four UMI tagging protocols were used in sample preparation (table 2.3). Each procedure differs in the method of UMI ligation or hybridization, the amount of UMI sequences per tagged molecule, and the number of bases per UMI tag as well as the enrichment of target regions.

NEBNext, QIASeq [45], and smMIP protocols involve either a single capture probe or gene specific primers. These protocols create an amplicon-like alignment distribution after read mapping. The ThruPLEX protocol, in contrast, is an untargeted tagging procedure and simply tags every DNA template. However, a set of genes was enriched using a custom Rapid Capture Panel (Illumina).

Table 2.3: summary of molecular tagging protocols.

Protocol Name	Manufacturer	Capturing Method	UMI Bases	Enrichment Cycles
NEBNext® Ultra II DNA Library Prep for Illumina	New England BioLabs®, Inc., Ipswich, USA	Hairpin loop adapter ligation, complementary UMI adapter ligation	8 or 12	4 - 9
QIASeq™ Targeted DNA Panel	QIAGEN, Venlo, NL	5'-end UMI adapter ligation, gene specific primer	12	7 - 9
smMIP	D & R Institute of Human Genetics, (in-house replicate)	Molecular inversion probe	10	26
ThruPLEX® Tag-seq Kit	Takara Bio USA, Inc., Mountain View, USA	None (5' and 3' stem-loop adapter ligation)	12	4 - 11

The number of PCR cycles required for sufficient amplification depends on the amount of input material. For example, the NEBNext protocol requires 9 enrichment cycles for 1 ng input material and 4 cycles in case of 100 ng input material for optimal yield. The amount of cycles required for the ThruPLEX and the smMIP protocol also needs adaptation depending on the amount of input DNA. The smMIP protocol was replicated at the D & R Institute of Human Genetics following instructions described by Hiatt and colleagues [40, p. 9ff].

2.4. Sequencing

All libraries were sequenced in a paired-end mode at the D & R Institute of Human Genetics on a MiSeq and/or NextSeq machine (both Illumina, Inc., San Diego, USA). Sequencing run and output parameters of these platforms are shown in table 2.4.

For NEBNext data sets, a workaround for *in silico* UMI sequence extraction was used. The functionality of Illumina devices to write sample index sequences to a separate file was adapted to extract the UMI sequence. To achieve an offset effect, upstream bases were 'N'-masked. For data

of other tagging protocols, the UMI sequence had to be extracted from the read-sequence during bioinformatic analysis. Both UMI sequence extraction methods disregarded the possibility of early indel artefacts or other sequence shift artefacts.

Table 2.4: expected run and output parameters of Illumina sequencing platforms and paired-end reads.

Platform[2]	Read Length	Approximate Total Time	Output	Reads Passing Filters[1]	Average Base Quality Scores Above Q30
MiSeq	2 x 150 bp	24 h	4.5 - 5.1 Gb	24-30 M	>80%
NextSeq	2 x 150 bp	26 h	32.0 -39.0 Gb	260 M	>70%

[1] Reads passing filters is based on Illumina PhiX control library with supported cluster densities at 865-965 k/mm² and 129-165 k/mm² for the MiSeq and NextSeq instrument respectively.
[2] Manufacturer information was taken from the Illumina website [46].

2.5. Data Sets

Depending on the six factors 'reference material type', 'amount of input material', 'tagging protocol', 'sequencing device', 'replicate', and 'dilution series', sequencing data was organised in several data sets (see table 2.5). In general, data sets with TruQ4 reference material were mainly used in the exploratory data analysis. Seraseq reference material data sets, which became available later during the thesis, were utilized in a reanalysis for parameter fine-tuning, performance assessment and analysis pipeline benchmarking.

Table 2.5: information on data sets created for this thesis.

Data Set Name	Reference Material	Input Material Amount	Tagging Protocol	Sequencing Device	Replicate	Mix Levels
QIASeq TruQ4	TruQ4	10 ng	QIASeq	MiSeq	-	-
NEBNext 10 ng	TruQ4	10 ng	NEBNext	MiSeq	-	-
NEBNext 100 ng	TruQ4	100 ng	NEBNext	MiSeq	-	-
NEBNext TruQ4	TruQ4	100 ng	NEBNext	MiSeq	-	4
ThruPLEX MiSeq	TruQ4	100 ng	ThruPLEX	MiSeq	-	-
ThruPLEX NextSeq	TruQ4	100 ng	ThruPLEX	NextSeq	-	-
smMIP	TruQ4	10 ng	smMIP	MiSeq	1	3
NEBNext Seraseq	Seraseq	100 ng	NEBNext	NextSeq	-	6
QIASeq Seraseq	Seraseq	100 ng	QIASeq	NextSeq	-	6

Dilution series were used to assess the performance degradation for decreasing GTV VAFs. Replicates were used to test concordance of repeated analyses. Control data was utilized to estimate

the number of false positive variant calls. The ten-fold input material difference between two NEBNext data sets was intended to show the influence of the input DNA amount on the variant calling outcome.

2.6. Development Environment

Hardware

The analysis software was developed on a Lenovo™ ThinkPad E470 notebook, subsequently named 'local hardware'. Computationally inexpensive read data quality checks and other statistical analyses including variant calling validation were performed on this hardware.

All full data analyses including variant calling were carried out on a HP DL580 server of the seventh generation (512 GB RAM, 32 cores, 1 PB conventional disc space, 12 TB SSD array in RAID-0 configuration). This server was property of the Institute for Computational Biotechnology (ICBT) at the University of Technology Graz (TUG). Contamination checks, which required mapping of reads to several genomes, were also carried out on the server.

The operating system (OS) for the local hardware was chosen according to the most common OS used for bioinformatic analyses at the D & R Institute of Human Genetics. An Ubuntu 16.04 OS (64-bit) was set up on the local hardware accordingly. The server's OS was CentOS Linux release 7.5.1804 (Core).

The programming language Python, version 2.7 [47], was a prerequisite defined by the D & R Institute of Human Genetics for the development of a variant calling pipeline. This software must be capable of processing UMI-tagged paired-end reads for variant calling. In this thesis, open-source software and freeware was preferred over closed source, proprietary software in application programming interfaces (APIs) selection.

Software and Requirements

An overview of software dependencies, licenses, versions, and references of tools and packages used in this thesis is given in table 2.7. References for non-standard library Python packages are listed in table 2.6.

A variety of publicly available tools were used for data quality examination that were not incorporated in the developed Python analysis pipeline. The sequencing quality of received paired-end read data and optional associated UMI data were checked using Babraham Bioinformatics' FastQC software [48]. The FastQ Screen software [49] was used for impurity screening, *i.e.*

searching for reads mapping only to a genome other than the human genome or which are completely unmapped. The Python-based summarising tool MultiQC [50] was used to aggregate FastQC and FastQscreen output. Quality checks were carried out manually using the UNIX command line or graphical user interfaces if available. Furthermore, the statistical programming language R [51] along with the visualisation R-package ggplot2 [52] were used in variant validation result visualization, which was implemented in a separate script.

There are four types of requirements for running the developed analysis pipeline: a computer with a UNIX OS, Python version 2.7, non-standard library Python packages accessible for the required Python version and their non-Python dependencies must be installed. The nine command-line tools mentioned below must be available (see table 2.7). The modified smCounter code must reside in a callable location from within a Python environment. Tools were made available from anywhere on the system by manually creating a soft link (or symbolic link) inside a 'bin' folder (*e.g.* '/usr/bin'), or by adding an export entry containing the path of the executable to the hidden 'bashrc' file in the user's home directory. Both measures require root access.

Table 2.6: Python packages used in the analysis pipeline code.

Package Name	Imported Functions[1]	Version	License	Reference
matplotlib	patheffects; pyplot, **ticker:** MaxNLocator; use	2.2.3	BSD, PSF[2]-based	[53]
natsort	humansorted	5.4.1	MIT	[54]
numpy	amax; amin; arrange; array; float64; isin; log2; log10; mean; median; std; unique; vectorize; where	1.15.1	BSD	[55]
pandas	DataFrame	0.23.4	BSD	[56]
pysam	AlignmentFile	0.15.0	MIT	[57]
pyvcf	*	0.6.8	BSD, MIT	[58]
scipy	**stats.stats:** kendalltau, linregress, pearsonr, spearmanr	1.1.0	BSD	[59]
seaborn	*	0.9.0	BSD	[60]

[1] An asterisk denotes an import of the whole package. A bold name preceding a colon followed by a list of imports denotes a module name inside a package which contains the list of imported functions and/or objects.

[2] Python Software Foundation

In case of smMIP data sets, adaptor trimming of reads was required. Trimming was carried out using the cutadapt software. Reads were aligned to the latest release of the human reference genome version GRCh37.p13 (hg19) with the BWA short read aligner [61]. The versatile SAMtools

software [62] was used for conversions between sequence alignment/map (SAM) and binary alignment/map (BAM) representations of mapped paired-end read data, for creation of BAM index files, and to gather mapping statistics. The BAMtools toolkit [63] was also used to gather mapping statistics of BAM files. The Picard toolkit [64], another set of useful command-line tools from the Broad Institute, was employed for SAM file sorting. Overlapping target regions as defined in the input browser extensible data (BED) file were merged using the utility software BEDTools [65]. Furthermore, the `sort` command of Ubuntu, a command included in the GNU Core Utilities package called 'coreutils', was applied to lexicographically sort regions inside BED files, which is required for chromosome files larger than 512 Mb, as stated on the BEDTools website [66]. Group-wise UMI sequence errors were corrected using the novel UMI-tools software.

Table 2.7: details of tools used in this thesis.

Tool/Software[1]	Version	Dependencies[2]	License	Reference
FastQC	0.11.5	Java RE	GPL[3] v3 or later	[48]
FastQ Screen	0.11.4	BWA, Bowtie, Bowtie2, or Bismark, Perl, GD-Graph*	GPL v3 or later	[49]
MultiQC	1.3	Python 2.7 or later	GPL v3 or later	[50]
cutadapt	1.14	Python 2.7 or later	MIT[4]	[67]
BWA[6]	0.7.12 - r1039	-	GPL	[61]
SAMtools	0.1.19 - 96b5f2294a	HTSlib*	MIT/Expat	[62]
BAMtools	2.5.1	CMake*, JsonCpp*	MIT	[63]
Picard toolkit	1.128	Java RE, HTSlib*	MIT	[64]
BEDTools	2.25.0	zlib*	GPL v2	[65]
UMI-tools	0.5.4	Python 2.7 or later, various Python packages	MIT	[68]
Mutect	4.0.8.1	Java 8 RE	BSD[5]	[69]
smCounter	Modified online version of August 8th 2017	Python 2.7 or later, BEDTools	MIT	[70]

[1] All GATK tools of the listed version require Java runtime environment version 8.

[2] Dependencies marked with an asterisk may not be required in all cases because they are either required only for newer versions, packed with a precompiled version of the software/toolkit, or are part of most UNIX-based operating systems.

[3] GNU General public license

[4] Massachusetts Institute of Technology

5 Berkeley Software Distribution

6 Burrows-Wheeler aligner

Two variant callers were tested: Mutect [69], distributed by the Broad Institute as part of the GATK and capable of tumour-only variant calling, and the novel smCounter software by Qiagen Sciences, Inc., Frederick, USA [70]. The original code of smCounter was adapted for reasons of pipeline output control. Minor changes were included to comply with the pipeline's mode of calling external tools and to redirect error output according to the pipeline's folder structure. The core code implementing the variant calling process was not changed.

Package imports were limited to keep a clean namespace for faster lookup in Python. Two non-standard packages were used for visualization of analysis statistics: matplotlib and seaborn. The latter requires data organization using pandas `DataFrame` objects. The `humansorted` function imported from the natsort package allowed for sorting more complex strings following a human-like logic. The pysam package is a Python port of the HTSlib C programming language API [71] which was useful for SAM/BAM file processing. It required the HTSlib 1.7 to be installed. The scipy package, a library for scientific computing, was used in variant call validation for linear regression analysis and for calculating different correlation measures. The validation of called variants was implemented in a separate script due to its intended use of validating multiple variant calling format (VCF) files which form a combined result of a dilution series analysis.

2.7. Data Quality Control

Prior to exploratory analysis, data quality checks were carried out on all data sets. Automatic evaluations of FastQC quality check modules must be put into perspective in case of atypical library construction or special sequencing conditions and interpreted accordingly.

Though screening for contaminants of known sequence could have been carried out by using the FastQC tool, an impurity screening for non-human organism DNA required a more potent approach. Therefore, the software FastQ Screen was used for screening for common wet lab contaminants. Reference genomes of the following organisms along with the latest patch release of the human reference genome GRCh37.p13 (original release February 2009, patch release June 2013) [72] were downloaded from the University of California, Santa Cruz (UCSC) genome browser file transfer protocol (FTP) server [73]: cat (*Felis catus*), mouse (*Mus musculus*), dog (*Canis familiaris*), *Escherichia coli*, lettuce (*Lactuca sativa*), tomato (*Solanum lycopersicum*). Furthermore, the Coliphage phi-X174 *sensu lato* genome (accession.version: NV_001422.1) was downloaded from the National Center for Biotechnology Information (NCBI) nucleotide database [74].

Index files for downloaded FASTA reference sequences were created using Picard and SAMtools. FastQ Screen was configured to use the BWA aligner to map reads of query data sets to each of the reference genomes.

2.8. Software Development Process

The freeware community edition of the PyCharm Integrated development environment (IDE) [75] was used for creating Python code. For local software development, data sets were subsampled to a custom amount of per mill using a Python script in an automated fashion. Subsampling kept correspondences between mate pairs and UMI sequence files. MiSeq data sets were subsampled to one and ten per mill. NextSeq data sets were subsampled to one per mill only.

Access was granted to the private github [76] account of the D & R Institute of Human Genetics. A repository was created for this thesis using the account. The git [77] (GPLv2 license) version-control system was used for pipeline development. The analysis pipeline code and accessory scripts were regularly uploaded to the repository.

Code was transferred between the local hardware and the server hardware using the open source FileZilla freeware [78]. It is distributed under the GPL license.

Mapping results and manipulated BAM files were visualized manually for consistency checks and means of debugging using the open source integrative genomics viewer (IGV) [79] [80] version 2.4.10 which requires Java version 8. The use of the IGV software is granted under the MIT license.

2.9. Performance Measures

All measures described below collectively form the performance of the analysis approach from sample drawing to variant calling. In this thesis, the focus was placed on the total amount of detected GTVs. This number divided by the total amount of possible GTV calls can be interpreted as a sample-based estimate of the classical (Bayesian) probability of calling a GTV using a certain analysis approach.

A more frequently used measure for variant calling performance assessment is recall. Recall is simply defined as the number of called GTVs (true positives) over the total amount of GTVs in the dataset (true positive calls plus false positives). Hence, it can be viewed as the representation of GTV calling probability as a fraction (*i.e.* value range: 0 to 1). GTV recall will be used subsequently.

The VAF-dependent deviation (precision and trueness) of the called VAF from the target VAF represents another performance measure, simply termed 'VAF accuracy'. This measure is mostly investigated using plots since variability is comprehended best visually.

Two thresholds were assessed: the lowest called VAF in a data set which was used as an estimate for the lowest VAF of a detectable GTV (limited by coverage), and the VAF at which GTV recall was interpolated to be 0.5. The last measure was used to describe the usability of a sample analysis approach. In case of few GTVs, the last measure required the expected GTV VAFs to be identical and, thus, could not be assessed for the TruQ4 reference material. The TruQ4 reference material only covered GTV VAFs from 5% to 30%. The 50% GTV observation VAF threshold was estimated from a linear interpolation using GTV observation counts and target VAF levels of the two data sets achieving more and less than 50% GTV detection.

Lastly, the overall amount of variant calls was assessed to describe the susceptibility to noise. Fewer calls indicate fewer calls due to uncleared base noise especially in data with short amplicons.

2.10. Exploratory Data Analysis

The exploratory data analysis investigated short indel and SNV calling performance of both Broad Institute's Mutect and Qiagen Science's smCounter variant callers. The smCounter variant caller was specifically developed to make use of the UMI-annotation of alignments while the Mutect caller does not regard any kind of UMI annotation. Therefore, either a basic removal of technical DNA template replicates ('deduplication') or a UMI group consensus formation was required for the Mutect analysis to obtain a representative single alignment per UMI. Sequence-based deduplication could be carried out like in conventional NGS analysis (which does not make use of UMIs) via the Picard toolkit. Nevertheless, a consensus formation approach for UMI groups was pursued to achieve noise suppression and to obtain a less distorted picture of the original state of the sequencing library. In contrast, no deduplication was required for the smCounter analysis. Instead, a proper alignment annotation was of utmost importance.

The exploratory data analysis was implemented as a Python module. The `UMIRead` and `UMIReadGroup` classes were implemented in a separate file. The module provides a `UMIanalyst` class implementing all analysis steps. The pipeline is started through the analyst's null parameter method `run_analysis`. The whole module can also be used through a command line interface. An argument parser was implemented to convey attributes to the `UMIanalyst`'s `__init__` method.

Checks were implemented for received user arguments. For certain parameters, advanced checks trying to compensate user errors were implemented. The analysis was shut down in case of violations which could not be circumvented. Meaningful warning and error messages were prompted to the user to correct erroneous arguments more easily.

Paired-end read data line consistency was also checked. Read pairs, where one or more reads failed this check, were omitted.

Besides the main analysis steps, file and folder control methods were implemented to give the analysis module a structured output. All files produced by the analysis code were saved to a destination defined by the user. This folder – the project directory - was named after the analysis sample. Logfiles were saved to a separate folder inside the project directory. Descriptive statistics of read groups and alignment maps were saved to separate text files in the project directory.

References to reads and read groups were deleted along with a globally accessible UMI list of the `UMIRead` class at the end of every analysis.

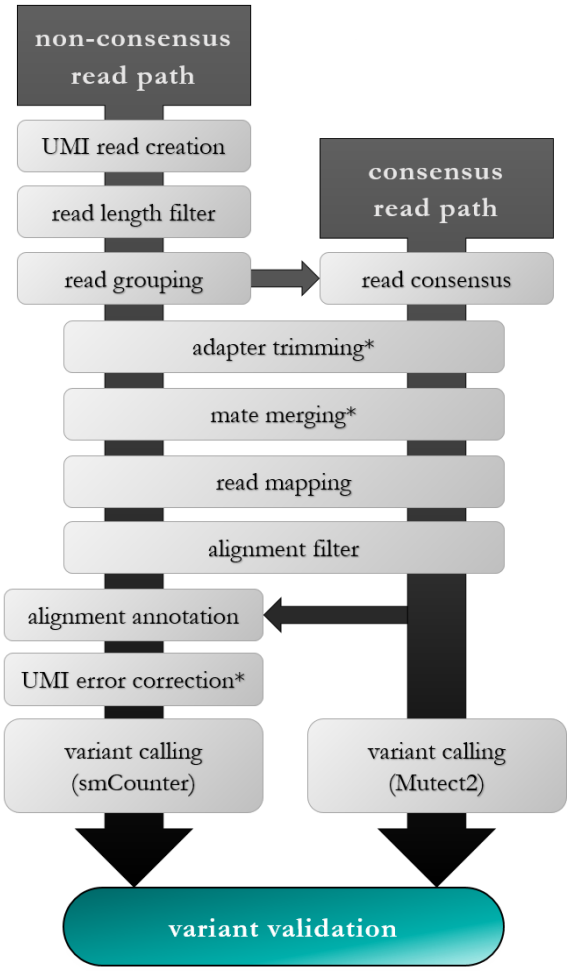


Figure 2.1: structure of the exploratory data analysis. *The turquoise box depicts the last variant validation step which was implemented in a separate script. Steps marked with an asterisk are either optional depending on the analysed data set (adapter trimming) or were disabled during optimization (mate merging). For alignment annotation of the non-consensus read path, information from the consensus read path was used. Figure created with Microsoft Word.*

Structure

The exploratory minimal approach (see figure 2.1) tried to exploit the UMI tags as early as possible so the rest of the bioinformatic analysis could be carried out as usual. For this purpose, `UMIRead` objects were created linking the UMI information to the paired-end read data. Read groups were formed based on UMI sequences afterwards. A multiple sequence alignment of reads forming a group was omitted for simplicity.

Subsequently, a consensus sequence was computed for each read group. Consensus formation was implemented as a majority vote over the observed bases at a certain position within the read group.

After computing the relative portions of each base, the first type of nucleobase exceeding the threshold count in the order of adenine (A), thymine (T), guanine (G), and cytosine (C) was declared as the consensus base type. In case no base type exceeded the threshold count, the base was left undefined using an 'N' symbol in the consensus read sequence.

After read pre-processing, reads were mapped to the human reference genome using the BWA mem algorithm. Alignments were filtered based on their mapping score and the pair orientation (*i.e.* proper pairs).

The implementation of the alignment annotation in the non-consensus read path followed the requirements defined in the supplementary materials of the smCounter paper [36, p. 3]. A sequence of strings was added to the query name of every alignment describing the mapping location of the plus-strand-relative, leftmost mapped base of the entire UMI group of an alignment.

A template for the annotation string is given between quotation marks which are not part of the annotation:

'`:reference_contig_name-strand-position-UMI_after_clustering:UMI_before_clustering`'

Individual strand annotations were extracted from the SAM flag of each alignment. The leftmost mapping position inside the UMI group was estimated by using the mapping position of the corresponding UMI's consensus read alignment.

The UMI-tools software was used to compensate for UMI errors. Four error correction approaches 'percentile', 'cluster', 'adjacency', and 'directional adjacency' were tested on the NEBNext 100% TruQ4 data set. The best performing 'directional adjacency' approach was then applied to all data sets. In short, a UMI network was created for every covered genome position and solved by merging two UMI alignment groups, if their UMIs were within edit distance of one and if the larger group's size was at least twice the size of the smaller group minus one. Detailed statistics about the Hamming distance of observed UMI sequences per position were created along with the UMI reassignment suggestions. These suggestions were used to change alignment annotations. Commands for UMI error correction and variant calling were manually applied to filtered BAM files. Mutect variant calling results of the non-consensus BAM file before and after UMI-error correction with UMI-tools were compared.

Finally, variant calling with specific callers for each analysis path was carried out. Variants of collapsed reads were called with Mutect with default parameters. Using the smCounter variant caller required computation of statistics concerning mean UMI depth and mean read pairs per UMI. The BEDTools' `genomecov` command was utilized for UMI depth computation on the

alignment file of the collapsed analysis path. Mean read pairs per UMI were calculated from the non-consensus alignment file using custom code. Furthermore, a file for tandem repeat regions and a repeat masker subset file had to be provided to the caller which were used for identifying and masking repeat regions. These files can be downloaded from the University of California, Santa Cruz (UCSC) website for each release of the human genome. A BED file containing regions of interest was passed to both variant callers. VCF files were obtained from the variant calling process.

Variant Validation and Performance Assessment

Variant calls of dilution series were validated using the tumour levels in the corresponding sample and the ground truth variant table corresponding to the sequenced reference DNA. Variant ground truth tables were taken from the manufacturers' websites. Stand-alone VCF files were only analysed regarding GTVs.

For dilution series data sets, a variant trace was carried out saving the occurrences (trace depth) and VAFs of a given variant in VCF files of a dilution series. The ambiguous background portion of variant calls was estimated as the number of calls occurring only once throughout a dilution series. These variants were labelled 'untraceable'. Correlation measures and a linear regression were computed. In case a variant occurred only twice, the regression result was labelled 'trivial' (*i.e.* a line defined by two points). The linear regression slope was compared to the expected decrease in VAF (based on the tumour DNA percentage in the sample) given the GTV VAF decrease relative to the last data set that called the same variant. Variants were categorized as either tumour (observed VAF decreases with decreasing tumour level), wildtype (vice versa), or left uncategorized in case the regression slope was too flat (default: 20% slope or lower). Afterwards, the regression slopes of categorized variants were compared to the expected increase or decrease.

Variants were omitted for the final steps of validation, if they were located outside of target regions. This was done to disregard low coverage artefacts at the margins of enriched regions.

For smMIP, the concordance between variant calling results of replicates was assessed using the intersect command of BEDTools.

2.11. Reanalysis Software

The code of the final analysis pipeline is based on the exploratory analysis code. It makes use of the `UMIRead` and `UMIReadGroup` classes as well as the `UMIalignment`, `UMIalignmentGroup`, and `UMIalignmentGroupCluster` classes. Improvements mentioned in subsections 2.12 and 2.13

were implemented among other usability improvements including enhanced parameter customization, standardized message formatting, a central master logfile, more descriptive statistics and plots, exception handling, optional debugging output, improved folder structure, and file handling of intermediate and final results. Clustering was implemented as a single-pass algorithm iterating over clusters only once. In contrast, an algorithm described by Peng *et al.* in 2015 [81, p. 11] clusters more rigorously by using more passes and gradually increasing the allowed UMI sequence difference.

Structure

Individual steps of the reanalysis software roughly follow the exploratory analysis' structure but are called consecutively (figure 2.2). Schematic representations of UMI-enhanced bioinformatic information units used in the analysis are depicted in table 2.8.

Reads were filtered according to their length. Reads below the minimum trimmed read length threshold (default: 45 bases) were omitted. `UMIRead` and `UMIReadGroup` objects were created subsequently. Read groups were primarily used for creating the initial alignment annotation and were deleted afterwards. Following read grouping, optional adapter trimming and subsequent mapping using the BWA mem algorithm [82, p. 1f] was carried out.

Alignment filtering was reimplemented as a two-step filter to increase remaining coverage and to avoid region border artefacts (*i.e.* artefacts resulting from decreasing coverage). Alignments were processed as pairs and categorized as either leftmost or rightmost alignment relative to the chromosome's plus strand. A separate alignment file was created for every filtering category to simplify debugging. In case an alignment failed a primary filter, it was immediately rejected along with its mate. Failing the secondary filter criterion lead to pausing the alignment. In case the second alignment in pair also failed the secondary filter, the whole pair was

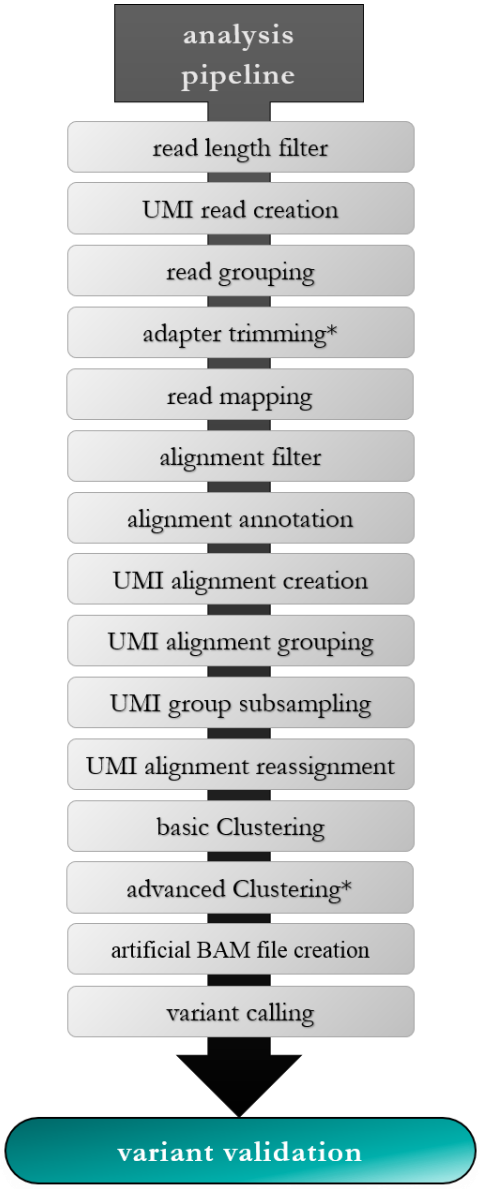


Figure 2.2: consecutive steps of the final variant calling pipeline. *Steps marked with an asterisk are optional. Variant validation was implemented in a separate script. Figure created with Microsoft Word.*

rejected. Primary filtering categories were: alignment not mapped to reference, alignment unpaired or not in a proper pair, is secondary alignment, reads of pair mapped to different chromosomes/contigs. The secondary filter tested alignment pairs for the following cases: alignment outside target regions, alignment exhibits inferior quality. Statistics about chimeric alignment tags observed in primary alignments as well as statistics concerning alignments of a pair mapping to different chromosomes were saved.

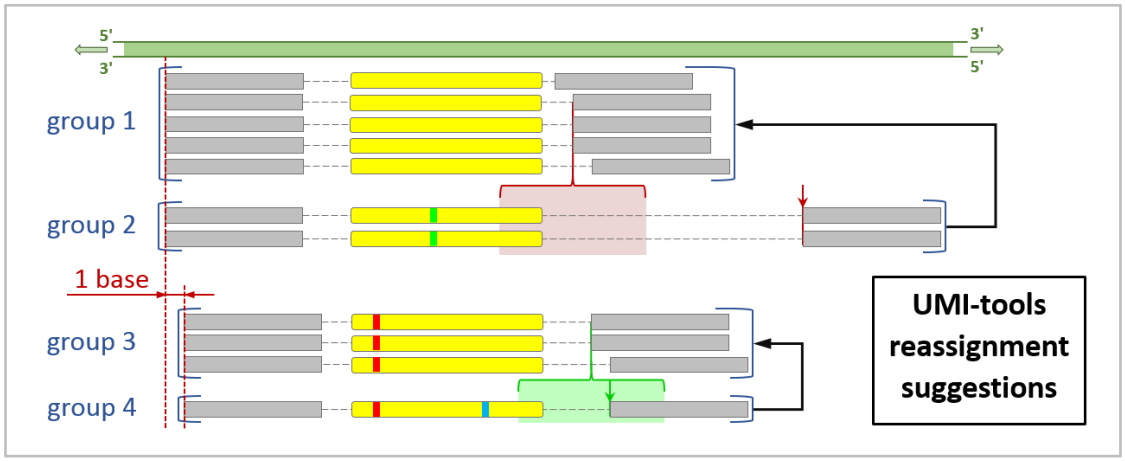
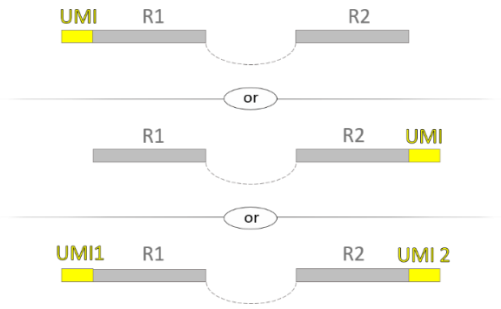
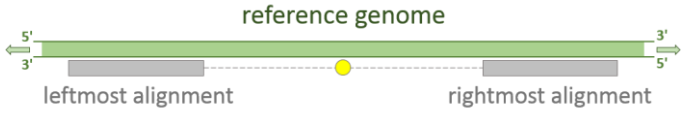
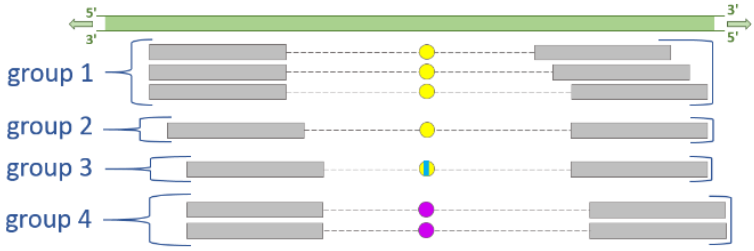
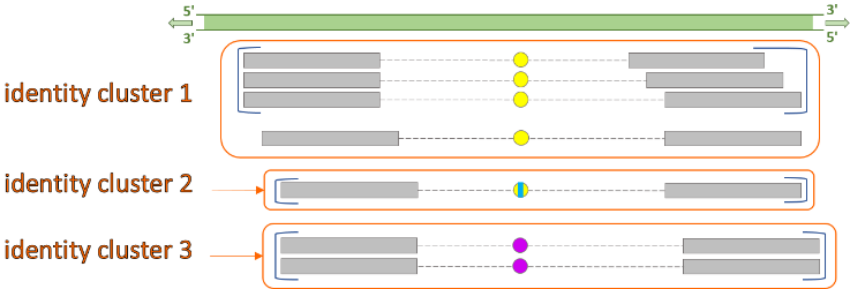
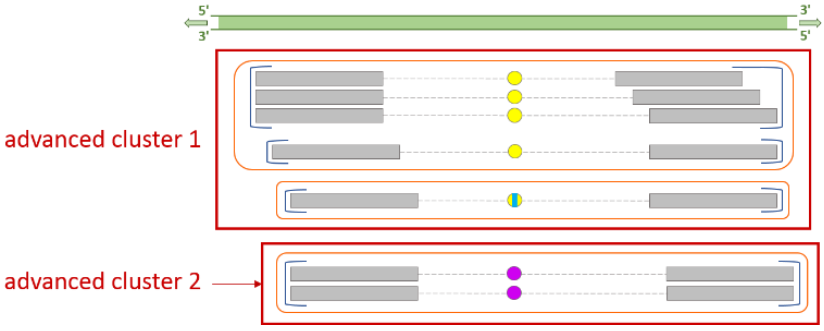


Figure 2.3: alignment group-based UMI correction with UMI-tools. *The UMI sequences are depicted as rounded yellow rectangles between alignments of a pair. Base substitution errors inside the UMI sequence are displayed as coloured vertical lines. Alignment group reassignment suggestions produced by UMI-tools are indicated by black arrows pointing towards the absorbing group. Regrouping suggestions were filtered by mate mapping position (vertical arrows). Red and green areas indicate mate mapping position windows which were calculated from the standard deviations of mate mapping positions of the absorbing group. Red/green area plus the corresponding arrow denote a rejected/accepted reassignment. In the depicted scenario, only the UMI of group 4 would be reassigned. Although group 3 could originate from group 1, UMI-tools disregards this option due to the different leftmost mapping positions.*

`UMIRead` objects were enriched by their corresponding alignment pair's leftmost mapping position taken from the filtered BAM file. Alignments were assigned to their UMI groups according to the UMI information retrieved from the corresponding `UMIRead` objects using a list of object references. Afterwards, references to `UMIalignment` objects were sorted according to their expected sequencing errors in ascending order. The number of expected sequencing errors was calculated from read quality values following equation (2.1). All alignments listed after the user-defined subsampling value (default: 32) were deleted from the group's alignment list. A subsampled alignment file was written to disc for subsequent UMI sequence error correction with UMI-tools (figure 2.3).

$$P_{err} = 10^{-\frac{MAPQ}{10}} \tag{2.1}$$

Table 2.8: information unit schematics for the alignment reducing procedure.

UMI Information Unit	Schematic
read pair	
alignment[1]	
alignment groups[2]	
identity clusters	
advanced clusters	

[1] Extracted UMIs act as labels and are indicated by coloured dots between alignments.

[2] The cyan line within the yellow dot denotes a single base error within the UMI. Group diversity and UMI similarity found in real data is more closely mimicked by groups one, two, and three.

The size relation between absorbing and absorbed alignment group was defined in the UMI-tools paper [68, p. 3]. Groups of size one can also absorb other groups. Mean mate mapping position

windows were calculated from the absorbing group's mean mate mapping position (rounded to the nearest integer). For every reassignment suggestion, the mate mapping position of the alignment in question was checked. The mate mapping position had to be inside the window for accepting a reassignment. A new alignment file with updated UMI fields in the alignment name annotations was written to disc.

Following UMI correction, alignment groups were clustered (figure 2.4). Basic clustering searched for alignment groups with identical UMI sequence with deviating leftmost GTV mapping positions in close vicinity. `UMIalignmentGroupCluster` objects were created which held references to successfully clustered `UMIalignmentGroup` objects.

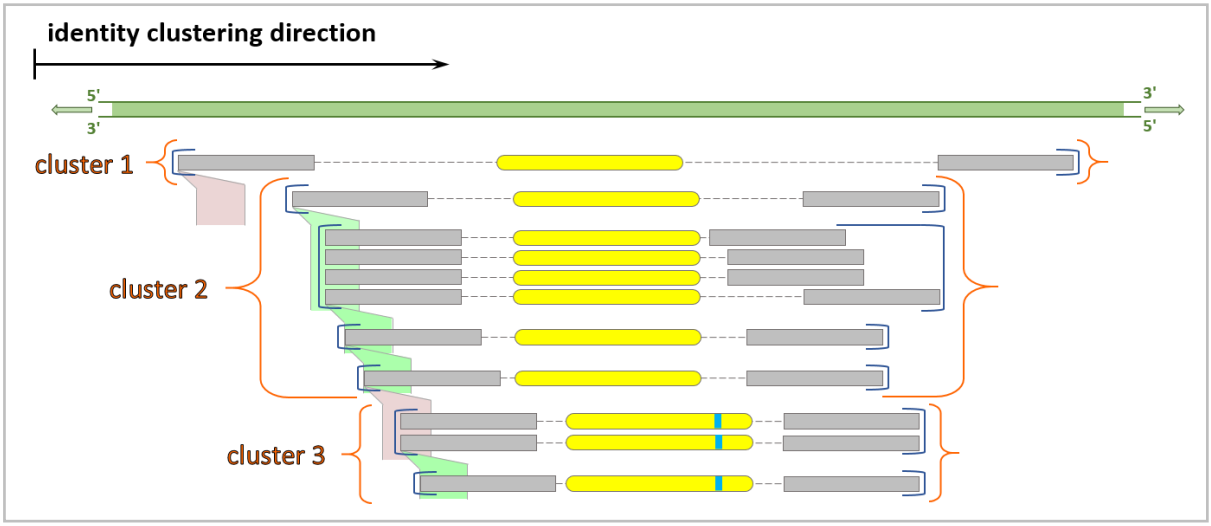


Figure 2.4: clustering procedure for merging alignment groups into identity clusters. *UMI sequences are represented by yellow rounded boxes. Red or green areas underneath leftmost-in-pair alignments indicate identity windows. Using position windows allows for bridging mapping position deviations. Position-sorted alignment groups with a leftmost mapping position inside this window and identical UMI sequences are merged into an identity cluster (green area). If an identity cluster already exists for the seed group, newly found groups are merged to that cluster instead of creating a new one. Red areas indicate that no alignment group was found for merging. After a match is found, the search process for the current seed group is halted at the current position. After all alignment groups mapped to the position are processed, the focus is passed on to the next alignment group. The focus can also be passed to alignment groups that are already inside a cluster.*

The optional advanced clustering used a list of identity clusters which was sorted in descending order regarding the non-subsampled alignment count (figure 2.5). Three specific UMI error categories, which can be disabled or customized by the user, were implemented. The first category was a shift of the UMI sequence towards the 5'-end of the plus strand ('left shift'). The second error category described UMI sequences being shifted towards the 3'-end of the plus strand ('right shift'). Left shifted UMI sequences are expected to occur more frequently than right shifted UMI sequences. Both categories should not be dominant in a standard experiment though. Shift artefacts

are depicted in table 2.9. Finally, clusters were also checked for random PCR errors in their UMI sequences in advanced clustering.

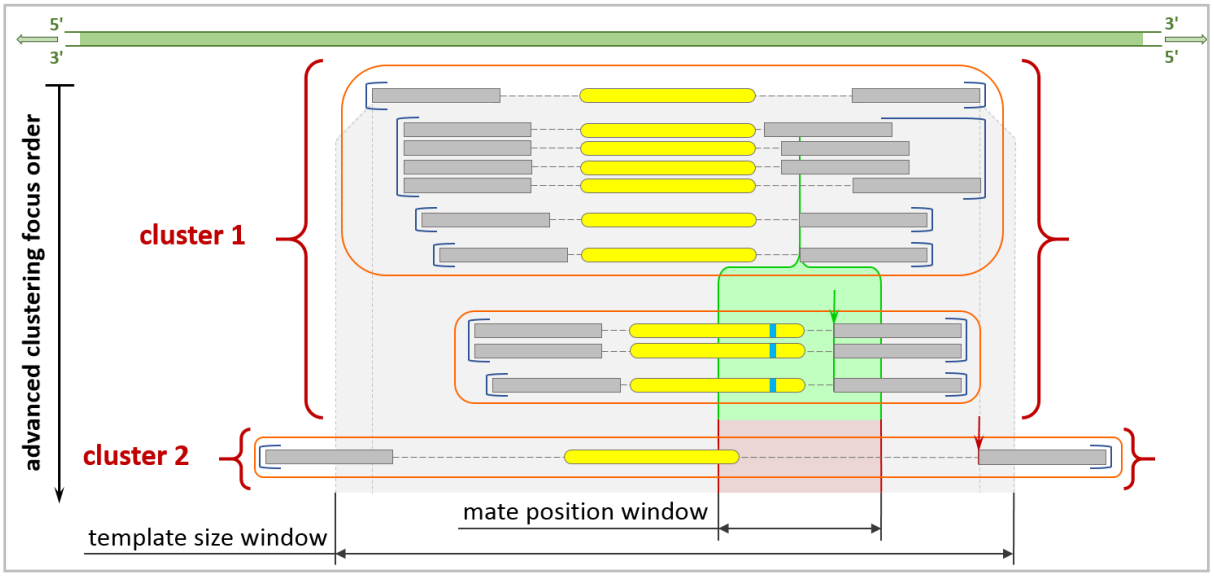


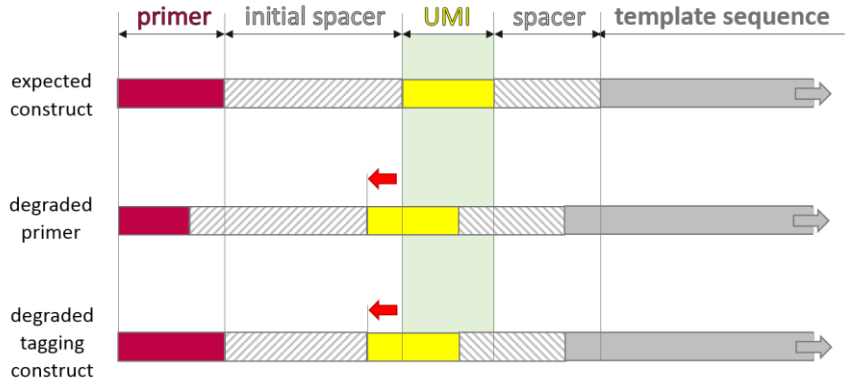
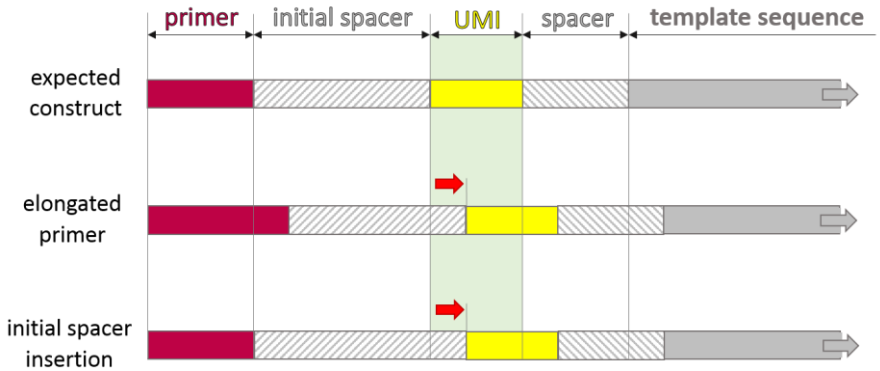
Figure 2.5: optional advanced clustering procedure. *Identity clusters are indicated by rounded orange boxes. The grey background visualizes the template window. Vertical dotted grey lines indicate different window settings. The largest identity cluster is used as the initial focus cluster. All identity clusters inside the template size window smaller than a certain threshold are tested for merging. These clusters are checked for mapping position and template size similarity. For this purpose, a template window is computed. The green area indicates a passed check for mate mapping position while the red area shows a hypothetical case where a mate mapping position check would fail. However, the bottom identity cluster would fail the previous template size window check first. If both checks are passed, the smaller identity cluster is investigated for UMI sequence artefacts. After all merge candidate clusters are processed, the focus is passed on to the next smaller identity cluster until the first cluster of size one is reached. In the shown example, the top and the middle identity clusters would be merged to an advanced cluster correcting the putative substitution error in the absorbed cluster's UMI sequence.*

After final alignment clusters had been determined, clusters were initialized. Initialization produced several descriptive statistics about the cluster, calculated the consensus sequence, and calculated an optimistic estimate of the consensus base quality values. The latter was carried out following the considerations on quality computation for read pair merging based on Bayes' theorem as stated in the USEARCH algorithm paper [83, p. 2f]. The base quality values were limited to a maximum error probability of one (quality value zero) and a minimum error probability of 6 substitution errors in 100,000 bases which corresponds to a quality value of 42. Concerning CIGAR notation (CIGAR stands for 'Compact Idiosyncratic Gapped Alignment Report'), all non-match alignment information was included in the cluster consensus formation primarily following an information maximization principle and an error suppression principle secondly.

Sequence consensus formation was carried out by counting the most frequent nucleobase at a certain position inside the insertion-free base matrix. In case of deletions, the respective positions were padded by Ns. If a deletion was most frequently observed, the deletion was added to the

consensus CIGAR. Consensus insertions were calculated separately. An insertion was included in the consensus sequence, if and only if the exact same insertion sequence was present at the same base position relative to the reference genome in a number of alignments equalling or exceeding the required consensus threshold. If more than one distinct insertion sequence was observed equally often at a certain position, the longest insertion sequence was chosen. In the unlikely event of observing two equally long and equally often observed insertion sequences, the one exhibiting the lowest expected error was chosen. If there would be still more than one candidate insertion sequence, a sequence was chosen at random. In case of initial or terminally soft-clipped sequences, a 'consensus soft-clip' was created. The longest, most frequently observed soft-clip was chosen if any was present. Hard-clipped bases were omitted by default since the BWA mem algorithm only marked hard-clipped bases in the CIGAR string and removed them in the base and quality sequences. The CIGAR string was created according to the consensus soft-clip, consensus bases, consensus deletions and consensus insertion sequences.

Table 2.9: UMI shift artefacts corrected by advanced clustering.

UMI Artefact	Construct Schematic
left shift	
right shift	

Representative values for the remaining eight fields of the SAM format specification [62, p. 2] were either chosen or computed from the information present in the cluster. If specified, a bitmap representation of cluster consensus matrices without insertion sequences and with both types of clipped sequence parts removed were computed.

An annotated or an artificial BAM file was created according to the variant caller selected by the user. In case of an activated 'omit error-prone clusters'-flag, only alignments of clusters surpassing the user-defined 'single error non-ambiguous consensus base'-threshold were written to disc. If the experimental mode was activated, all four versions were output for variant calling with both callers.

Validation of single VCF files and calling results for whole dilution series was carried out by calling the corresponding analysis script.

2.12. Parameter Optimization

Mapping, Mate Merging, and Omitting Error-Prone Clusters

To optimize data pre-processing prior to variant calling with respect to GTV recall, the influence of selecting different quality thresholds and mapping parameters, and the influence of certain optional steps in the analysis procedure were tested.

The effect of mate merging on the outcome of read mapping was tested. The portion of read pairs which could be merged was assessed using the `fastq_mergepairs` command of the USEARCH software [83]. To avoid coverage problems caused by mapping which might reduce GTV recall and/or decrease VAF accuracy, BWA versions 0.6 and 0.7 were tested. The mapping quality threshold of the alignment filter and the minimum allowed trimmed read length was varied to find satisfactory values for excluding alignments that negatively affect performance measures.

The trimmomatic software [84] was used to assess the portion of reads that can be successfully trimmed for adapter contaminants.

Variant calling with Mutect on the NEBNext dilution series was carried out using version 4 and the last release of version 3. Furthermore, Mutect calling was tested on the hundred percent tumour sample of the TruQ4 dilution series NEBNext data set with a variation of the following settings: variant calling regarding and disregarding soft-clipped bases, and optimizations either activated or deactivated.

Clustering

Several settings of identity and advanced clustering parameters were tested to optimize default parameters with respect to GTV recall (summary in table 2.10). The optimization relied on Seraseq data sets only since they had the largest dilution series with the lowest VAFs. Both QIASeq and NEBNext tagging protocol data sets were included. Parameters tested for optimization were: required cluster size difference in advanced clustering, clustering permissiveness (*i.e.* sizes of identity and advanced clustering windows), the option to disable advanced clustering, and the option to remove clusters containing an insufficient amount of alignments for the formation of non-ambiguous consensus bases in case of a single base error. Clusters of size one and two were omitted in the latter case.

Statistics about observed advanced UMI error category cases were recorded along with the actual number of corrected artefacts. An error category prioritization was implemented for the advanced clustering procedure: 5'-shift first, then 3'-shift, and lastly random PCR errors. This artefact prioritization order was chosen to have an optimistic estimate of the shift error occurrences since it was unclear whether these kind of UMI alterations occur at all.

Table 2.10: clustering settings and associated parameter values.

Clustering Setting	Identity Clustering Window Size	Mate Mapping Position Window Size[1]	Advanced Clustering Window Size	Advanced Error Extent Corrected[2]	Shift Artefacts Corrected
moderate	3 bases	1SD	0 bases	1	no
permissive	5 bases	3SD	3 bases	2	yes

[1] Parameters other than identity clustering window size and mate mapping position window take effect in advanced clustering.

[2] The mate mapping position window extends in both 5' and 3' directions and defaults to 10 bases in each direction, if the focus cluster's mate mapping position standard deviation (SD) is below 5.

[3] The advanced error extent defines the number of allowed UMI base differences and bases a UMI may be shifted and will still be merged with the focus cluster.

Variant Caller

For smCounter, the automated detection threshold selection was deactivated to use a more sensitive setting which allows for a single alternative base observation leading to a variant call. This setting was used after initial calling tests on NEBNext resulted in a much lower GTV recall than was achieved by the most sensitive setting. This also allows for a more direct investigation of the effects of clustering efforts on noise suppression.

The decision on the best suited variant caller for low VAF variant calling was made based on the results of the Seraseq dilution series analyses, which were computed with the reanalysis pipeline. Performance measures were prioritized in the following order from most important to least

important: GTV recall, VAF accuracy, lowest callable VAF, susceptibility to noise, and 0.5 recall VAF threshold.

2.13. Computational Optimization

The use of the C-based pysam module was a countermeasure to improve alignment file interfacing performance in cases where commands of external tools were not suited for a desired task. Custom code was required for successfully applying filters which used the UMI tag and the mean mate mapping position as additional criteria. In case of the read length filter, the `islice` command of Python itertools module was used to exploit the regular structure of FASTQ files.

To speed up clustering, a decision was made to create statistical plots only in cases where a cluster reached a certain alignment or alignment group threshold. Consensus matrices were only output as bitmap pictures, if there were more than 19 alignments or more than 4 alignment groups in a cluster. Thresholds were chosen rather permissive and combined with an upper threshold for the number of plots that were allowed to be created for every reference contig.

Alignment groups were subsampled to a maximum of 32 alignments per default to reduce the runtime of UMI-tools and to limit group absorption to smaller alignment groups which was reasoned to keep the impact on the original allele distribution to a minimum.

A fast Python solution was implemented to compute consensus representations of alignment group clusters. The implicit element-by-element nature (*i.e.* 'broadcasting') of operations of the numpy module along with vectorization were utilized by using the `array` data structure and the vectorization function `vectorize.` This function allows to translate the standard Python code of the per-position consensus bases computation into code that uses numpy broadcasting. Since it was unclear whether a cluster will end up containing only one alignment or several thousand, the consensus computation had to be extensively optimized.

Main memory usage was reduced by enforcing full garbage collection during pipeline execution in situations where objects occupying loads of memory had just been deleted. In addition, multiple analyses were called successively through a separate Python process through a driver script to avoid any kind of incomplete garbage collection which might have unnecessarily increased memory usage over time.

Alignments were extensively filtered prior to alignment object creation. Only alignments mapping to target regions and matching further criteria were kept. To reduce the impact of alignment object instantiation on main memory, read objects were deleted after the alignment annotation step. To

avoid a state of the analysis pipeline where read objects and alignment objects coexist in main memory, the alignment enrichment procedure was carried out without the use of alignment objects. To avoid data redundancies, alignment group and alignment group cluster objects only referred to alignment instances rather than copying their contained information.

In general, when operating on huge data structures, an emphasis was given to reduce the steps required for carrying out a task. List comprehensions and generator expressions were used for faster iteration whenever possible. The creation of large objects for iteration like lists or dictionaries was circumvented by using generator expressions and iterators over existing iterable objects.

2.14. Noise, False Positive Estimation and Correction

To establish a minimalistic model for polymerase errors during DNA synthesis, the error base incorporation characteristic was approximated as white noise. This very simple model was used to describe how noise affects GTV calls with low VAF. An error occurring at a GTV position in a DNA fragment carrying the reference allele can lead to one of the following situations: the error either supports the GTV allele or a non-GTV allele. In the first case, the error increases the GTV VAF. In the second case, it supports one of two non-GTV alternative alleles equally likely because of the adopted noise characteristic. This means that observing any non-GTV alternative allele is twice as likely as observing the expected alternative allele in case of a polymerase error.

Table 2.11: possible combinations for alternative allele observations in the event of two base errors at GTV-supporting positions.

Non-reference Allele Observations		**Cases Supporting Outcome Type**	**Variant Called (assumed)**
Base 1	Base 2		
GTV	GTV	1	Yes
Non-GTV A	Non-GTV A	2	Yes
Non-GTV B	Non-GTV B		
GTV	Non-GTV A	6	No
GTV	Non-GTV B		
Non-GTV A	GTV		
Non-GTV B	Non-GTV A		
Non-GTV B	GTV		
Non-GTV A	Non-GTV B		

In case an error occurs in a template which supported the expected GTV allele, the similar considerations can be made. According to the model, the incorporation of a base supporting a

non-GTV allele would be twice as likely as the incorporation of the reference allele-supporting base. If the GTV allele frequency is small (≤1%), the portion of errors resulting from these GTV allele supporting reads can be omitted as a first approximation. Hence, the level of molecular noise could be determined solely by the number of non-GTV alternative allele observations. The practical signal level would then be the observed VAF of the GTV allele minus half the number of observations supporting other alternative alleles.

Considerations like the above can be made for variant calls which are supported by two alternative allele observations under the assumption that they resulted from uncleared polymerase errors (see table 2.11). Again, observing any non-GTV, non-reference allele is twice as likely as observing the GTV allele. Under the assumption that variants are only called if the same nucleobase was observed twice, the portion of base error combinations resulting in calls is only one third of the portion for single alternative allele observation calls. Thus, dual observation-supported variant calls due to polymerase errors are expected to be made less frequent.

A simple noise level classification for false positive estimation was applied to VCF data using to two binary criteria. These criteria consider the relation between GTV signal level (VAF) and coverage dictated noise level. The first criterion is met if the noise level is approximately equal to (≥90%) or higher than the signal level (SNLR criterion). The second criterion is met if the noise level is higher than half the signal level (HSNLR criterion). The signal level was defined as the data set's
expected (theoretical) GTV VAF. The coverage-dictated noise level was defined as the lowest observed VAF of any variant call in the corresponding VCF file.

A trade-off based on observed variant calls was made to reduce the number of GTV calls by a reasonable amount but still allow for GTV calls in situations where expected variant signal and coverage dictated (estimated) noise levels are comparable (*i.e.* SNLR criterion met). The number of true positive ground truth variant calls of a VCF file (part of a dilution series), was reduced based on:

- the number of alternative calls at GTV sites in the control data set of the dilution series
- the estimated noise level of a data set belonging to the same dilution series (SNLR and HSNLR criteria)
- the amount of UMIs supporting a GTV call of the given non-control data set

In case the SNLR and the HSNLR criteria were not met (*i.e.* noise level below half signal level), simply all single and double observation supported GTV calls were regarded as false positives due to the large gap between signal and estimated noise. If the HSNLR criterion was met but the SNLR

criterion not (*i.e.* noise level between 50% and 90% of signal level), all single UMI-supported GTV calls were regarded as false positives along with a portion of the dually supported GTV calls. This portion was estimated to be half the number of dual UMI-supported alternative calls at GTV sites in the control data set. This was observed to be a good estimate for false positive GTV calls in both control data sets of NEBNext and QIASeq Seraseq analyses which did not make use of advanced clustering and, thus, were thought to have resulted in more noisy calls. The number of removed dual UMI-supported GTV calls was limited to the total amount of dual UMI-supported GTV calls in a given data set. For example, if the noise level in a given data set was 65% of the expected signal level and there were twelve dual UMI-supported GTV calls in said data set, and if there were ten alternative calls at GTV sites in the control data set, the number of dually supported GTVs to be removed would be five, leaving seven dually-supported GTV calls as true positives in the data set.

If the noise level is approximately or above the signal level, instead of removing all single UMI-supported GTV calls, the procedure for estimating the false positive portion of dually supported GTV calls was also applied to single UMI-supported GTV calls. The number of false positive single supported GTV calls was estimated from half of the number of single UMI-supported alternative calls at GTV sites in the control data set. The portions of false positive single and dually UMI-supported GTV calls are subtracted from the data set's number of single and dually supported GTV calls.

Although this procedure gives a better estimate of the actual number of alternative allele-carrying templates captured by the tagging approach, it cannot be used to select the true positive GTV calls from the pool of single and dually-supported GTV calls of a data set. Also, this procedure fails, if the number of alternative calls at GTV sites in the control data set is rendered non-informative due to successful error suppression.

To regard rare false positive GTV observations in the control data analysis which do not stem from the single and dual observation portion of variant calls, a supplementary approach was added. In addition to the false positive estimation based on the signal-to-noise level relation, if a non-SOC, non-DOC GTV was observed in the control data set, the exact same variant is removed as systematic false positive call from any non-control data set in which it occurs at similar VAF. The maximum VAF deviation allowed for a GTV to be counted as systematic false positive was defined as 25% to allow for UMI coverage variations between data sets.

The combined GTV recall correction measures described above are named 'estimated false positives filter' subsequently.

3. Results

3.1. Quality Control

A minor quality degradation towards the 3'-end of reads was frequently observed for all data sets. Second in pair reads exhibited lower overall quality compared to first in pair reads. Notable abnormal results are presented below (see figure 3.1).

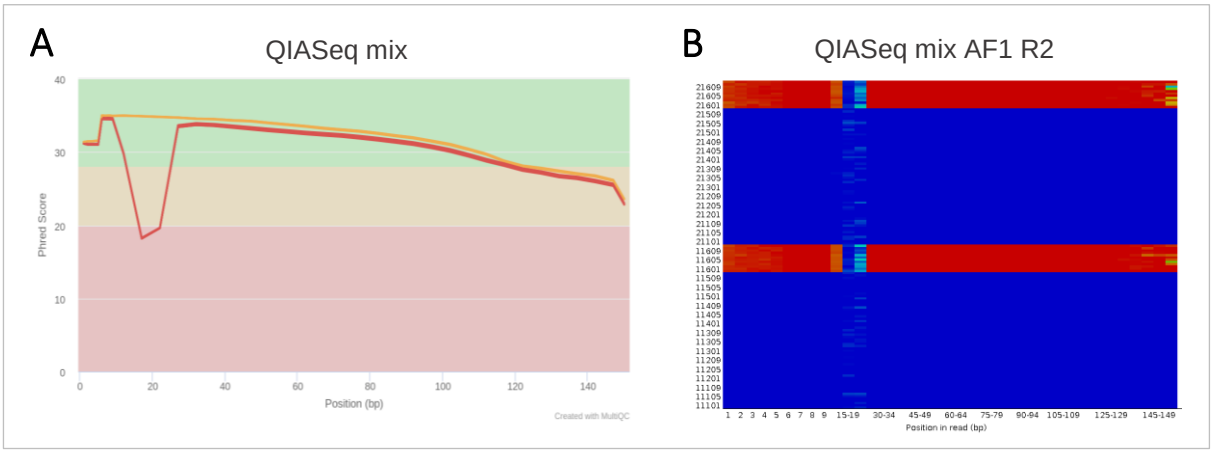


Figure 3.1: abnormal quality check result for QIASeq Seraseq dilution series. *(**A**) Mean quality scores per position of QIASeq mix data sets. A warning was output for every QIASeq R1 FASTQ file of the Seraseq dilution series (orange lines) because read quality degraded towards the 3'-end below a quality value of 30. The R2 FASTQ files failed the quality check (red lines) because of the steep decline in quality after the 9th base to a mean quality value to 18. The quality degradation was very similar for all QIASeq Seraseq samples because they were sequenced simultaneously. The plot was created with MultiQC. (**B**) Per tile quality score deviation from the average quality score per base position of one R2 FASTQ file of the QIASeq Seraseq dilution series. Blue colours denote tiles with mean quality scores equal to or higher than the average score. Red colours depict lower mean quality scores. The plot was created with FastQC.*

Per sequence GC-content distributions for different tagging and targeting methods are shown in figures 3.2, and 3.3. The spread of per sequence GC-content distributions was shortest for smMIP (30-40%), followed by QIASeq (around 40%). ThruPLEX showed an intermediate GC-content coverage of about 50%. NEBNext exhibited the largest GC-content spread of approximately 55%.

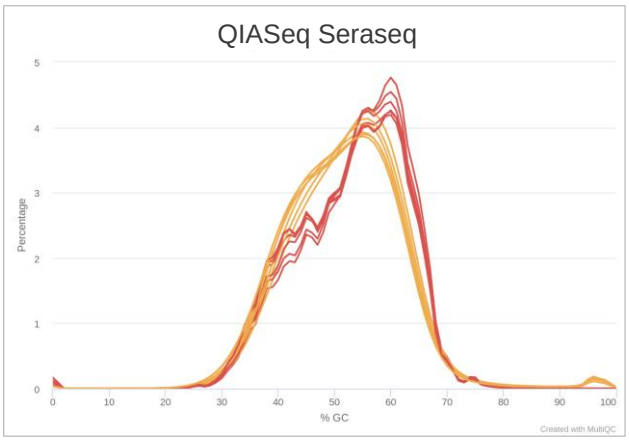


Figure 3.2: per sequence GC-content of QIASeq Seraseq data sets*. All QIASeq Seraseq R2 data (orange curves) exhibit a more Gaussian-like, smoothed version of red R1 distributions. QIASeq distributions exhibit a negative skew. The plot was created with MultiQC.*

As expected, sequence duplication statistics showed alternating base composition per position and non-uniform amplification of a multitude of sequences (and k-mers). The estimated sequence duplication levels for

NEBNext read pairs (figure 3.4) differed from the estimated duplication levels for corresponding UMI sequences which were available as separate FASTQ files. Usually, UMIs showed a three to five percent higher duplication level compared to read sequences of the same sample.

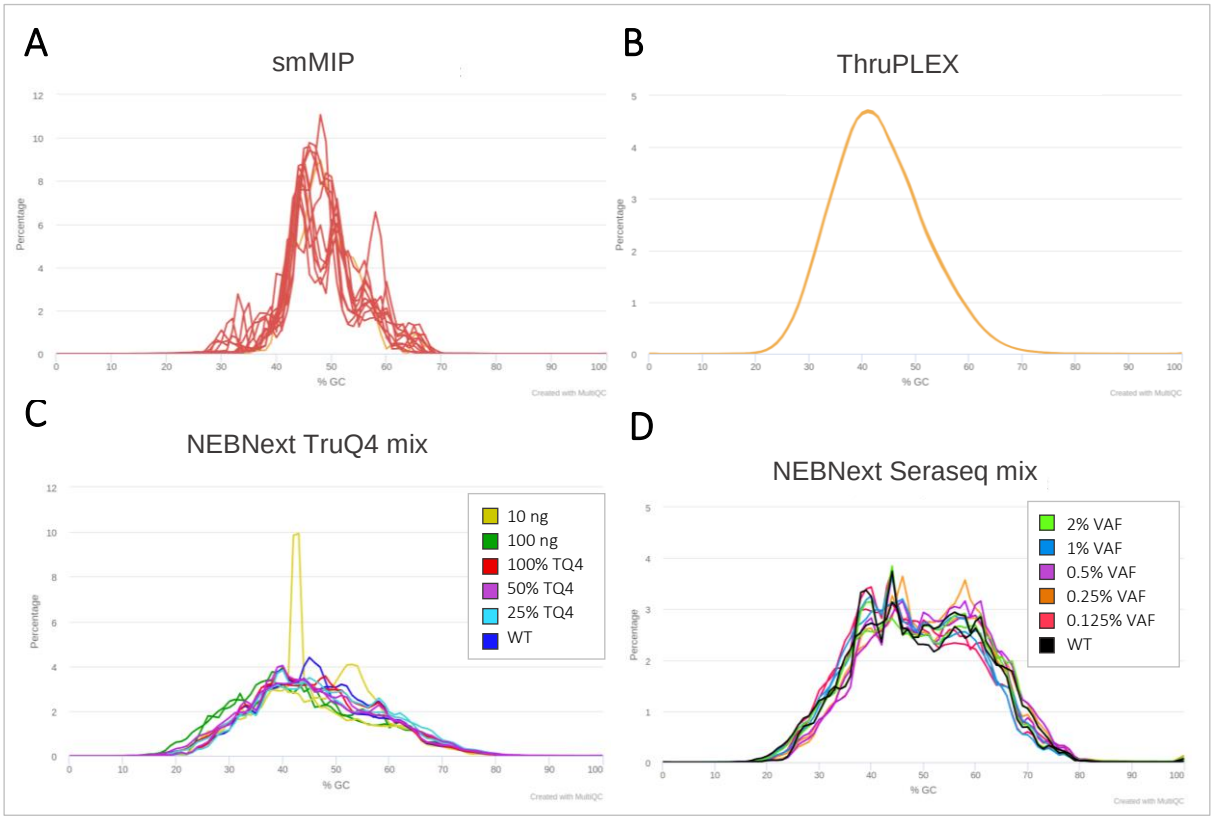


Figure 3.3: per sequence GC-content distributions. *(**A**) smMIP data sets: no normal distribution was observed for any of the smMIP data sets and, thus, the red colouring due to the failed quality check. (**B**) ThruPLEX data sets: all distributions of ThruPLEX data are almost identical, smooth and exhibit a mode at 41% GC. Nevertheless, a warning was output by the MultiQC software (orange colour) because the distributions were shifted relative to a reference normal distribution which was calculated from the data set. (**C**) NEBNext TruQ4 data sets: the yellow distribution of the NEBNext 10 ng input material data set shows a sharp peak around 43% GC and an additional mode at 53% GC. Distributions for NEBNext data sets using the TruQ4 reference DNA are slightly positively skewed. All data sets failed the quality check module. Curves with identical colour represent data from R1 and corresponding R2 FASTQ files. (**D**) NEBNext Seraseq data sets: distributions for NEBNext Seraseq data sets are more symmetrical than the GC-content distributions of NEBNext TruQ4 data sets. Deviations from a normal distribution due to region targeting and template amplification are also clearly visible. All data sets failed the quality check module. Plots were created with MultiQC.*

Notable contamination with Illumina universal adapter sequence was found in the NEBNext 10 ng data set (~5%) and the NEBNext 100%TruQ4 mix data set (~2.5%). Traces (≤ 1%) of the same sequence were found in all other NEBNext and both ThruPLEX data sets. Traces of the nextera transposase sequence were found in the QIASeq TruQ4 data set and all non-control data sets of the QIASeq Seraseq dilution series.

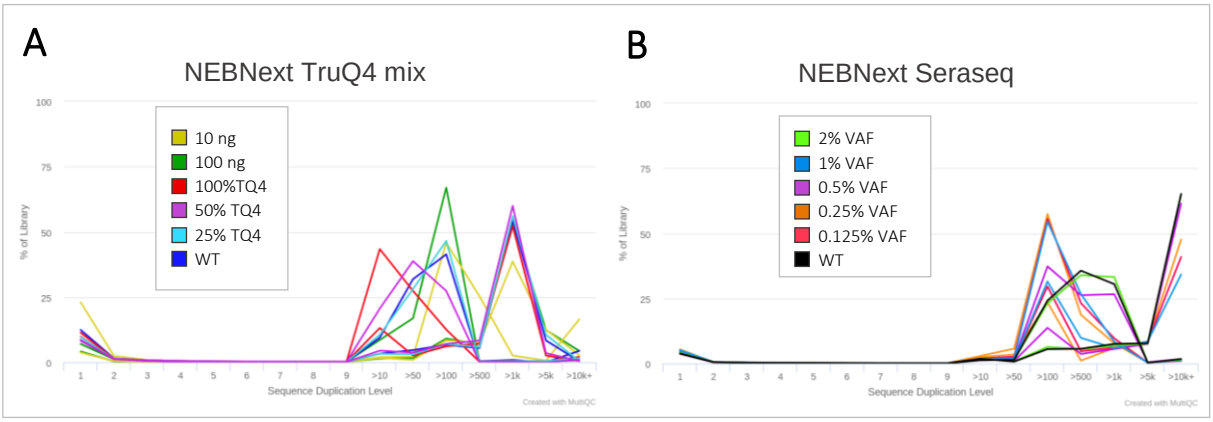


Figure 3.4: estimated duplication levels for NEBNext data sets of TruQ4 and Seraseq dilution series. *Curves with identical colour represent data from R1 and corresponding R2 FASTQ files. For both dilution series, R2 reads were estimated to have higher sequence duplication levels than R1 reads. The initial portion of supposedly unique sequences stemmed from both first and second-in-pair reads. (**A**) NEBNext TruQ4 mix data sets: duplication estimates were more homogenous for R2 data. (**B**) NEBNext Seraseq mix data sets: the initial portion of unamplified sequences was lower than for TruQ4 mix data sets. There were more highly duplicated sequences than for the TruQ4 dilution series. Plots were created with MultiQC.*

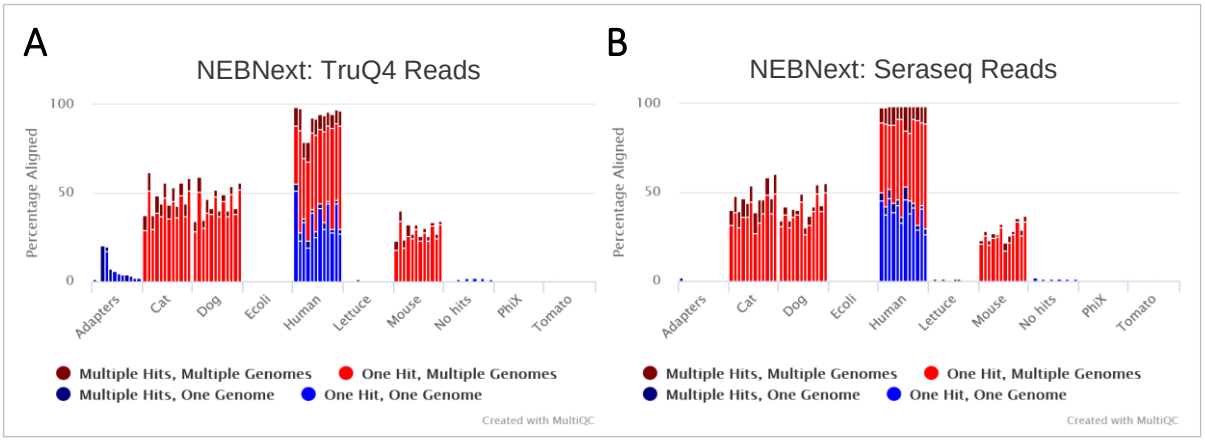


Figure 3.5: impurity screening results for NEBNext reads of both TruQ4 and Seraseq reference material. *Red portions of reads shared between mammalian genomes indicate homologous captured regions and do not indicate contamination. The few reads aligning against lettuce in both cases were either single or multiple hits on multiple genomes and were therefore not viewed as true impurities. (**A**) Sample order (left to right) is: 100 ng, 10 ng, WT, 100% TruQ4, 25% TruQ4, and 50% TruQ4. Mate order is R1 reads before R2 reads. The portion of reads mapping to the human genome varied between 78.9% (10 ng data set) and 98.7% (100 ng data set). The number of R2 reads without any hits were systematically higher than the number for R1 reads but did not exceed 2% of all reads of the individual data set. Notable Illumina universal adapter contaminations are clearly visible. (**B**) Sample order (left to right) is: 100 ng, 10 ng, WT, 100% TruQ4, 25% TruQ4, and 50% TruQ4. Mate order is the same as for TruQ4 results. Adapter contamination did not exceed 2%. The portion of reads mapping to the human genome varied only between 97.8% and 98.9%. Plots were created with MultiQC.*

3.2. Impurity Screening

Impurity screening results of ThruPLEX reads, NEBNext Seraseq reads (figure 3.5, B), and QIASeq TruQ4 reads showed adapter contamination and reads without hits on any of the screened genomes below 2% of all reads. NEBNext TruQ4 reads showed higher adapter contamination

with Illumina universal adapter than Seraseq data sets. This resulted in a lower yield of reads aligning against the human genome (see figure 3.5, A).

A small subset of reads of the NEBNext TruQ4 100 ng sample which showed no hits on any screened genome were used in a web basic Local Alignment Search Tool (BLAST) search on the NCBI website. Searches with the BLASTN program [85] version 2.8.1 were optimized for different sequence similarities (megablast, discontiguous megablast, and blastn). The best search result consisted of 100 hits on artificial sequences with the top 40 hits having scores between 121 and 135. Query coverages ranged from 65% to 100%. Among hits were mammalian expression vectors of type pCI-NEO-FLAG-SMNdelta, and cloning vectors of types L1R1-neo-Scal, 1446_neo, pCI-neo, and GloSponge. Other cloning and expression vectors were also found but exhibited lower quality values and query coverages.

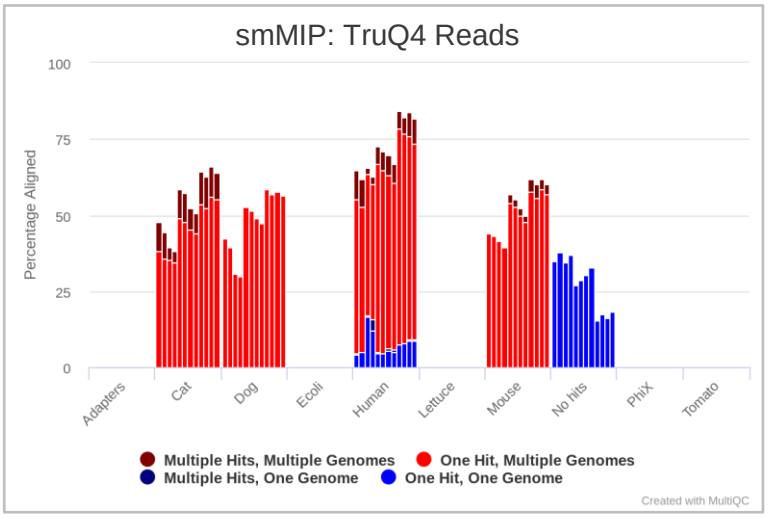


Figure 3.6: impurity screening results for smMIP data sets. *Sample order from left to right is: 100% TruQ4, 25% TruQ4, and 50% TruQ4. Replicate one before replicate two. Mate order inside replicate is read 1 before read 2. Between 15.6% and 37.7% of each FASTQ file did not map to any genome. The plot was created with MultiQC.*

Between 4.2% and 5.9% of R2 data of each QIASeq Seraseq data sets did not align against any screened genome. R1 data aligned better with 1.8% up to 3.1% not aligning to any genome. The fraction of reads mapping to the human genome varied between 96.9% and 98.2% for R1 data and between 93.9% and 95.7% for R2 data.

Screening results for smMIP (see figure 3.6) showed unusually high percentages of unmapped reads per data set which reached almost 40% in some cases. Web BLAST searches using a small subset of unmapped smMIP reads against the human genome and non-human nucleotide databases with the BLASTN program yielded hits with low scores between 35 and 75. Query coverages were also low (between 20% and 27%). Organism results ranged from bacteria and fungi to fish.

3.3. Data Processing Optimization

Read Pre-Processing

Merging read pairs resulted in a decrease of mapped reads for all data sets. The overall number of mapped reads also decreased which ultimately resulted in a disadvantageous higher detection limit in variant calling (see table 3.1). One reason for this was the amount for mergeable read pairs, especially in case of NEBNext for which only approximately 10% to 20% of all read pairs could be merged. Approximate percentages of mergeable read pairs for other data sets were 90% for QIASeq, 85% for smMIP, and 75% for ThruPLEX. Using a combination of merged reads and paired reads in a single analysis was avoided for simplicity. Based on characteristic amplicon structures, the mate merging was expected to work best for smMIP data sets.

Illumina Adaptor Trimming and Mapping

For QIASeq and ThruPLEX MiSeq data sets, the attempt to trim adapter sequences from reads resulted in no trimmed bases. For the NEBNext 100% TruQ4 data set, 2.71% of R1 bases and 0.15% of R2 bases were trimmed.

Table 3.1: influence of mate merging with FLASH on mapping.

Sample[1]	Read Information	Mapped Reads Count	Mapped Reads Percentage[2]
NEBNext 25% TruQ4	pairs	5,363,350	95.82
	merged	232,462	54.14
smMIP 25% TruQ4 replicate 2	pairs	1,980,374	77.63
	merged	41,551	31.50
QIASeq	pairs	3,690,037	99.42
	merged	104,532	83.97
ThruPLEX MiSeq	pairs	2,908,475	99.88
	merged	1,709,115	98.48
ThruPLEX NextSeq	pairs	27,572,258	99.84
	merged	5,158,400	97.61

[1] Only samples yielding the maximum number of variant calls per tagging procedure are shown.
[2] For merged reads: mapped reads percentage calculated from portion of successfully merged read pairs.

The choice of the best suited BWA version was made based on the capability to use hard-clipping to remove substantial terminal portions of reads which did not align against the reference. Hard-clipping also renders short read trimming by potential adaptor contaminants obsolete (assumption: short adapter portion in reads). No notable differences in GTV recall could be reported after

comparing BWA version 0.7 to the latest release of version 0.6. Thus, the newer version 0.7 of BWA was used to map reads.

The mapping quality threshold of the alignment filter was lowered from 17 to 15 and the minimum allowed trimmed read length was lowered from 30 to 25 bases. Both measures resulted in a mild increase of alignments after applying the filter. No effect on GTV recall or GTV VAF variability was observed though. The lower values were used for both thresholds in subsequent analyses involving TruQ4 data sets. For Seraseq data sets, a minimum read length of 45 and a minimum mapping quality of 17 was used.

Mutect Variant Calling Parameters

There was no difference of GTV recall observed when comparing variant calling results for the latest releases of the Mutect-containing GATK versions 3 and 4. The amount of alternative calls at sites of uncalled GTVs decreased from two to zero. This reduction was an improvement because issuing variant calls other than expected GTV calls indicated noise levels at the GTV site being higher than the signal level (*i.e.* higher than the frequency of alternative allele observations supporting the GTV). Based on these results, the latest release available of GATK version 4 was used for Mutect variant calling.

Omitting soft-clipped bases for Mutect variant calling resulted in fewer variant calls compared to the results obtained using default settings (for details see table 3.2). Insertion calls decreased by 62.5% and deletion calls decreased by 31.8%. Thus, the option to disregard soft-clipped bases was not used.

Table 3.2: effect of disregarding soft-clipped bases on variant calling outcome for Mutect version 4.

Activated Mutect Options	SNVs	Insertions	Deletions	Total Calls
Default	271	8	22	301
Disregard soft-clipped bases	254	3	15	272

The option to deactivate optimizations had no effect on GTV recall. Effects on GTV VAF variability were minimal. Therefore, the option to disable optimizations was also not used in subsequent variant calling experiments.

3.4. Exploratory Analysis

Profiling

A listing of execution times per data set measured by the Python driver script is shown in table 3.3. All analyses were started with a maximal usage of 8 threads. The ThruPLEX MiSeq data set showed the longest and the smMIP data sets the shortest execution durations (total and per mega base).

The total execution time of the exploratory analysis can be divided according to the following main tasks: read pre-processing, mapping, alignment processing, and variant calling. The total execution time was approximately distributed over these tasks as follows: 20% read pre-processing, 13% mapping, 20% alignment processing, and 47% variant calling. Times were computed from time stamps of the console output of conducted exploratory analyses. Because both variant calling paths were executed per default, the variant calling task could be further divided according to the utilized variant caller. The smCounter variant calling made up 30% of total execution time while the Mutect caller required 17%. By using a BED file for region filtering, computation time required for variant calling with the smCounter software was drastically reduced.

Read pre-processing, alignment processing, and variant calling using smCounter made up about 70% of the total execution time. These tasks were carried out mainly using Python code.

Table 3.3: total execution times of exploratory analyses without UMI-tools error correction.

Data Set[1]	Read Pairs	Max. Read Length	Total Duration[2]	Duration per Read Mega Base
QIASeq TruQ4	1,808,103	151 bp	1 h 32 m 33 s	10.17 s
NEBNext 10 ng	3,976,440	75 bp	1 h 03 m 24 s	6.38 s
NEBNext 100 ng	2,886,909	75 bp	53 m 00 s	7.34 s
NEBNext 100% TruQ4 mix	2,271,218	76 bp	1 h 01 m 19 s	10.66 s
NEBNext 50% TruQ4 mix	2,182,469	76 bp	58 m 54 s	10.49 s
NEBNext 25% TruQ4 mix	2,798,019	76 bp	1 h 12 m 19 s	10.20 s
*NEBNext 0% TruQ4 mix	2,731,760	76 bp	23 m 08 s	3.33 s
ThruPLEX MiSeq	1,451,625	151 bp	3 h 59 m 39 s	32.80 s
ThruPLEX NextSeq	13,762,371	151 bp	20 h 40 m 47 s	17.91 s
smMIP 100% TruQ4 mix rep1	788,977	151 bp	18 m 52 s	4.75 s
smMIP 50% TruQ4 mix rep1	767,418	151 bp	18 m 32 s	4.80 s
smMIP 25% TruQ4 mix rep1	810,676	151 bp	18 m 59 s	4.65 s
smMIP 100% TruQ4 mix rep2	1,097,521	151 bp	25 m 53 s	4.69 s
smMIP 50% TruQ4 mix rep2	732,203	151 bp	19 m 02 s	5.16 s
smMIP 25% TruQ4 mix rep2	1,460,267	151 bp	29 m 36 s	4.03 s

[1] Terminated analyses are marked with an asterisk.

[2] Total durations were measured by a Python driver script (server loading not monitored).

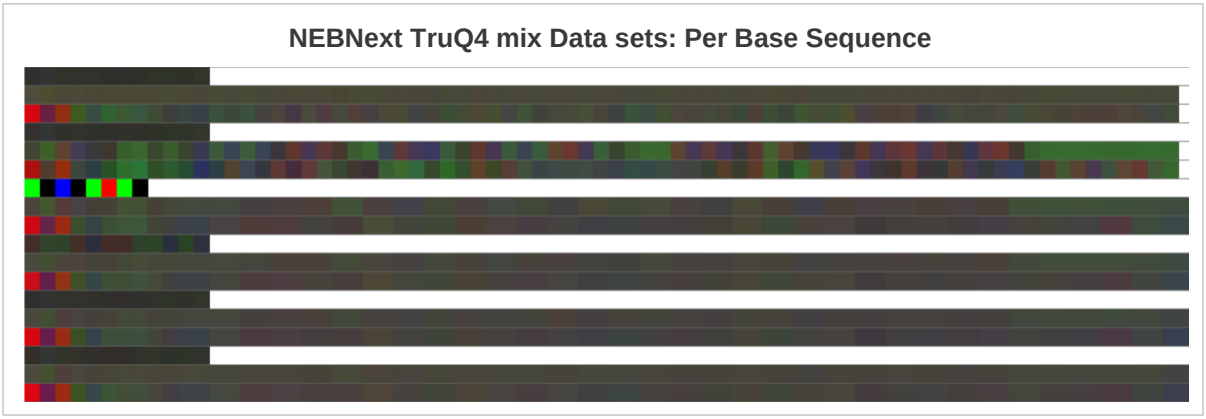


Figure 3.7: per base sequence content of the NEBNext TruQ4 mix data sets. *The base colouring scheme was: G=black, A=green, T=red, C=blue. The resulting colour of every tile in the plot is a weighted additive mixture of all four base colours according to their occurrence in the data set at the corresponding position. Sample order from top to bottom is: 100 ng, 10 ng, WT control, 100% TruQ4, 25% TruQ4, and 50% TruQ4. Three successive lines belong to one sample and are displayed in the following order (top to bottom): UMI sequences, first-in-pair read sequences, and second-in-pair read sequences. The UMI sequences of the WT control were extremely homogeneous. The colour code reveals the sequence: AGCGATAG. The plot was created with MultiQC.*

The NEBNext control data set terminated prematurely. There were only 25 different UMIs for this sample with 24 UMIs occurring only once (see figure 3.7). All other read pairs were associated with the same UMI sequence. This sequence was identical to the Illumina TruSeq CD Index 1 (i7) adapter sequence D712 [86, p. 19]. The missing UMI sequence diversity caused the exploratory analysis of the whole sample to fail because nearly all reads were collapsed into a single sequence. After mapping, the remaining 25 consensus read alignments did not pass the alignment filter which resulted in analysis termination.

An attempt to use UMI error correction on data sets which were not part of this thesis by applying the directional-adjacency approach of UMI-tools failed due to an extremely high run time compared to data sets used in the exploratory analysis. UMI sequence error correction was found to take at least two days for a deep sequenced amplicon data set. The execution of the program was terminated manually after 48 hours.

Amplicon structure

The smMIP data sets exhibited the shortest amplicons which were approximately only one read long (*i.e.* 150 bases). Each region was targeted by a forward and a reverse primer. The smMIP coverage track can be interpreted as the addition of two slightly shifted uniform coverages (figure 3.8).

The NEBNext and the ThruPLEX data sets exhibited the largest amplicons of about 800 bases. The size of displayed ThruPLEX targeted regions were approximately half the size of regions targeted by the NEBNext protocol (for details see table 3.4).

Amplicons of NEBNext and QIASeq data sets were created by targeting overlapping sub-regions using multiple primers which led to large contiguous regions (figure 3.8). Coverage tracks of both tagging protocols showed additively overlapping individual coverages of truncated, approximately Gaussian-like shape.

The highest region targeting efficiency was observed in smMIP data sets followed by QIASeq data sets. Alignments in the immediate vicinity of targeted regions stretched out approximately 100 bases in QIASeq data sets. For the NEBNext protocol, coverage was still observed up to 250 bases away from target regions. In ThruPLEX data sets, coverage was still observed 300 bases away from targeted regions.

The ratio of observed covered bases over targeted bases was highest for NEBNext (*e.g.* 37.7), followed by QIASeq (*e.g.* 11.7), ThruPLEX (*e.g.* 7.7), and finally smMIP (*e.g.* 2.3). These values correspond to the examples used in table 3.4.

Table 3.4: statistics of targeted regions for all tagging procedures.

Tagging Protocol	Initial Regions	Regions After Merging	Mean Region Base Width[1]	Total Bases Inside Regions	Example of Covered Bases[2]
QIASeq	4,832	1,832	309	566,976	6,643,582
NEBNext	190	12	3,062	36,747	1,385,086
ThruPLEX	745	671	752	504,652	3,880,446
smMIP	11	11	17	1,584	3,610

[1] Mean region base width and total bases refers to merged regions.

[2] Data sets used as coverage examples: QIASeq TruQ4, NEBNext 100% TruQ4, ThruPLEX MiSeq, and smMIP 100% TruQ4, replicate 1. The non-consensus, alignment filtered version of each BAM file was used.

Read Grouping

Read group sizes were visualized as rank size plots as described in [87, p. 4]. The size of a read group was defined as the number of read pairs that shared the group's UMI sequence. The height difference of hypothetical plateau regions observed for NEBNext and QIASeq is likely due to the 10-fold difference in input DNA material (figure 3.9).

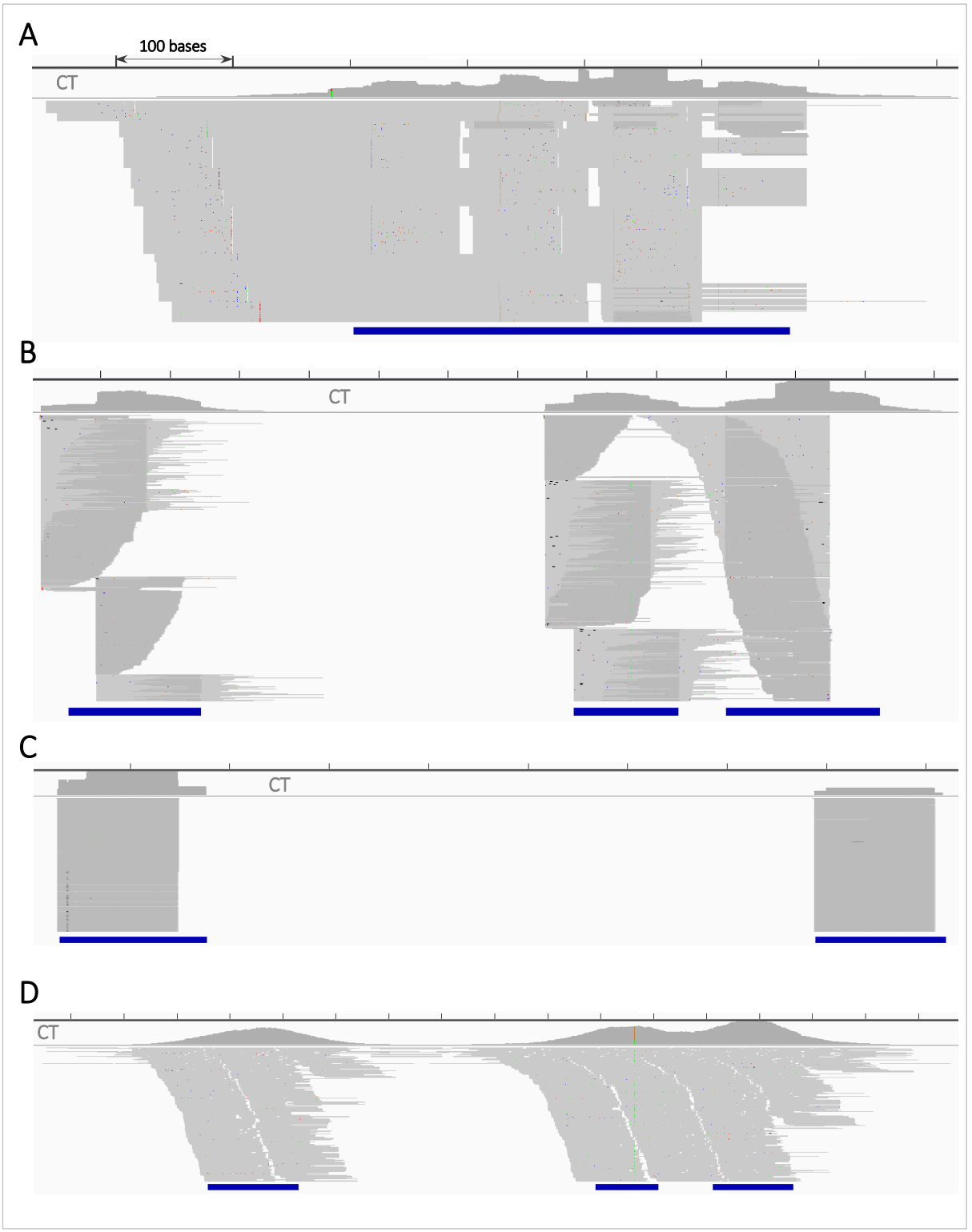


Figure 3.8: amplicon structures of investigated tagging protocols. *Alignment filters were applied. Reads are displayed as pairs (dark grey overlap) except for ThruPLEX. Only a portion of alignments is shown. Coverage tracks are marked with 'CT'. Blue bars below alignments indicate targeted regions. Tick marks are placed at a 100-base distance. Coloured dots label alternative base observations. (**A**) Structure of a NEBNext amplicon of the 100% TruQ4 sample. (**B**) Three amplicons from the QIASeq TruQ4 data set. (**C**) Two amplicons of the smMIP 100% TruQ4 sample, replicate one. (**D**) Three partially overlapping amplicons of the ThruPLEX MiSeq data set. Amplicon depictions were adapted from BAM visualizations from the IGV software.*

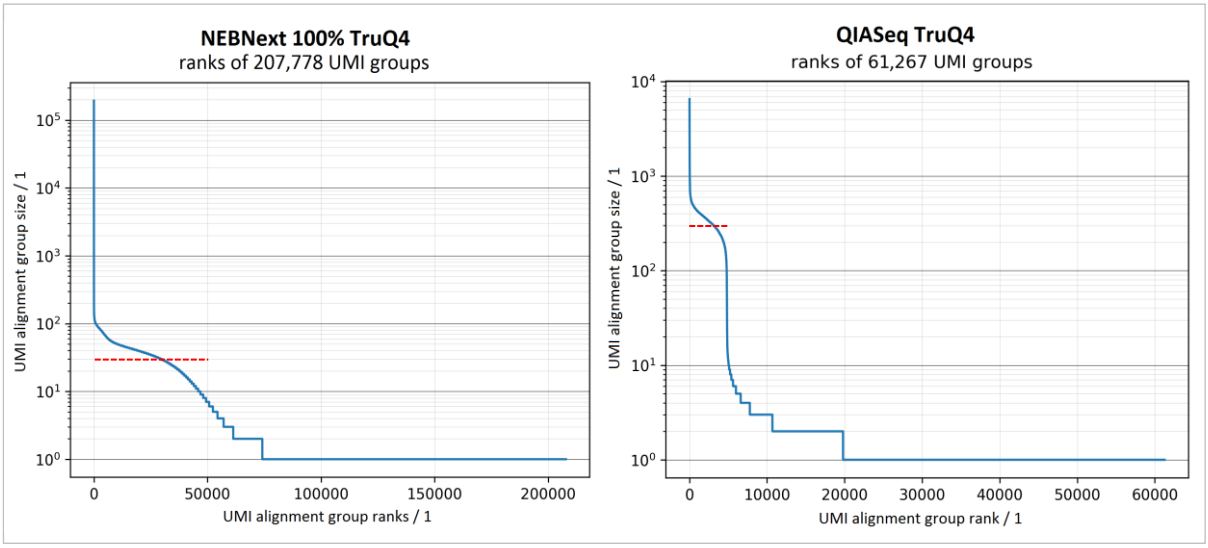


Figure 3.9: typical UMI group size rank plots for NEBNext and QIASeq data sets. *Grouping was based on UMI sequence only. Hypothetical plateau regions as would result from perfect amplification are drawn as red dashed lines. An extreme initial peak is present for NEBNext which is over three orders of magnitude higher than the pronounced shoulder region. The peak for QIASeq is only one order of magnitude higher than UMI groups forming the less pronounced shoulder region. The hypothetical plateau region of the QIASeq data set is shorter (approximately 6,000 read groups) and higher (around 300 copies) compared to the NEBNext plateau. This ideal NEBNext plateau is formed by 50,000 UMI groups and 30 copies per read group. Long tail regions are present for both data sets. These make up approximately 75% of NEBNext read groups and 90% of QIASeq read groups. Plots were created in Python using matplotlib.*

Based on the group rank plots for ThruPLEX and smMIP (figure 3.10), the number of observed read groups can be viewed as wildly inflated compared to the underlying true amount which can only be estimated.

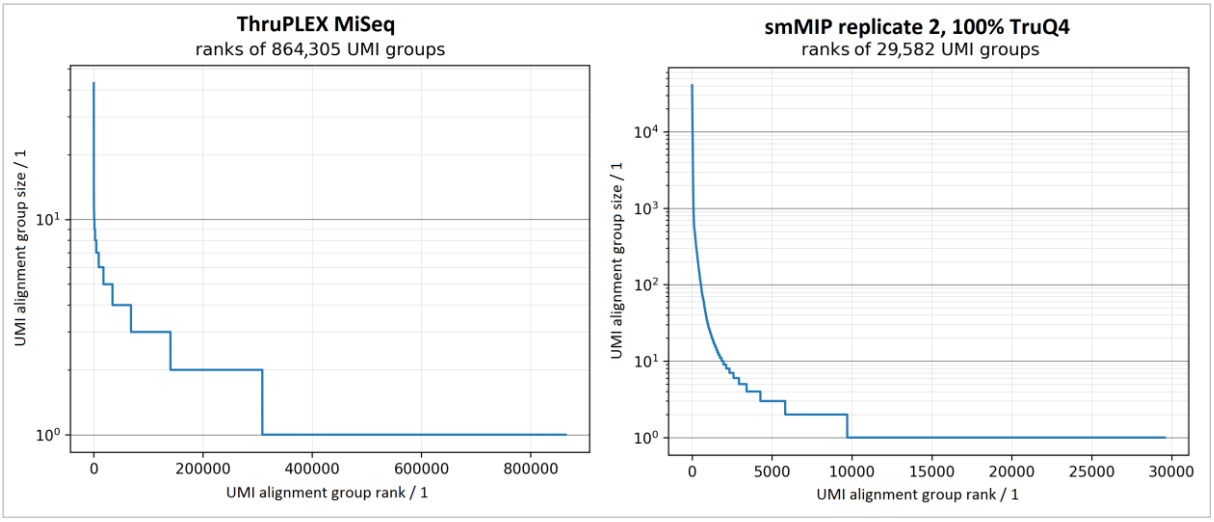


Figure 3.10: typical UMI group size rank plots of ThruPLEX and smMIP data sets. *Grouping was based on UMI sequences only. No plateau or shoulder region is visible. The sequence abundances can be fully characterized by a steep initial peak which directly goes over to a dominant tail region. The maximum observed read group size was around 43 for ThruPLEX while smMIP exhibited read group sizes up to 40,000. More than 50% of all read groups were formed by a single read pair. Plots were created in Python using matplotlib.*

Tagging Protocol Performance

The number of covered ground truth variants differed for each of the tagging procedures (for details see table 3.5). Statements about performance measures are based on Mutect variant calling results because Mutect outperformed smCounter in the exploratory analysis. The exception to this was smMIP which exhibited better variant calling performance with smCounter.

The highest GTV detection probabilities were observed for data sets of the ThruPLEX and the NEBNext protocols. Target regions of NEBNext exhibited the largest GTV coverage. Accordingly, the NEBNext pool of callable GTVs across all samples was approximately 8-times larger than the number of ThruPLEX GTVs (70 vs. 9). In summary, 13 out of 14 possible GTVs were detected in the NEBNext dilution series. The lowest VAFs of detected GTVs were observed for the 25% TruQ4 samples of smMIP and NEBNext at 1.00% and 1.05% for smCounter and Mutect respectively.

Table 3.5: GTV-based validation results for single data sets and dilution series.

Data Set	Variant Caller	Covered GTVs	GTV Recall[2]	Lowest VAF of Detected GTV	GTVs Never Called	GTVs Classified as Tumour
QIASeq	Mutect	7	0.286	16.70%	5	-
	smCounter		0.000	-	7	-
NEBNext 10 ng	Mutect	14	0.571	4.00%	6	-
	smCounter		0.429	5.00%	8	-
NEBNext 100 ng	Mutect	14	0.857	4.00%	2	-
	smCounter		0.357	5.00%	9	-
NEBNext TruQ4 dilution series	Mutect	14	0.738	1.05%	1	61.5%
	smCounter		0.310	1.05%	6	31.7%
ThruPLEX MiSeq	Mutect	9	1.000	4.00%	0	-
	smCounter		0.111	30.00%	8	-
ThruPLEX NextSeq	Mutect	9	1.000	4.00%	0	-
	smCounter		0.667	5.00%	3	-
smMIP[1]	Mutect	11	0.091	2.50%	10	0.0%
	smCounter		0.424	1.00%	6	68.8%

[1] Average values across replicates were used for smMIP.

[2] Recall values were not corrected.

The portion of initial alignments, which were still available for variant calling after applying the alignment filter, was 91.9% for the NEBNext 100 ng data set. In the ThruPLEX MiSeq analysis, approximately one third of initial alignments were removed by the alignment filter (*i.e.* off-target,

unmapped, not in proper pair, or reads of a pair mapped to different chromosomes). The portion of off-target reads made up the majority of removed reads.

For the NEBNext 100 ng input DNA material data set in combination with Mutect variant calling, GTV recall was 1.5-times higher than for the 10 ng Mutect analysis. There was no observable effect on the GTV VAF detection limit for these two data sets though. Compared to both NEBNext data sets, the GTV detection performance was worse for the 10 ng QIASeq data set. Mutect could only detect 2 GTVs, both exhibiting a VAF deviation of +15%.

The ThruPLEX MiSeq sample showed GTV VAF deviations below 6% around the expected values. Maximum deviations called for the ThruPLEX NextSeq sample were slightly higher. The NEBNext dilution series exhibited GTV VAF deviations similar to the ThruPLEX MiSeq sample. Few exceptions of higher deviations might be attributed to the higher overall number of issued GTV calls. The largest deviation of +9.5% VAF for this dilution series was observed for a GTV which was expected at 8% VAF. However, the main portion of GTVs were of acceptable accuracy in terms of VAF variability (see figure 3.11).

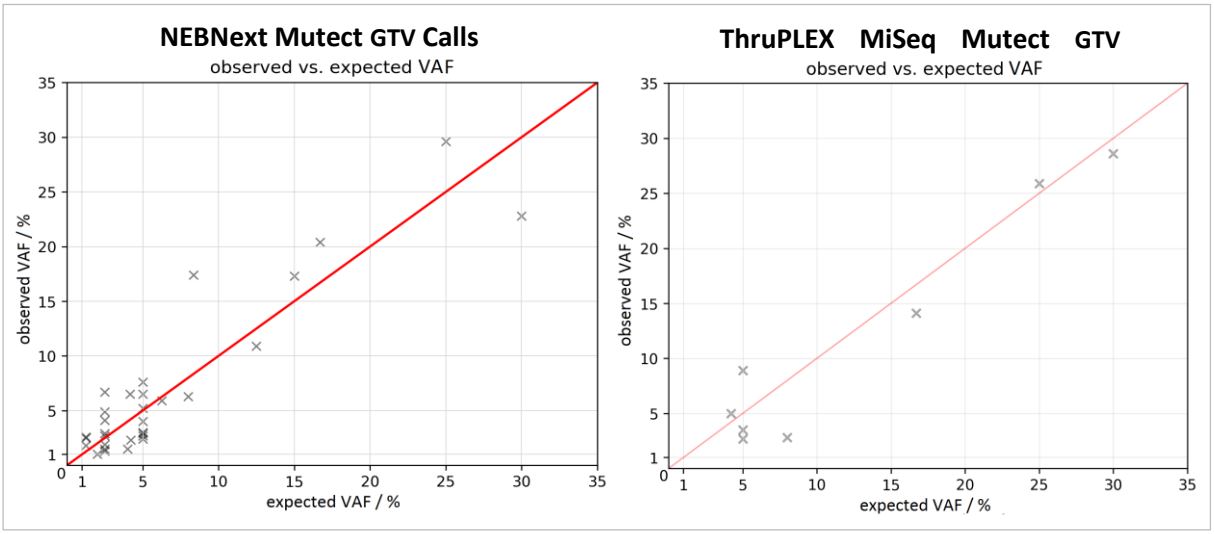


Figure 3.11: VAF of GTVs called by Mutect for the NEBNext dilution series and the ThruPLEX MiSeq data set. *Red lines indicate perfect concordance. Only GTV calls which could be identified automatically are shown. Expected variant allele frequencies taken from manufacturer homepage. Plots were created in Python using matplotlib.*

In case of Mutect variant calling on smMIP data sets, only 2 GTVs were called for replicate one and 4 GTVs for replicate two. These are the lowest results in terms of GTV recall for any conducted analysis involving the Mutect caller. Deviations of called GTV VAF from the expected values were between -17.5% and +0% VAF for replicate one and between -21.5% and +37.5% VAF for replicate two. The latter dilution series showed the largest deviations from expected VAF.

For smMIP, the amount of Mutect variant calls per replicate and dilution series was also low: 14 for replicate one and 11 for replicate two. Only one GTV call of replicate two was observed in another data set of the dilution series. Replicate one yielded no traceable variants. For smCounter, 26 and 14 variants were observed more than once in dilution series results of replicate one and replicate two respectively.

The smCounter results showed an increased GTV recall compared to Mutect results for smMIP (see figure 3.12). For replicates one and two, 18 and 10 GTVs out of 33 possible were called. Observed VAF variability was maximal for replicate two with deviations between -10% and +70%. The lowest expected VAFs of detected GTVs were 1% and 2% for replicates one and two respectively.

In replicate two, the lowest called variant allele frequency of a variant, which was observed more than once and classified as putative tumour, was 0.32% for Mutect and 0.26% for smCounter.

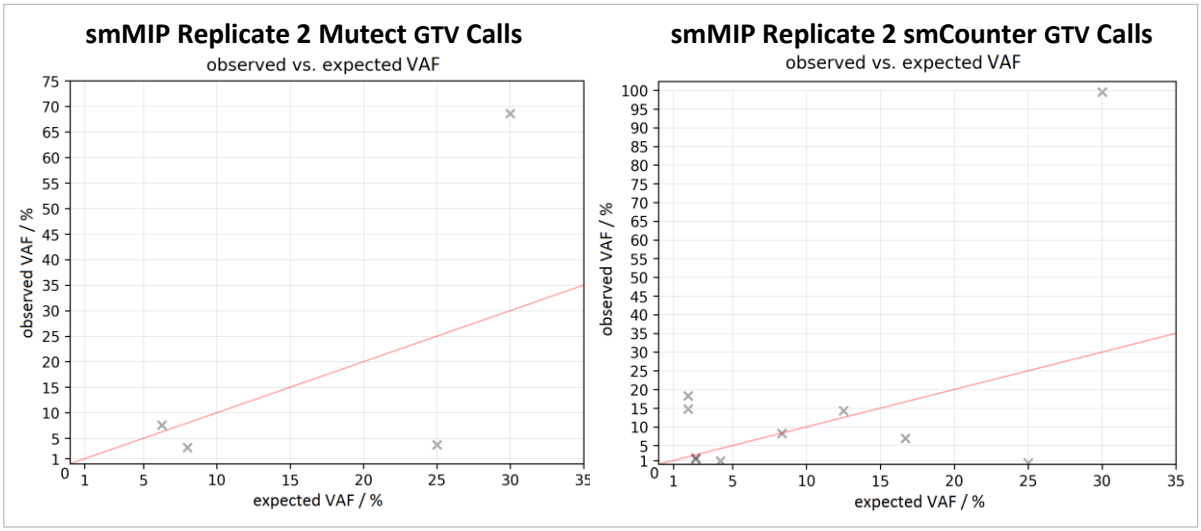


Figure 3.12: VAF of called GTVs of the smMIP replicate two dilution series. *Mutect and smCounter results shown. Red lines indicate perfect concordance. Only calls which could be identified automatically are shown. Expected variant allele frequencies taken from manufacturer homepage. Plots were created in Python using matplotlib.*

Concordance checks between Mutect calls of replicates resulted in a single shared variant call between replicates of the 100% and the 25% TruQ4 data sets. This equalled 23.8% and 25.0% on average of all made calls of the 100% and the 25% TruQ4 data sets. No shared variant was found between the 50% TruQ4 data sets. Shared calls were remarkably higher for the smCounter calling results. Eleven calls were shared between replicates of each TruQ4 dilution level which equals 8.8%, 7.3%, and 7.8% of all calls on average for the 100%, 50%, and 25% data sets respectively.

Validation results not based on GTV calls are listed in table 3.6. The highest amount of variant calls was observed for the ThruPLEX NextSeq data set. The number of total variant calls was 1.3-times higher than for the NEBNext dilution series, which exhibited the second highest amount of variant calls. NEBNext yielded more variant calls than the ThruPLEX MiSeq analysis despite regions targeted by ThruPLEX covered 13.7-times more bases than targeted regions of the NEBNext protocol (see table 3.4). The number of variant calls made with the QIASeq-Mutect combination was comparable to the amount of NEBNext 100 ng Mutect calls. The smMIP dilution series experiment resulted in the lowest number of variant calls albeit including results of three VCF files.

Table 3.6: variant call validation results based on ambiguous calls.

Tagging Protocol	Best Variant Caller	Unique Variants[1]	Calls per Region Kilo Base and VCF File	Traceable Variants[2]	Putative Tumour	Very Strong Linear Relation[3]	Flat Reg. Slope
QIASeq	Mutect	470	0.8	-	-	-	-
ThruPLEX MiSeq	Mutect	2,689	5.3	-	-	-	-
ThruPLEX NextSeq	Mutect	4,468	8.9	-	-	-	-
NEBNext 10 ng	Mutect	380	10.3	-	-	-	-
NEBNext 100 ng	Mutect	439	11.9	-	-	-	-
NEBNext	Mutect	3,532	32.0	17.1%	39.3%	8.3%	14.9%
smMIP[4]	Sm-Counter	176	37.0	11.1%	42.3%	25.0%	9.6%

1 The term 'unique variant' refers to the indicated variant itself and not the individual variant call.

2 Trace-based results are only available for dilution series.

3 Strength of linear relation refers to VAF of traceable variants with more than two observations.

4 The average across replicates is shown for smMIP.

The number of variants which were called more than once over different tumour dilution levels was higher for the NEBNext dilution series than for the smMIP series but fell short of expectations. Likewise, the percentage of variants categorized as originating from the tumour portion was below expectations for both series. The portion of variant traces which showed a very strong linear relation between dilution levels was low in all cases. The average number of unfavourable variant traces which exhibited only minor changes in VAF over dilution levels (*i.e.* flat regression slope) was 1.5-times higher for NEBNext than the smMIP average.

Due to the length of UMI sequences used by the NEBNext tagging protocol, 4^{12} equalling 16,777,216 different UMI sequences were possible for every data set. Only 79,072 UMI sequences were observed for the NEBNext 100% TruQ4 data set which corresponds to 0.47% of all possible UMI sequences. Therefore, observations of highly similar UMI sequences at the same mapping location (*i.e.* low edit distance at the same 5'-most covered base) should be rare.

UMI Error Correction

The error reduction performance of four approaches was tested on the NEBNext 100% TruQ4 data set. An edit distance distributions per mapping position between UMI sequences of alignments was calculated by UMI-tools (figure 3.13). The directional-adjacency performed best in reducing UMI sequence errors in this thesis and in the UMI-tools paper.

Corrected results for all approaches showed a bi-modal edit distance distribution. Results for the two best performing approaches (directional-adjacency and cluster) were highly similar.

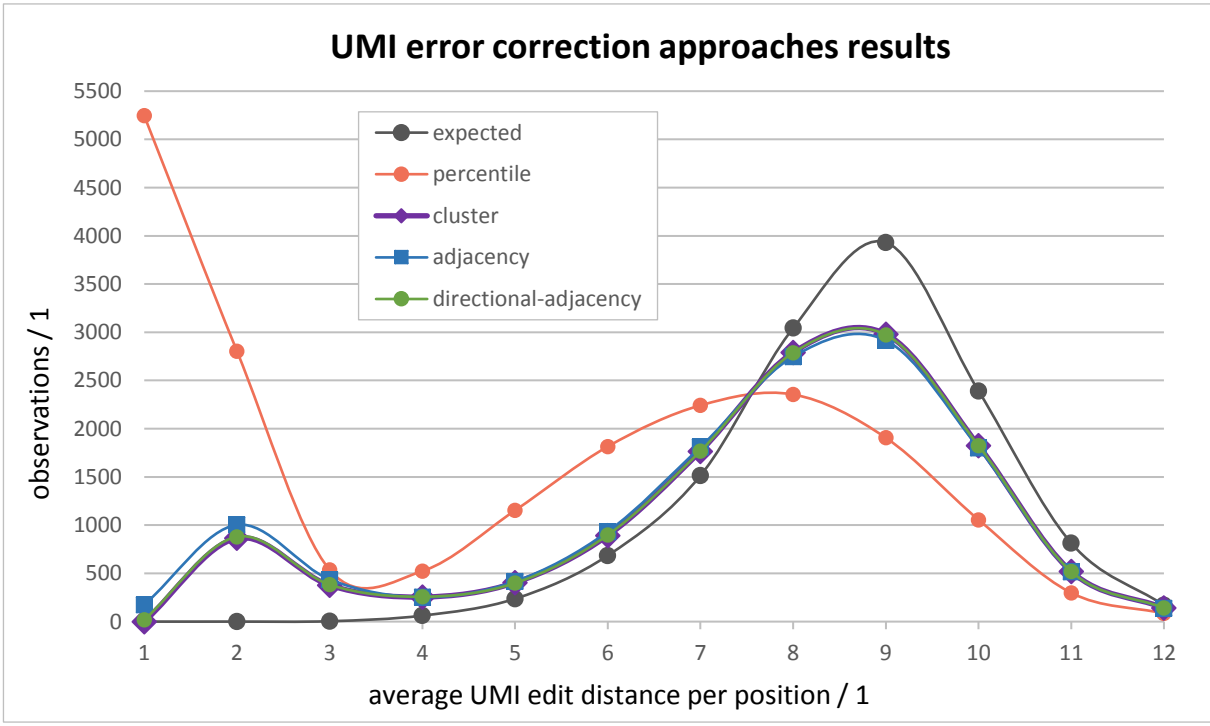


Figure 3.13: results of the test of UMI error correction approaches on the NEBNext 100% TruQ4 data set. *Spline interpolation curves were drawn to increase visibility of individual edit distances trends. The UMI edit distances per position after correction are shown. The expected distribution was assumed to be ideal. The plot was created with Microsoft Excel.*

After applying the directional-adjacency correction method, the number of observed UMI groups of size one and two decreased the most (figure 3.15). Observations of group sizes above 5 changed

only to a minor extent (figure 3.14). Larger variations observed for higher group sizes are due to smaller absolute numbers of observations.

Based on the findings above and the recommendation by the UMI-tools paper, the directional-adjacency approach was used in the final variant calling pipeline.

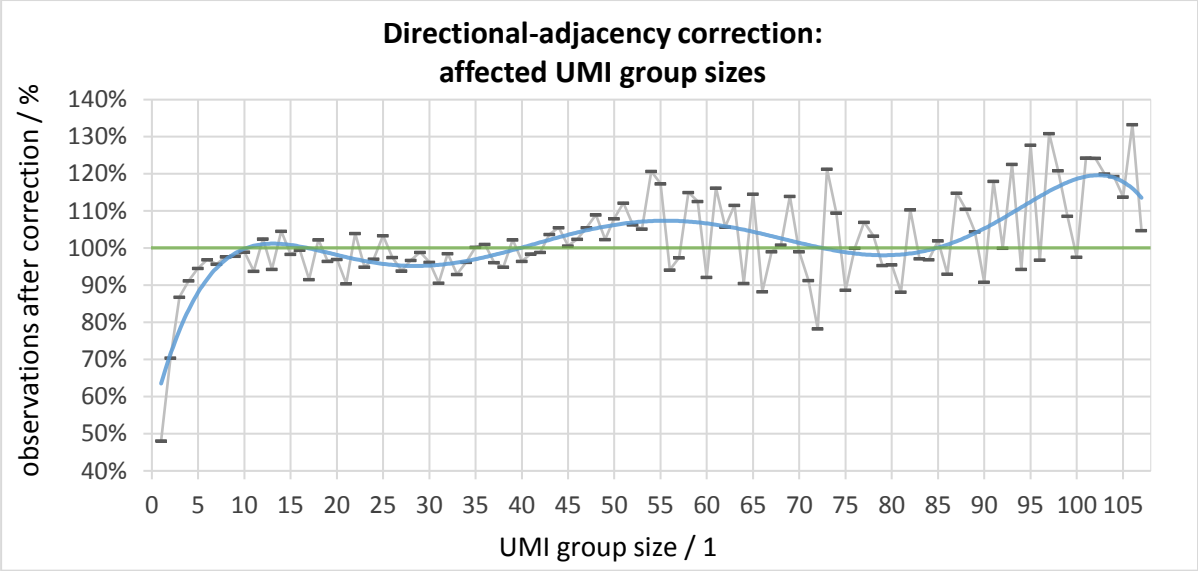


Figure 3.14: changes in UMI group size observations after applying the directional-adjacency UMI error correction. *Group size is the number of alignment pairs with identical mapping positions that share the group's UMI. Observations are relative to the size count before UMI-tools correction. The green 100% line denotes no change in observations of UMI group size. The blue curve is a polynomial fit of degree 6 to indicate the size change trend. Light grey lines were drawn to increase the visibility of the reduction trend. The plot was created with Microsoft Excel.*

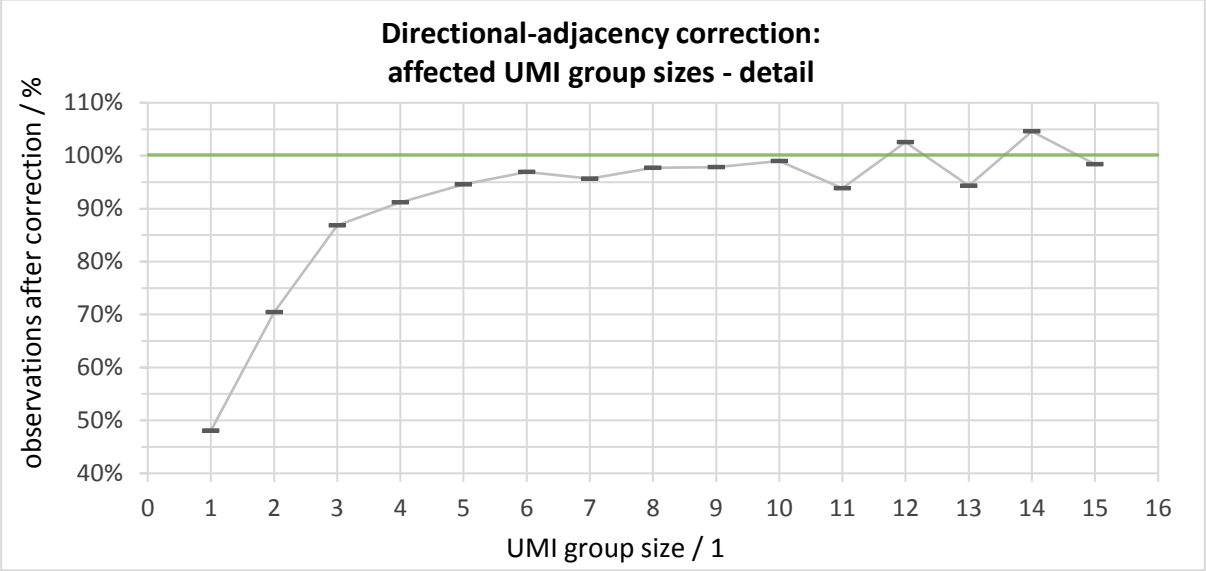


Figure 3.15: detail of changes in low UMI group size observations due to applying the directional-adjacency correction method of UMI-tools. *Group size is the number of alignment pairs with identical mapping positions that share the group's UMI. Observations are relative to the size count before UMI-tools correction. The green 100% line denotes no change in observations of UMI group size. Light grey lines were drawn to increase the visibility of the reduction trend. Observations of UMI groups of size one are reduced to approximately 50% compared to the uncorrected BAM file. The plot was created with Microsoft Excel.*

Recall decreased for the NEBNext dilution series from 0.881 for the uncorrected BAM file to 0.643 for the corrected BAM file after deduplication by UMI-tools. Instead of 5 missed calls, 15 GTV calls were missed after the correction which equals an increase in GTV detection loss of a factor of 3. The amount of variant calls was reduced from 883 for the non-consensus Mutect analysis to 272 for the corrected and deduplicated analysis. This equals a reduction factor of 3.2. For comparison, the exploratory analysis for NEBNext without UMI error correction of the collapsed reads resulted in 3,532 unique variants (table 3.6) and a GTV recall of 0.738 (table 3.5).

The GTV VAF variability mildly decreased on average after correction for variants with an expected VAF over 5% (figure 3.16). The average decrease was less pronounced for variants below 5%.

Background Estimation

Variant calling results of the NEBNext TruQ4 dilution series were used to investigate the large portion of ambiguous variant calls. Two categories of these ambiguous calls were of special interest: variants which were called only once in all data sets of the dilution series ('untraceable' variants) and 'traceable' variants (*i.e.* observed more than once) which exhibited only a minor change in VAF over different tumour levels. This means that linear regressions for the VAF of these traceable variants exhibited flat-angled slopes below 20%. For example, if between the 100% TruQ4 data set and the 50% TruQ4 data set a variant's VAF changed by less than 10% (a change of 50% would have be expected), the variant was labelled unclassifiable.

A major portion of approximately 70% of all 4,438 variant calls were untraceable, *i.e.* background variants (figure 3.17). In contrast, only one out of 13 GTVs fell into the untraceable category.

Of 605 traceable variants, 90 exhibited a flat regression slope and could not be classified as either tumour or wildtype. This equals 14.9% (see table 3.6). Four out of 13 GTVs exhibited a linear regression slope below 20%.

Therefore, the untraceable category might be a better classification of erroneous calls than the unclassifiable category.

Other source-classifiable ambiguous variants were investigated to answer the question whether their observed VAF could be predicted by a downward trace linear regression (figure 3.18).

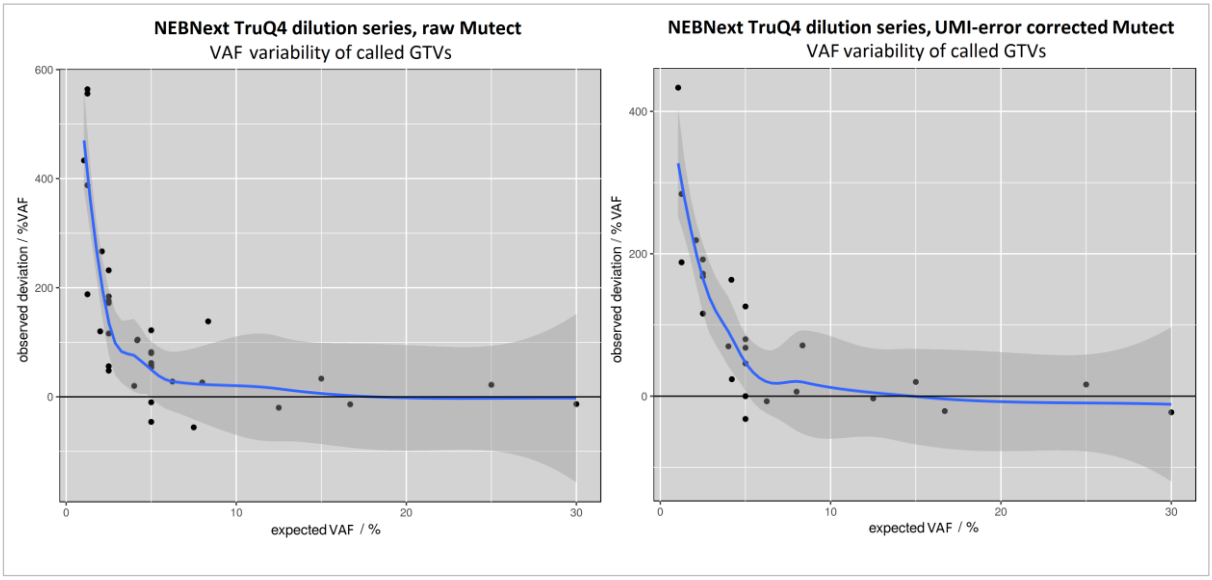


Figure 3.16: VAF deviations before and after UMI sequence error correction with UMI-tools' directional-adjacency approach. *The blue curve was interpolated using locally estimated scatterplot smoothing (LOESS). The dark grey area around the interpolated curve is the curve's 95% confidence interval. The results on the left were obtained with the annotated non-consensus BAM file. Error corrected results were obtained from the deduplicated BAM file which was created with UMI-tools. Plots were created in R using the ggplot2 package.*

Variants were observed more frequently at higher VAFs than at lower VAFs compared to the expected VAF as calculated by regression. This is because of a simplification that was made during the estimation of the expected VAF by using a downward trace that assumed perfect accuracy for the highest VAF of a traceable variant.

The variability of expected VAFs of ambiguous variants was higher than for NEBNext dilution series GTVs (compare results of figures 3.18 and 3.11, left side).

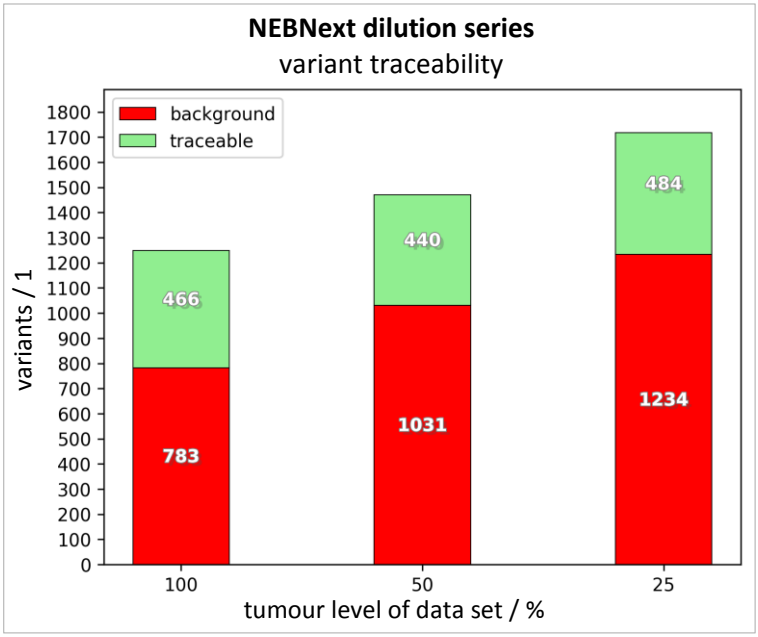


Figure 3.17: traceability of Mutect variant calls per data set of the NEBNext dilution series. *Variants occurring only once throughout the dilution series were labelled 'background'. The plot was created in Python using matplotlib.*

A slight majority of 53.6% of source-classifiable variants were classified as putative wildtype variants. In contrast, all traceable and source-classifiable GTVs were correctly classified as putative tumour.

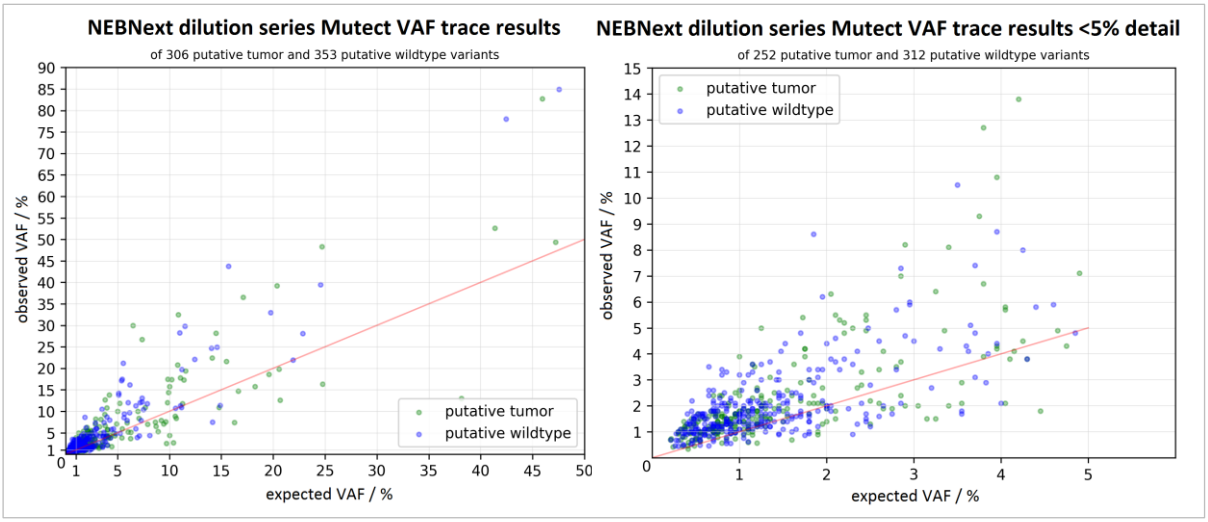


Figure 3.18: observed and expected VAF of traceable ambiguous variants. *Expected VAF estimated from the highest VAF of each traceable variant. Only estimates for calls below the highest observed VAF are depicted. The red line marks perfect concordance. Blue dots mark putative wildtype variants. Putative tumour variants are green. Observed VAF deviations vary between approximately -50% and +400% of the expected VAF. Most ambiguous variants were observed below 15% VAF and of that subgroup, the majority was observed below 5% VAF. Plots were created in Python using matplotlib.*

Improvements for Reanalysis

The mate-merging step was removed to maximize coverage and to avoid an additional, potentially read sequence-altering error source. Alignments were filtered more lenient to increase coverage as described in the paragraph 'Structure' of subsection 2.11 and paragraph 'Illumina Adaptor Trimming and Mapping' of subsection 3.3.

To eliminate read group mapping position annotation errors, the UMI group mapping position annotation was applied after read mapping. Annotation errors manifested as unsatisfactory GTV detection performance for smCounter variant calls for data sets with a larger range of possible mapping positions per amplicon (*i.e.* data sets created with a tagging protocol other than smMIP).

UMI error correction using the UMI-tools software was made mandatory because of the positive impact on background variant call reduction and the notable reduction of single alignment pair clusters. The directional-adjacency error correction was refined by adding a mate mapping position criterion to enhance specificity.

The UMI coverage-dependent runtime of the UMI error correction posed a problem in cases where deeper sequenced amplicons had to be computed. Since this was the case for most data sets in this

thesis, the need for a practical workaround arose. A solution was found by sorting alignments according to their expected sequencing error in an ascending manner followed by subsampling.

To implement additional approaches for alignment reduction based on UMI group similarity, two additional clustering steps were added after UMI-tools alignment grouping correction. Clustering of alignment groups allowed for combined reduction of alignments with an identical UMI and slightly different template lengths. The first 'basic' clustering acted as a technical duplicate detection in case of mapping position deviations which were frequently observed in reads with an initial homopolymer sequence which is prone to indel errors. The second 'advanced' clustering allowed for correction of three types of template molecule altering error events that might occur during a single synthesis step. Since this was expected to happen very rarely, the advanced clustering was made optional.

3.5. Reanalysis

Profiling

The reanalysis procedure was divided into the following tasks for profiling: read pre-processing, mapping, alignment processing, UMI error correction, clustering, and variant calling. The proportions of every task's contribution to the overall execution time could only be approximated, because server load in terms of main memory usage and file access operations varied highly. This exerted a higher impact on the overall longer reanalyses compared to the exploratory analyses. Execution times were computed from time stamps which were documented in master log files (table 3.7).

Table 3.7: total execution times of reanalyses.

Data Set[1]	Read Pairs	Mean Read Length After Filtering	Total Duration[2]	Duration Per Read Mega Base	Duration Increase Factor[3]
QIASeq TruQ4	1,808,103	144.4 bp	1 d 17 h 41 m 40 s	287.4 s	28.3
QIASeq 2% VAF Seraseq	28,634,852	138.4 bp	15 h 23 m 46 s	7.0 s	-
QIASeq 1% VAF Seraseq	20,747,543	136.8 bp	10 h 29 m 24 s	6.7 s	-
QIASeq 0.5% VAF Seraseq	23,177,862	138.3 bp	11 h 49 m 39 s	6.6 s	-
QIASeq 0.25% VAF Seraseq	18,369,432	136.3 bp	9 h 11 m 05 s	6.6 s	-
QIASeq 0.125% VAF Seraseq	18,862,323	138.1 bp	9 h 52 m 26 s	6.8 s	-
QIASeq WT Seraseq	24,576,543	137.5 bp	12 h 43 m 32 s	6.8 s	-

NEBNext 10 ng	3,976,440	75 bp	2 h 03 m 20 s	12.4 s	1.9
NEBNext 100 ng	2,886,909	75 bp	2 h 21 m 42 s	19.6 s	2.7
NEBNext 100% TruQ4 mix	2,271,218	76 bp	3 h 20 m 53 s	34.9 s	3.3
NEBNext 50% TruQ4	2,182,469	76 bp	3 h 06 m 15 s	33.7 s	3.2
NEBNext 25% TruQ4	2,798,019	76 bp	3 h 44 m 29 s	31.7 s	3.1
NEBNext 0% TruQ4	2,731,760	76 bp	4 h 55 m 32 s	42.7 s	-
NEBNext 2% VAF Seraseq	26,397,020	74 bp	5 h 39 m 30 s	5.2 s	-
NEBNext 1% VAF Seraseq	24,788,285	74 bp	5 h 14 m 25 s	5.1 s	-
NEBNext 0.5% VAF Seraseq	26,344,790	74 bp	6 h 05 m 50 s	5.6 s	-
NEBNext 0.25% VAF Seraseq	23,817,414	74 bp	5 h 02 m 44 s	5.2 s	-
NEBNext 0.125% VAF Seraseq	23,806,898	74 bp	5 h 10 m 57 s	5.3 s	-
NEBNext WT Seraseq	23,532,247	74 bp	5 h 10 m 38 s	5.4 s	-
ThruPLEX MiSeq	1,451,625	147.3 bp	2 d 07 h 40 m 08 s	468.6 s	14.3
*ThruPLEX NextSeq	13,762,371	-	-	-	-
smMIP 100% TruQ4 rep1	788,977	129.2 bp	31 m 00 s	9.1 s	1.9
smMIP 50% TruQ4 rep1	767,418	129.9 bp	47 m 54 s	14.4 s	3.0
smMIP 25% TruQ4 rep1	810,676	129.7 bp	49 m 39 s	14.2 s	3.1
smMIP 100% TruQ4 rep2	1,097,521	129.3 bp	37 m 09 s	7.9 s	1.7
smMIP 50% TruQ4 rep2	732,203	130.0 bp	44 m 34 s	14.0 s	2.7
smMIP 25% TruQ4 rep2	1,460,267	129.7 bp	1 h 05 m 19 s	10.3 s	2.6

1 Terminated analyses are marked with an asterisk.

2 Total durations were measured by a Python driver script (server loading not monitored).

3 Duration increase was computed relative to exploratory analyses.

The total execution time per analysis was approximately distributed over these tasks as follows: 5% read pre-processing, 5% mapping, 20% alignment processing, 5% UMI error correction, 50% clustering, and 15% variant calling. These percentages varied by up to 65% due to server loading which frequently led to suspending the main analysis thread. For example, variant calling step took longest (approximately 70%) in the ThruPLEX MiSeq reanalysis, while read processing took

longest for the QIASeq 0.5% GTV VAF Seraseq reanalysis. In contrast, clustering took about 90% of the QIASeq TruQ4 reanalysis. Factors of overall duration increases ranged from 1.7 to 28.3.

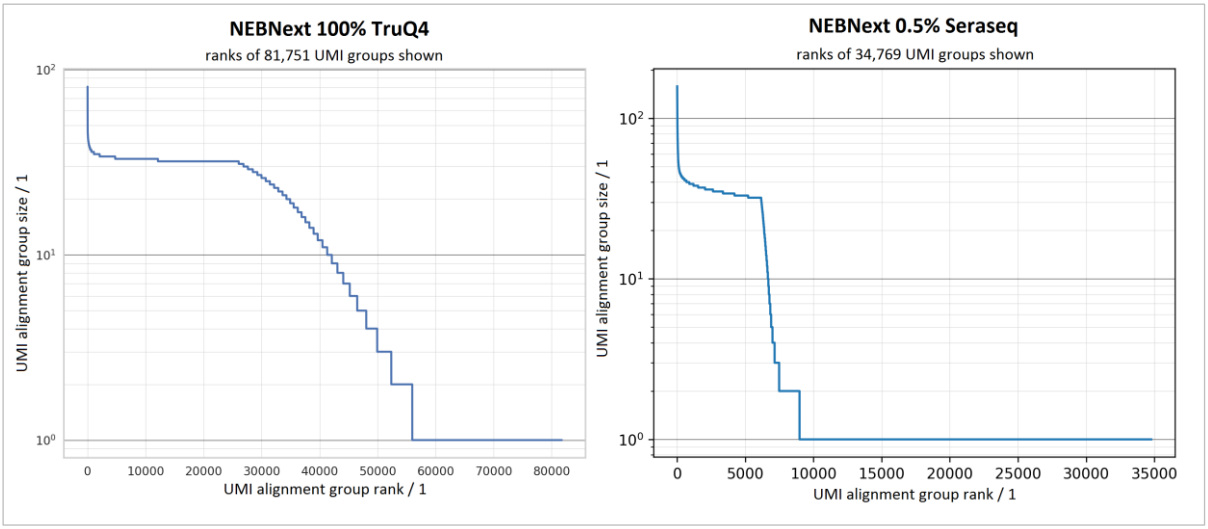


Figure 3.19: ranked UMI group sizes of subsampled and UMI error corrected TruQ4 and Seraseq NEBNext data sets. *Grouping was based on mapping position and UMI sequence. Plots were created in Python using matplotlib.*

The variant calling duration was reduced by using a separate BED file which defined targeted regions around GTVs including a border region width of 10 bases.

The ThruPLEX NextSeq analysis was terminated by the server because of excessive main memory usage during clustering. At the time of termination, the analysis process accumulated nearly 500 GB of memory usage.

Seraseq Alignment Group Sizes

After UMI error correction, more than 60% of alignment groups of the NEBNext 100% TruQ4 data set were of size 3 or more and, thus, could safely compensate a single base error per position during consensus formation (*i.e.* clusters were not error-prone). The ranks of UMI error corrected alignment groups of two NEBNext data sets are displayed in figure 3.19.

In the ThruPLEX MiSeq data set, the UMI error correction had no effect on the shape of the size ranking besides the reduction of UMI groups of size one. In the QIASeq TruQ4 data set, the UMI error correction and subsequent clustering led to exponentially decreasing ranked sizes of UMI groups which did not exhibit any kind of plateau region. In smMIP data sets, the alignment reduction steps resulted only in minor changes. The majority of UMI groups were still formed by a single alignment pair. An example for UMI group size ranks of smMIP and QIASeq data sets is displayed in figure 3.20.

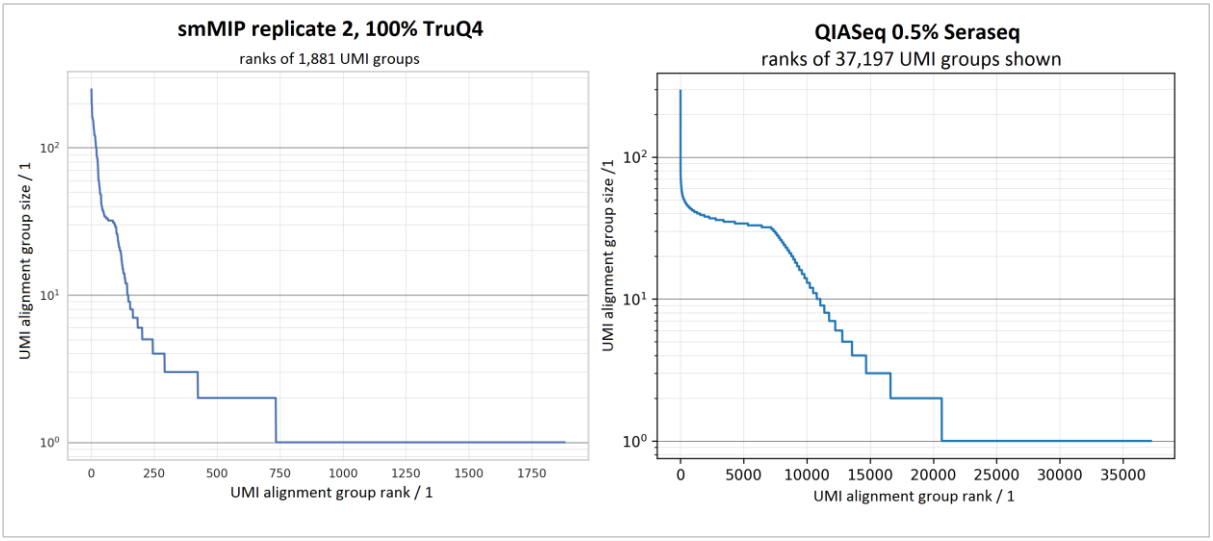


Figure 3.20: ranked UMI group sizes of a subsampled and UMI error corrected smMIP and a QIASeq Seraseq data set. *Grouping was based on mapping position and UMI sequence. Plots were created in Python using matplotlib.*

The results for Seraseq data sets either created with the QIASeq or the NEBNext tagging procedure responded better to alignment reduction by grouping, UMI error correction, and clustering. Besides the TruQ4 NEBNext data set, data sets of both Seraseq dilution series exhibited the most pronounced plateau regions after subsampling and UMI error correction. Approximately 20% of all NEBNext and 45% of all QIASeq alignment groups and subsequently created alignment clusters were of size three or larger and, thus, not error prone in consensus formation. Nevertheless, this still left 80% final NEBNext and 55% final QIASeq clusters error-prone.

Although some of the presented UMI group size ranking results responded quite well to UMI error correction, a notable portion of all data sets remained in tail regions which consisted mostly of UMI groups of size one.

Tagging Protocol Performance

Results of variant call validations are listed in table 3.8. No results could be obtained for the ThruPLEX NextSeq data set because of excessive usage of main memory which repeatedly led to the termination of the analysis.

For TruQ4 analyses with Mutect, no change in GTV recall was observed for the NEBNext 100 ng data set. GTV recall increased by 0.143, 0.095, and 0.016 for the QIASeq, NEBNext TruQ4 dilution series, and smMIP data sets respectively for Mutect. Minor changes of minimum called GTV VAF (detection limit) were observed. In summary, the NEBNext and ThruPLEX tagging

protocols outperformed QIASeq and smMIP for TruQ4 data sets. The lowest detection threshold was achieved with NEBNext and smMIP in combination with the smCounter caller.

Table 3.8: GTV-based validation results for single data sets and dilution series.

Data Set	Variant Caller	Error-Prone Clusters[1]	Covered GTVs	GTV Recall[2]	Lowest VAF of Detected GTV	GTVs Never Called	GTVs Classified as Tumour
QIASeq	Mutect	Included	7	0.429	16.700%	4	-
	smCounter	Included		0.714	4.200%	2	-
NEBNext 10 ng	Mutect	Included	14	0.500	4.000%	8	-
	smCounter	Included		0.929	4.000%	1	-
NEBNext 100 ng	Mutect	Included	14	0.857	4.000%	2	-
	smCounter	Included		0.857	4.000%	2	-
NEBNext TruQ4 dilution series	Mutect	Omitted	14	0.833	1.050%	0	76.9%
	smCounter	Included		0.952	1.000%	0	100.0%
ThruPLEX MiSeq	Mutect	Included	9	0.666	4.200%	3	-
	smCounter	Included		1.000	4.200%	0	-
smMIP[3]	Mutect	Included	11	0.107	1.250%	8	50.0%
	smCounter	Included		0.727	1.000%	1	75.0%
NEBNext Seraseq dilution series	smCounter	Included	37	0.530	0.125%	4	49.2%
QIASeq Seraseq dilution series	smCounter	Included	26	0.677	0.125%	3	95.7%

1 The best results regarding parameter settings are displayed. These results were achieved with an optimized cluster merging
size threshold of 1.4. Optimal results of Seraseq analyses were created with deactivated advanced clustering. All TruQ4 analyses were carried out with permissive advanced clustering.

2 Recall values were not corrected.

3 The average across replicates is shown for smMIP.

For Seraseq analyses, the QIASeq protocol outperformed the NEBNext tagging procedure in terms of GTV recall, detection limit, fraction of traceable variants, percentage of variants classified as tumour, and 50% GTV observation VAF threshold (see table 3.10). The NEBNext Seraseq analysis yielded better results for false positive GTV observations in the control data set and number of total calls. It is debatable whether a smaller number of absolute calls in case of lower GTV recall indicates only lower noise or is also a reason for the recall difference.

The initial read length filter affected tagging procedure data sets differently. For QIASeq data sets including the Seraseq reference material (NextSeq sequencing platform), about 30,000 to 70,000 read pairs per data set (0.28% of QIASeq Seraseq WT and 0.16% of QIASeq Seraseq 1% VAF) were observed with at least one read sequence in pair being shorter than the read length threshold of 45 bases. All reads of NEBNext, smMIP, ThruPLEX, and the QIASeq TruQ4 data sets passed the read length filter.

Variant Caller Performance

As presented in table 3.8, the results for all data sets obtained with smCounter greatly increased in GTV calling performance compared to the exploratory analysis. smCounter performed better than Mutect in most cases. An exception to this was the NEBNext 100 ng data set for which smCounter and Mutect performed equally well.

GTV recall achieved with Mutect decreased for the NEBNext 10 ng and the ThruPLEX MiSeq data set by 0.071 and 0.333 respectively.

As displayed in figure 3.21, variant calling with smCounter resulted in a much larger number of variant calls than Mutect variant calling for the same dilution series. Between 20 to 33-times more variants were called with smCounter. The fraction of traceable variants called with smCounter was lower for the NEBNext TruQ4 dilution series than for the Seraseq dilution series (see figure 3.24).

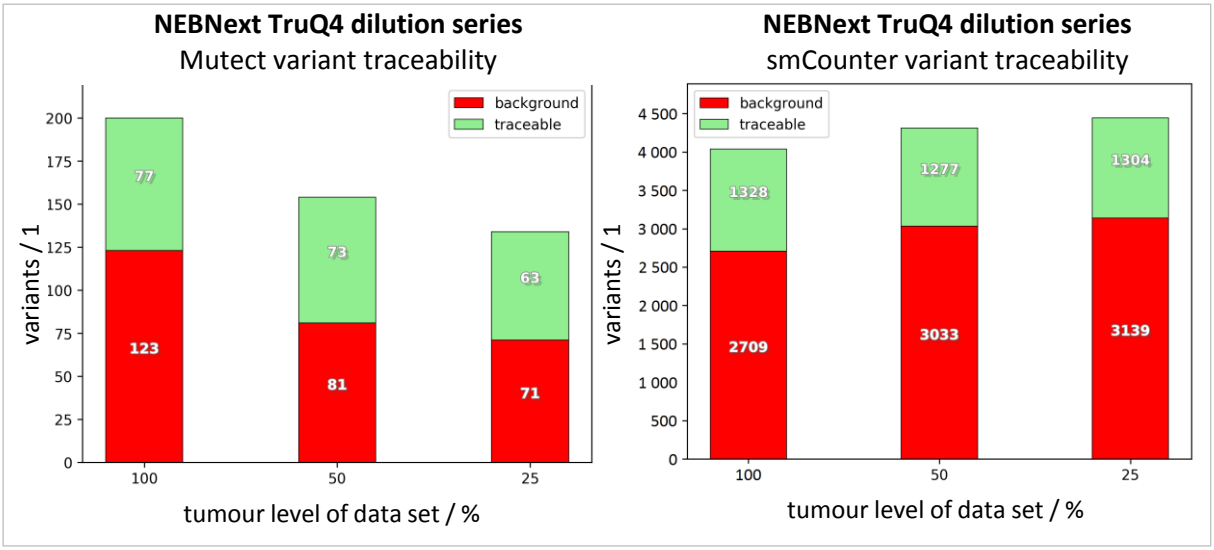


Figure 3.21: traceability of Mutect and smCounter variant calls per data set of the best performing NEBNext dilution series analysis with permissive advanced clustering. *Erroneous clusters were omitted for Mutect and included for smCounter. Variants occurring only once throughout the dilution series were labelled 'background'. Plots were created in Python using matplotlib.*

It is evident from trace results that the detection limit for smCounter was approximately 10-fold lower than for Mutect (figure 3.22). Also, the majority of smCounter variants were called at VAFs below 5%. Higher sensitivity of smCounter is also indicated by the portion of trace results of Seraseq analyses that exhibited a flat regression slope (see table 3.9).

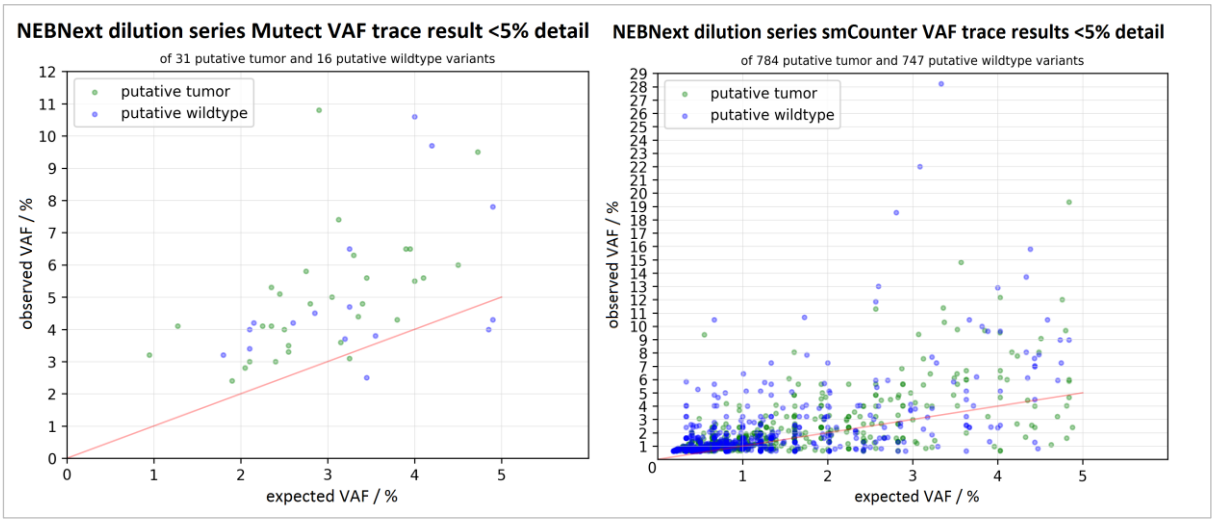


Figure 3.22: observed and expected VAF of traceable ambiguous variants of the Mutect and smCounter reanalyses. *Permissive advanced clustering was used. Error-prone clusters were omitted for Mutect and included for smCounter. Expected VAF estimated from the highest VAF of each traceable variant. Only estimates for calls below the highest observed VAF are depicted. The red line marks perfect concordance. Blue dots mark putative wildtype variants. Putative tumour variants are green. Plots were created in Python using matplotlib.*

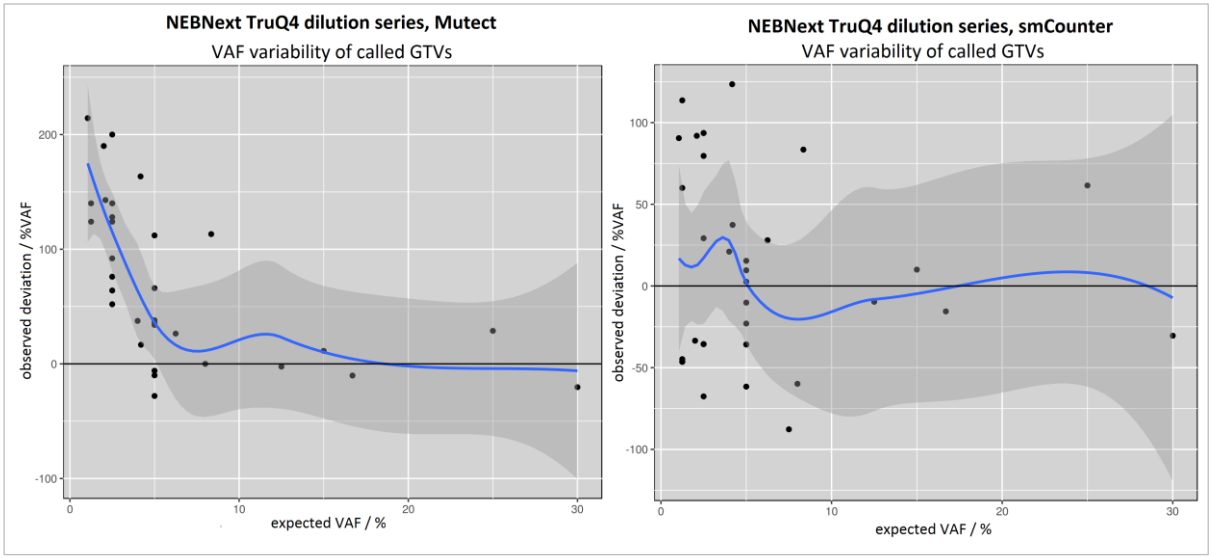


Figure 3.23: VAF deviations of called GTVs. *Best Mutect and smCounter results are shown for permissive clustering. The blue curve is an interpolation that uses LOESS. The dark grey area is the curve's 95% confidence interval. Error-prone clusters were omitted for Mutect and included for smCounter. Mutect frequently called GTVs with a deviation between 100% and 200% below 5% absolute VAF. SmCounter showed a more symmetrical deviation behaviour around the expected VAF. With few outliers, deviations were lower than 100% of the expected VAF. Plots were created in R using the ggplot2 package.*

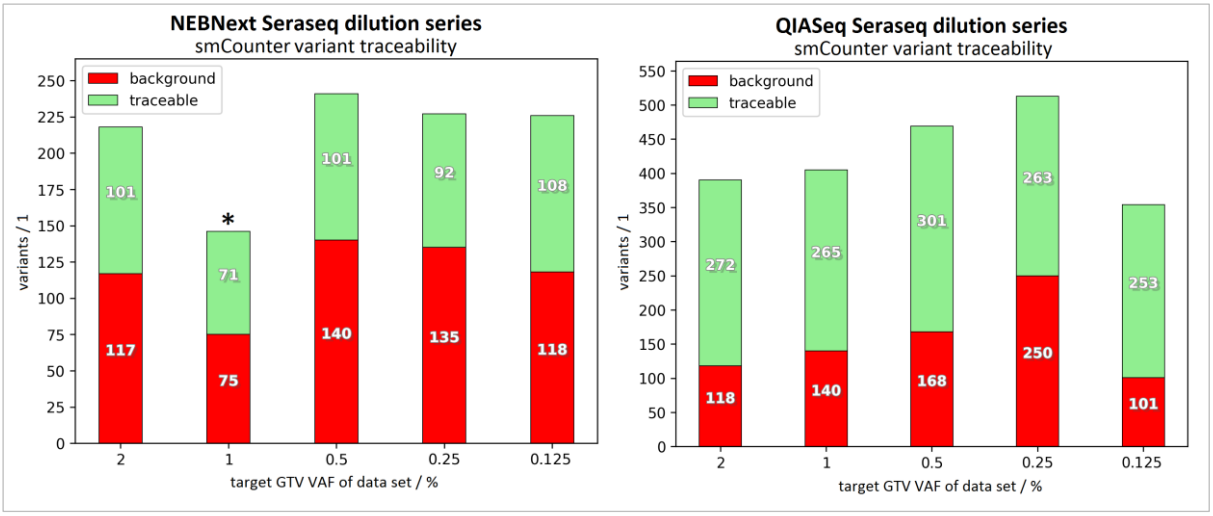


Figure 3.24: traceability of variant calls of the Seraseq analyses. *The NEBNext 1% GTV VAF data set was called with a narrower border region around the GTV loci of 3 bases instead of 10 for parameter optimization and should be disregarded for discussing the aspect of variant traceability. Variants occurring once in all dilution series data sets are labelled 'background'. Plots created in Python with matplotlib.*

Only the smCounter software was used for variant calling in Seraseq analyses because of the findings on caller performance obtained from TruQ4 analyses. The portion of background calls decreased immensely after pairing the smCounter caller with the QIASeq tagging procedure (figure 3.24). In general, VAF deviations and the detection limit were smaller for smCounter (figure 3.23).

Detection Limit Estimation

The absolute detection limit was estimated from the VAFs of single UMI-supported variant calls (SOCs) of a data set. It must be noted that at this VAF level, base noise and true GTVs were not distinguishable. Only statistical estimations about the portion of true SOC GTVs could be made. Individual detection limits of Seraseq data sets depending on the clustering settings (table 3.10), were estimated based on SOC VAF distributions (see figure 3.25).

The estimated detection limits of NEBNext Seraseq data sets increased and their VAF variability decreased only marginally by applying clustering settings with increased permissiveness. Only for the 0.125% Seraseq and the control data sets, an almost homogeneous detection limit with exception of a few outliers was observed.

The estimated detection limit increased by a factor of two between basic clustering and permissive clustering for QIASeq Seraseq data sets. In case of permissive clustering, UMI coverages were comparable over target regions. No variability was observed for the estimated detection limit.

While for basic and moderate clustering the main portion of NEBNext SOC VAFs reached the expected GTV VAF at 0.25%, the estimated detection limits for the same clustering settings of QIASeq reached the expected GTV VAF level one dilution level lower at 0.125% VAF. The detection limits for permissive clustering were approximately equal at 0.25% VAF for NEBNext and QIASeq.

Table 3.9: ambiguous variant call validation results.

Tagging Protocol	Variant Caller	Unique Variants[1]	Calls per Region Kilo Base and VCF File	Traceable Variants[2]	Putative Tumour	Very Strong Linear Relation[3]	Flat Regression Slope
QIASeq	Mutect	957	2.1	-	-	-	-
	smCounter	18,533	41.2	-	-	-	-
NEBNext 10 ng	Mutect	141	3.8	-	-	-	-
	smCounter	2,105	57.3	-	-	-	-
NEBNext 100 ng	Mutect	267	7.3	-	-	-	-
	smCounter	1,471	40.0	-	-	-	-
NEBNext TruQ4 dilution series	Mutect	330	4.2	25.5%	63.1%	12.8%	15.5%
	smCounter	10,185	112.4	16.7%	40.8%	4.2%	22.2%
ThruPLEX MiSeq	Mutect	8,226	16.6	-	-	-	-
	smCounter	299,883	605.8	-	-	-	-
smMIP[4]	Mutect	19	14.8	11.0%	75.0%	0.0%	25.0%
	smCounter	427	332.6	23.8%	53.2%	4.7%	15.3%
NEBNext Seraseq dilution series	smCounter	693	170.1	26.5%	49.2%	8.6%	16.0%
QIASeq Seraseq dilution series	smCounter	1,030	358.3	49.4%	95.7%	7.1%	21.2%

[1] The term 'unique variant' refers to the indicated variant itself and not individual variant calls.

[2] Trace results are only available for dilution series.

[3] The strength of linear relation refers to linear regression results of a variant's observed VAFs.

[4] The average across replicates is displayed for smMIP.

Thus, the estimated detection limits for basic and moderate clustering were lower for QIASeq than for NEBNext approximately by a factor of two.

Linear interpolations for the 50% GTV detection thresholds (table 3.10) resulting from the SOC-filtered GTVs were also found to be 3.5 to 4.0-times higher for NEBNext than for QIASeq.

Clustering

More permissive clustering settings increasingly reduced the amount of called variants as can be seen from the results listed in table 3.10. While clustering did not affect the estimated detection limit for NEBNext, the 50% GTV observation VAF threshold did increase with more permissive clustering settings for both tagging protocols. The amount of false positive GTV calls in control data sets was only reduced for QIASeq by 50% when comparing basic and permissive clustering results.

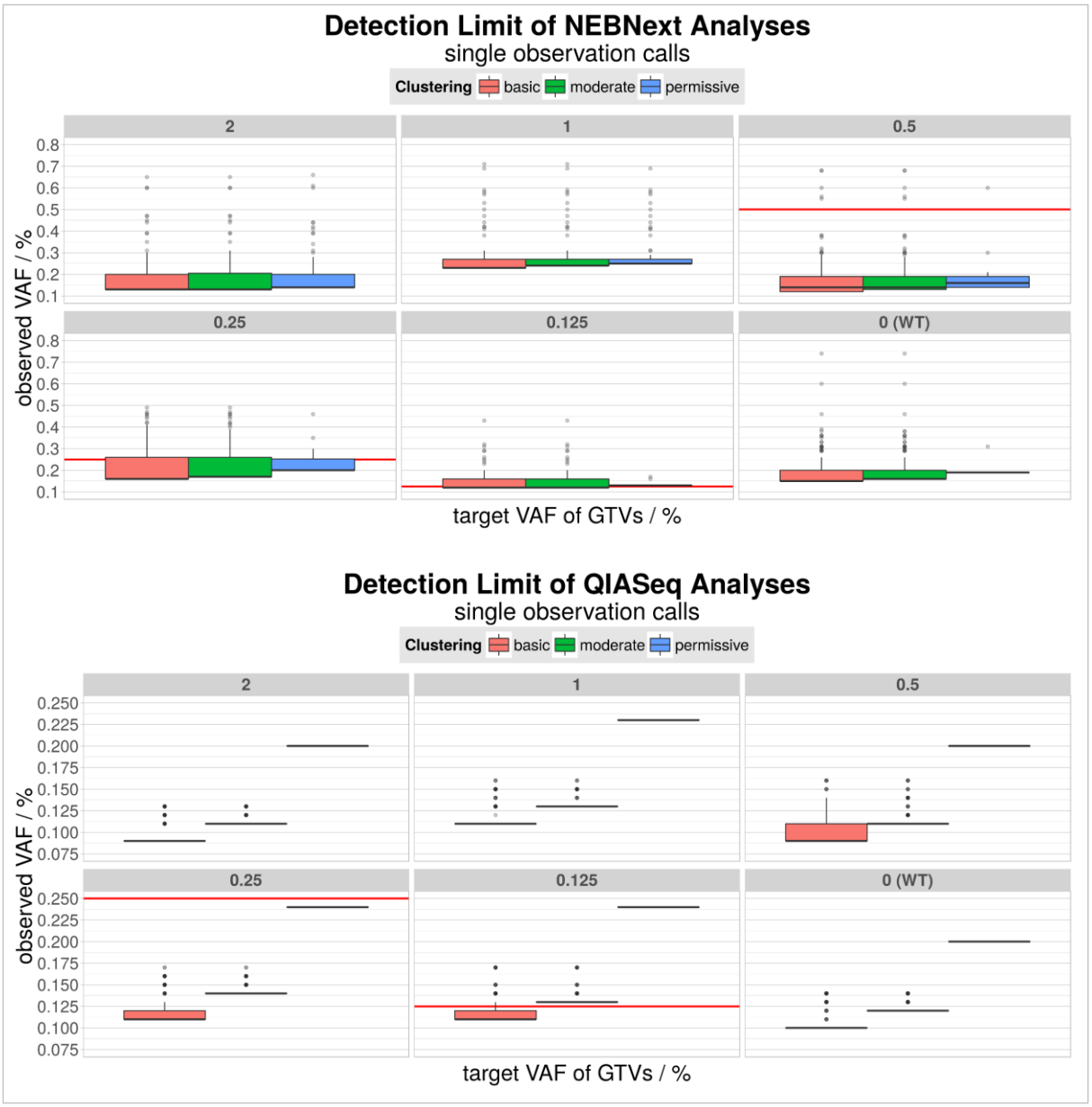


Figure 3.25: VAF distribution of SOCs of the NEBNext and the QIASeq Seraseq dilution series. *The expected VAF level of GTVs is displayed as a red line for each data set. In general, the estimated detection limit was lower for QIASeq analyses than for NEBNext analyses. Also, the VAF variability was generally lower for QIASeq SOCs than for NEBNext SOCs. Plots were created in R using the ggplot2 package.*

Measures describing corrected UMI artefacts, UMI coverage and cluster reduction are listed in table 3.11 for examples of Seraseq data sets. An estimate of expected errors per sequenced UMI was relatively constant 1 erroneous base in 100 sequences across NEBNext Seraseq data sets. For QIASeq Seraseq data sets, 1 in 30 UMIs was expected to carry a single base error. The estimation was based on the overall amount of expected base errors in filtered alignments, *i.e.* sum of translated base call quality values, and the ratio of UMI length to combined read length. Small leftmost mapping position deviations of groups with identical UMIs were observed approximately half as frequent up to equally frequent as random UMI base errors for QIASeq (results not listed in table 3.11). These mapping position deviations were observed less frequent for NEBNext Seraseq data sets. Mapping deviation observations ranged from 2 to 3 orders of magnitude lower up to one fourth of the number of corrected UMI base errors.

The estimated UMI base error frequencies for TruQ4 data sets were as follows: 1 base in 170 UMIs for ThruPLEX MiSeq and smMIP 100% TruQ4 replicate two, 1 base in 115 sequences for QIASeq TruQ4, and 1 base in 25 UMIs for NEBNext 100% TruQ4.

Table 3.10: performance measures for Seraseq dilution series and different clustering settings.

Data Set	Advanced Clustering	GTV Recall[1]	Number of Unique Variants	Estimated Detection Limit[2]	VAF threshold 50% GTV observations	WT False Positive GTV Calls
NEBNext	off	0.530	693	0.200%	0.687%	3 (8.1%)
	basic	0.524	682	0.200%	0.750%	3 (8.1%)
	permissive	0.481	336	0.200%	0.844%	3 (8.1%)
QIASeq	off	0.677	1,030	0.112%	0.188%	6 (23.1%)
	basic	0.654	1,006	0.125%	0.189%	5 (19.2%)
	permissive	0.615	870	0.225%	0.229%	3 (11.5%)

1 No noise filter was applied to GTV recall results.
2 The detection limit was estimated based on the SOC VAF distributions.

Shift artefacts were only frequently observed in QIASeq data sets and were negligible for other tagging procedures. In all cases, singleton clusters were reduced more than the overall reduction of cluster counts. The best overall cluster reduction and singleton cluster reduction was observed for QIASeq Seraseq data sets followed by smMIP. Also, the number of read pairs per UMI and the UMI coverage in case of moderate clustering was highest for QIASeq Seraseq data sets.

A comparison of the SOC-filtered Seraseq reanalysis results for different clustering settings and results obtained by standard analyses of the QIASeq and NEBNext Seraseq dilution series from the D & R Institute of Human Genetics are displayed in figures 3.26 and 3.27 respectively. For all clustering settings and QIASeq and NEBNext reanalyses, three single observation supported false positive calls were present at GTV sites in control data sets.

Table 3.11: corrected UMI artefacts and clustering measures.

Data Set	Advanced Clustering[1]	Mean UMI Coverage	Mean Read Pairs per UMI	Cluster Red.[2] Factor	Singleton Cluster Red. Factor	Random Base Errors	5'-Shift	3'-Shift
NEBNext 2% Seraseq	moderate	370.7	8.1	1.05	1.11	1,805	-	-
	permissive	359.5	8.4	1.09	1.17	2,270	36	0
NEBNext 0.125% Seraseq	moderate	418.6	9.7	1.03	1.05	1,124	-	-
	permissive	377.0	10.8	1.14	1.27	4,355	131	9
QIASeq 2% Seraseq	moderate	469.9	24.1	1.34	1.47	5,095	-	-
	permissive	252.8	47.7	2.25	2.81	7,022	1,298	0
QIASeq 0.125% Seraseq	moderate	375.1	25.0	1.34	1.45	3,840	-	-
	permissive	208.7	48.3	2.16	2.61	5,179	5,805	986
Thru-PLEX MiSeq	permissive	181.2	1.7	1.01	1.01	2,120	3	0
smMIP 100% TruQ4 replicate 2	permissive	141.3	9.1	1.62	1.72	514	11	3
NEBNext 100% TruQ4	permissive	78.1	16.9	1.08	1.15	4,761	112	4
QIASeq TruQ4	permissive	19.0	15.7	1.50	1.91	16,624	6,265	118

[1] No checks for shifted UMI sequences were carried out in moderate clustering.

[2] Cluster reduction factor

The QIASeq results from the D & R Institute of Human Genetics were produced with a data processing service offered by QIAGEN. The QIAGEN analysis results were validated and kindly provided by Sabrina Weber, MSc, from the D & R Institute of Human Genetics at the MUG. The NEBNext results were created with a standard in-house low VAF variant calling analysis pipeline using Mutect. Therefore, the NEBNext results lack comparability.

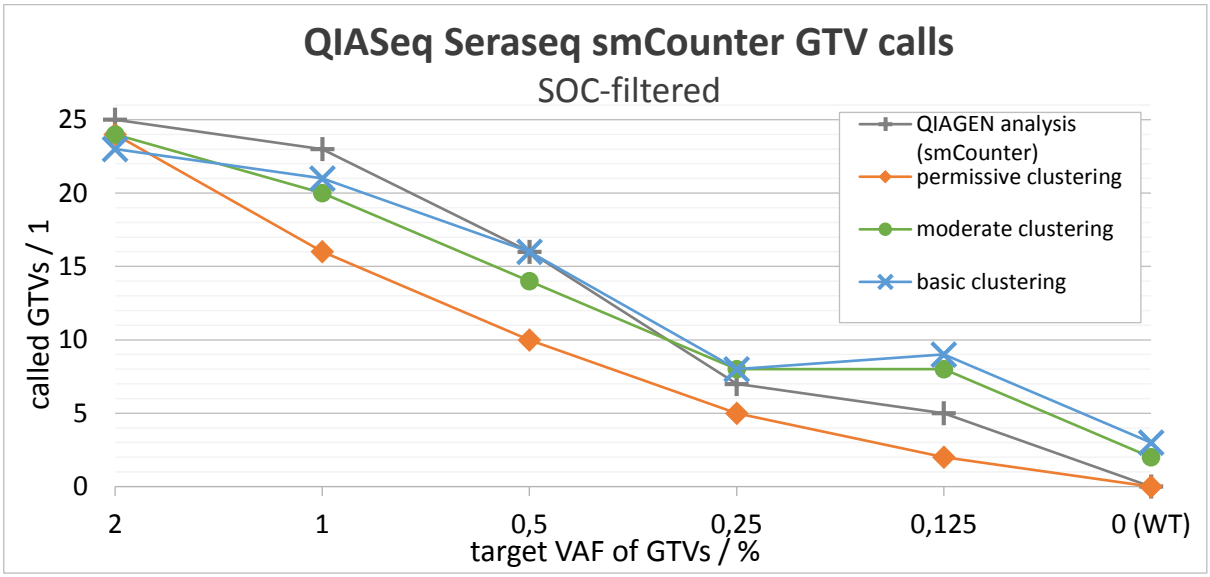


Figure 3.26: GTV call decay over GTV VAFs of the QIASeq Seraseq analyses. *Three advanced clustering settings were used. The grey data series is a QIAGEN analysis result using the smCounter software which was carried out for the D & R Institute of Human Genetics. Single observation GTV calls were removed from all results. Per dilution level results for QIASeq show a nearly linear GTV detection decline down to 0.25% VAF. Results of lower VAF show a contra intuitive increase in GTV calls. Control results have two and three false positive GTV calls for moderate and basic clustering settings respectively. Permissive clustering exhibits zero false positive GTV observations in the control but also an overall reduced capability to call GTVs except for the 2% VAF data set. The QIAGEN analysis has a better GTV recall in higher VAFs until 0.5% GTV VAF at which basic clustering of the reanalysis performs equally well. Results for the basic and moderate clustering setting exhibit slightly higher GTV detection performance than QIAGEN results below 0.5% VAF. Higher GTV recall obtained with the reanalysis for the 0.125% GTV VAF data set are questionable because of the higher false positive rates for basic and moderate clustering and the contra intuitive detection increase. The chart was created with Microsoft Excel.*

Based on these results, only moderate clustering settings can be recommended for QIASeq data sets in combination with the smCounter caller.

Details of VAF distributions of GTV calls and ambiguous multiple UMI-supported variant calls (MOCs) are displayed for NEBNext and QIASeq Seraseq data sets and for different clustering settings in figure 3.28 and figure 3.29 respectively. The SOC portion was removed due to the notion that SOCs are mainly due to uncleared base noise with no underlying true variant.

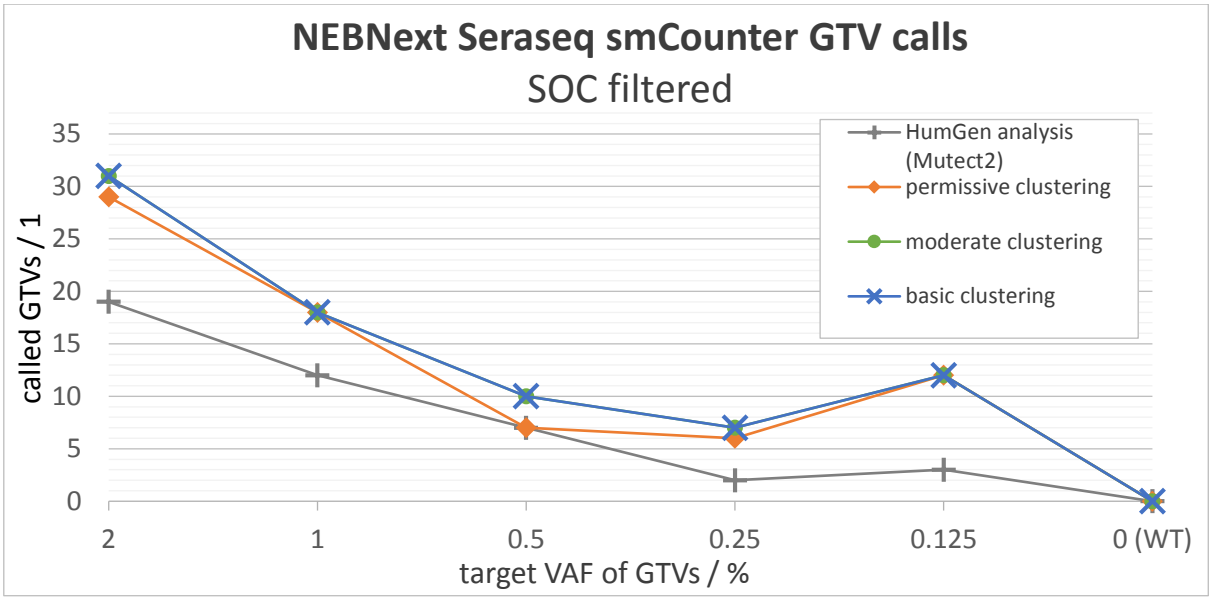


Figure 3.27: GTV call decay over GTV VAF of the NEBNext Seraseq analyses. *Three advanced clustering settings were used. The grey data series is an analysis result created at the D & R Institute of Human Genetics using the Mutect variant caller. This analysis was validated by an employee of the D & R Institute of Human Genetics. Single observation GTV calls were removed from all results. Results of the NEBNext Seraseq reanalysis show a better performance than the Mutect analysis of the D & R Institute of Human Genetics. Up to 12 additional GTVs could be detected by the reanalysis compared to the Mutect analysis. Basic and moderate clustering settings resulted in identical GTV call numbers. For permissive clustering, up to three GTVs are missing compared to other clustering settings. The chart was created with Microsoft Excel.*

For NEBNext, not only the lowest VAF of ambiguous calls was reduced by clustering (see 0.25% VAF data set results) but also calls of higher VAF were affected as evident from the reduced number of outliers in data sets of GTV VAFs below 1% especially in case of permissive clustering (see figure 3.28).

Results for QIASeq showed that ambiguous MOCs were dominated by dual UMI-supported calls and triple UMI-supported calls (see figure 3.29). The lower portion of ambiguous MOCs could be successfully reduced only by permissive clustering. In higher GTV VAF data sets, ambiguous calls of higher VAF persisted. Dual UMI-supported calls remained even after permissive clustering in the 0.25% GTV VAF data set, 0.125% GTV VAF data set, and the control data set.

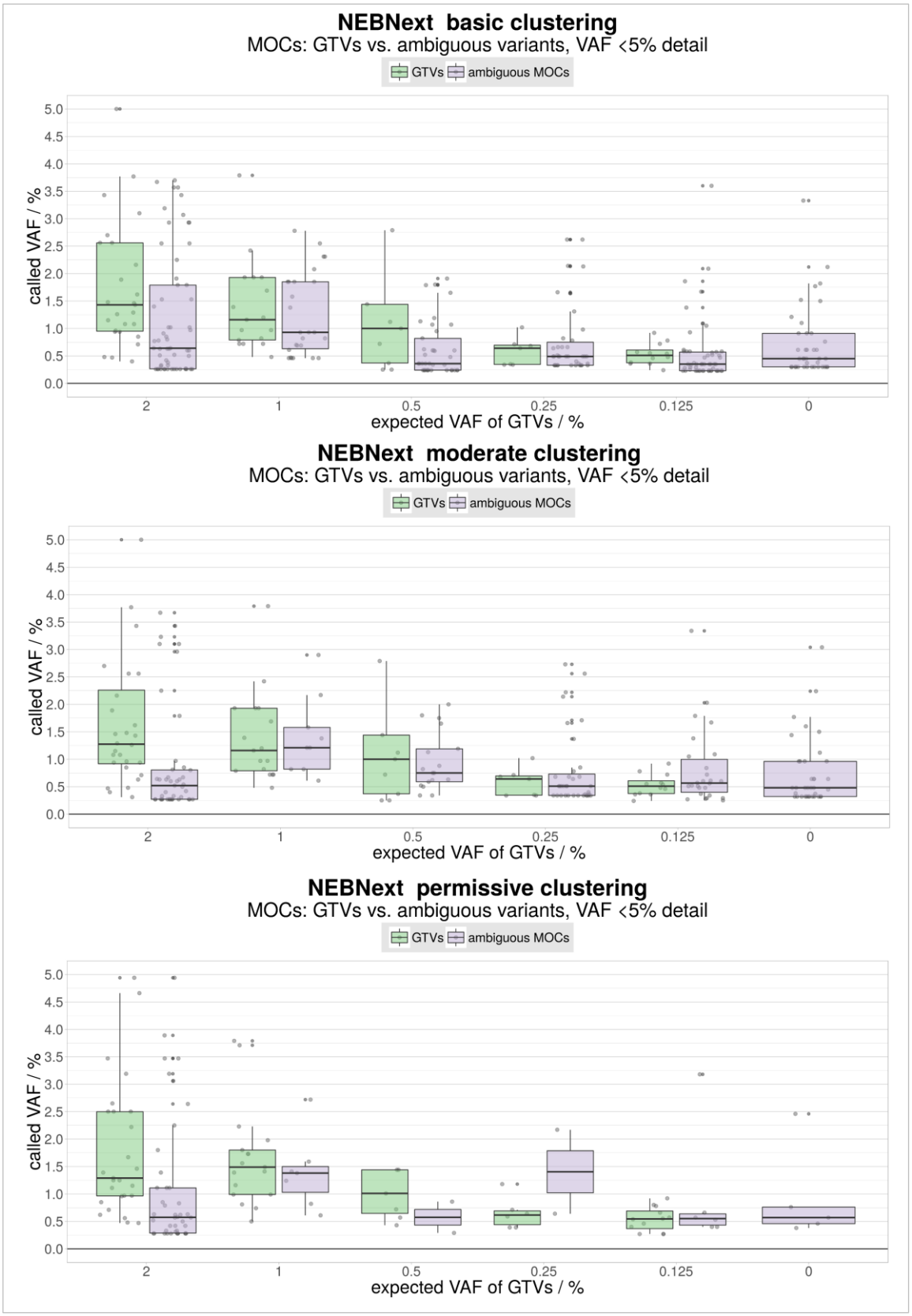


Figure 3.28: VAF distribution comparison of multiple observation GTV calls and ambiguous calls of the NEBNext Seraseq dilution series for different clustering settings. *Details below 5% VAF are shown. Boxplots and jittered dots represent identical data. Plots were created in* R *using the ggplot2 package.*

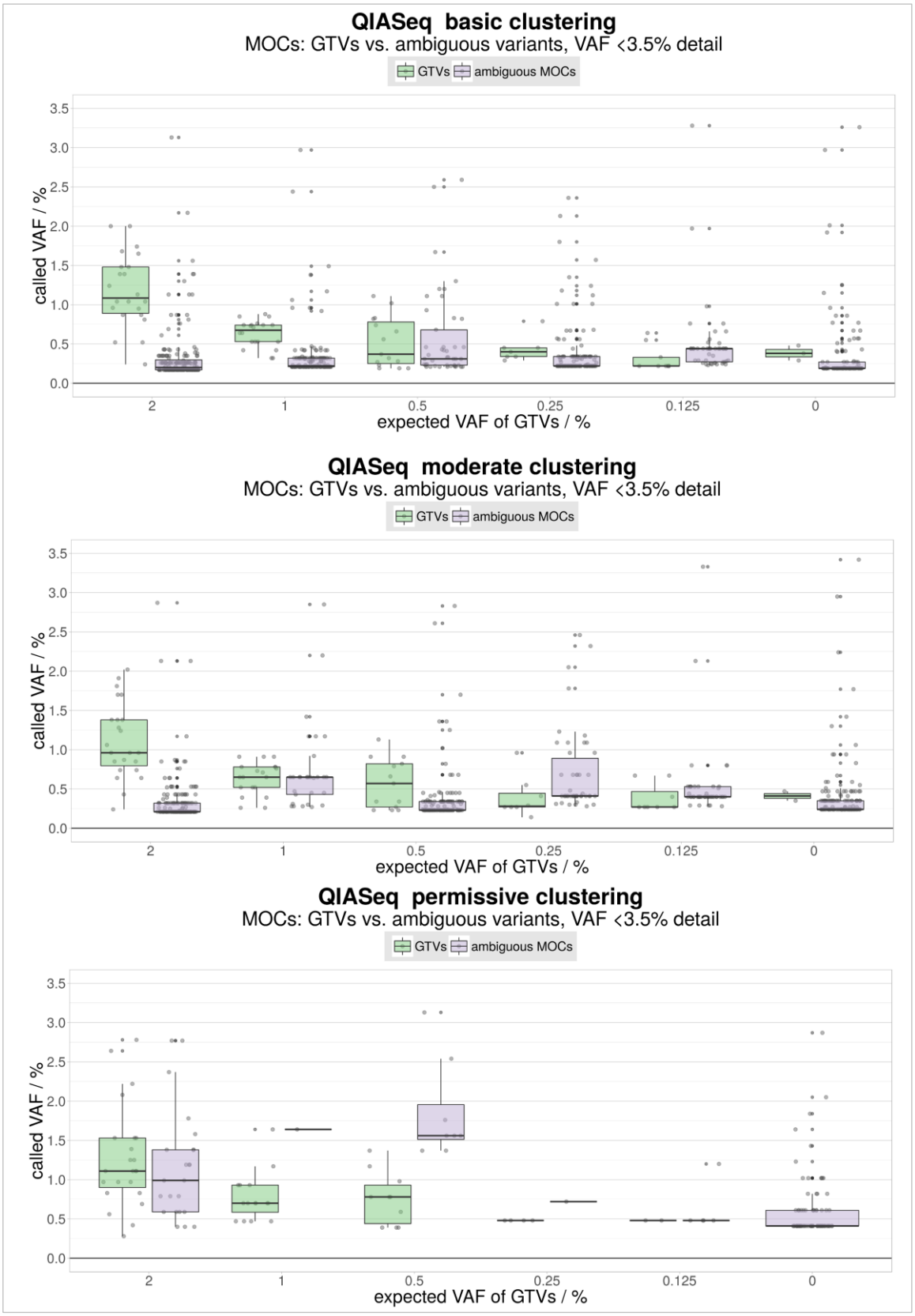


Figure 3.29: VAF distribution comparison of multiple observation GTV calls and ambiguous calls of the QIASeq Seraseq dilution series for different clustering settings. *Details below 3.5% VAF are shown. Boxplots and jittered dots represent identical data. Plots were created in R using the ggplot2 package.*

In total, the VAF variability of multiple observation GTV calls and ambiguous MOCs were higher for NEBNext than for QIASeq. QIASeq yielded more ambiguous calls than NEBNext. Permissive clustering was effective in reducing ambiguous calls while also exerting a disadvantageous effect on GTVs with VAFs below 2% especially for the QIASeq protocol.

False Positive Estimation

The equally distributed errors assumption would suggest that the abundance of pure noise dual UMI-supported variant calls (DOCs) can be determined by taking one third of SOC occurrences. An analysis of alignment counts supporting a variant call yielded a different result (figure 3.30). Moderate clustering resulted in DOCs being 35% to 45% as frequent as SOC calls. Permissive clustering resulted in the highest variability of the DOC to SOC ratio: DOCs were observed at 15% to 60% of the frequency of SOCs.

As expected, the majority of variant calls could be attributed to single or dual observation calls. In all cases, SOCs were the most frequent variant calls. Furthermore, the observed ratio of DOCs over SOCs was higher than would be expected from raw base noise based on the equally distributed base errors assumption in most cases. This is reasonable because of the employed template material amplification and subsequent *in silico* alignment reduction which alters the state of the sequencing library the equally distributed base errors assumption tries to describe.

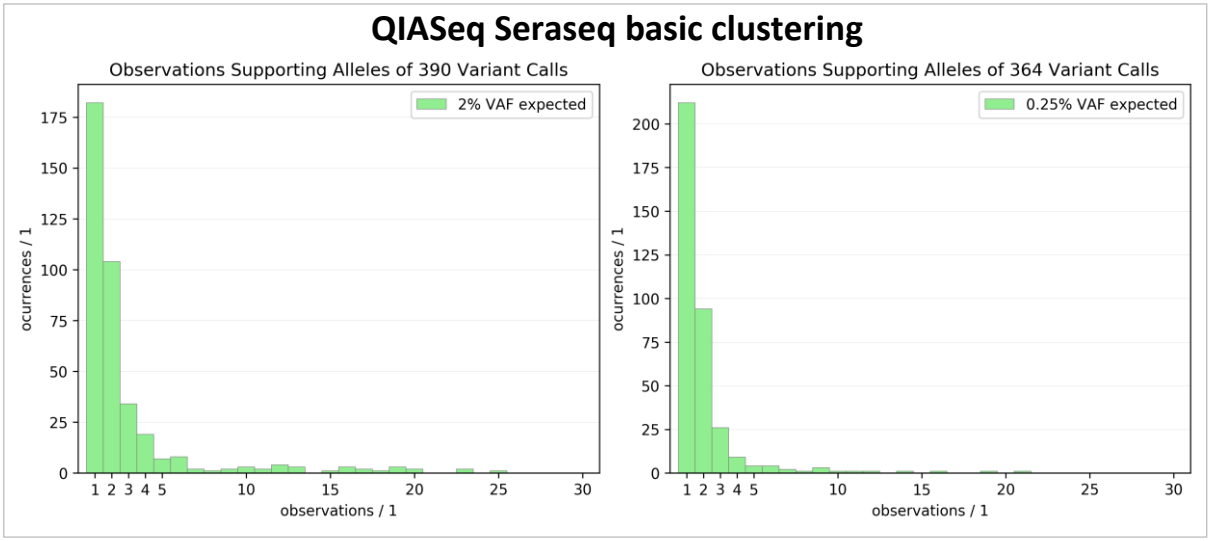


Figure 3.30: extremes for distributions of variant allele supporting observations for the QIASeq Seraseq dilution series with deactivated advanced clustering. *The number of DOCs was observed to be between 45% and 60% of SOCs contrasting the equally distributed errors assumption. Plots were created in Python with matplotlib.*

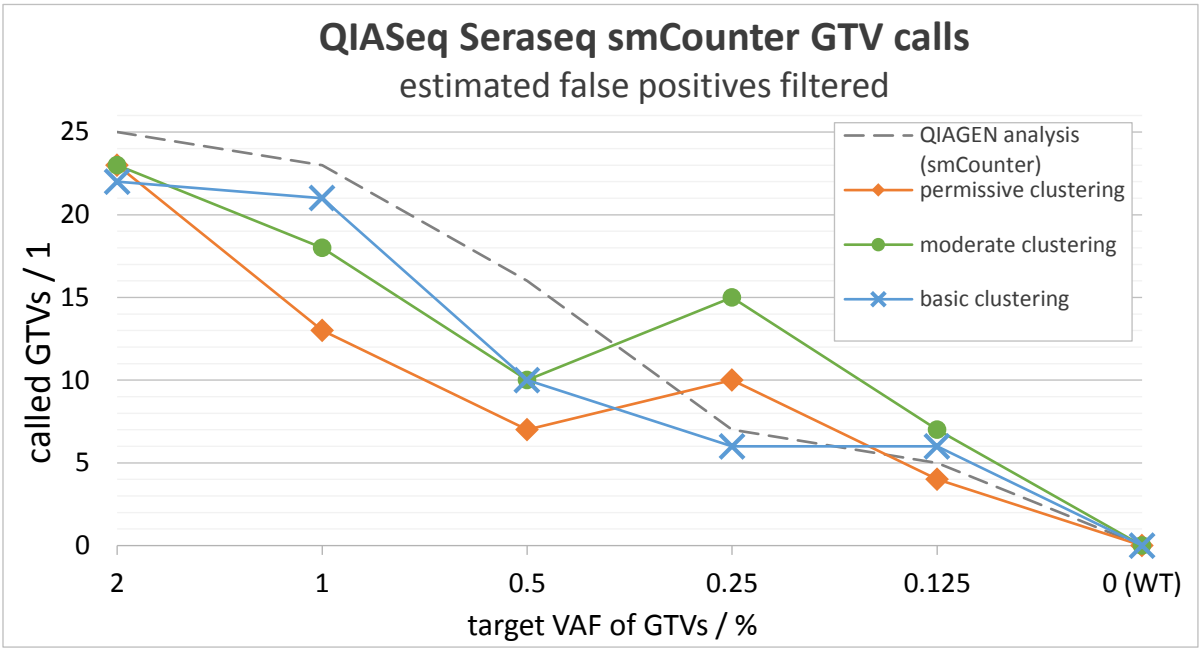


Figure 3.31: GTV call decay over GTV VAF of the QIASeq Seraseq analyses. *Three advanced clustering settings were used. The grey dashed data series depicts a QIAGEN analysis result created with the smCounter software for the D & R Institute of Human Genetics for which SOCs were removed. This analysis was validated by an employee of the D & R Institute of Human Genetics. The estimated false positives filter was used on results of different clustering settings. The chart was created with Microsoft Excel.*

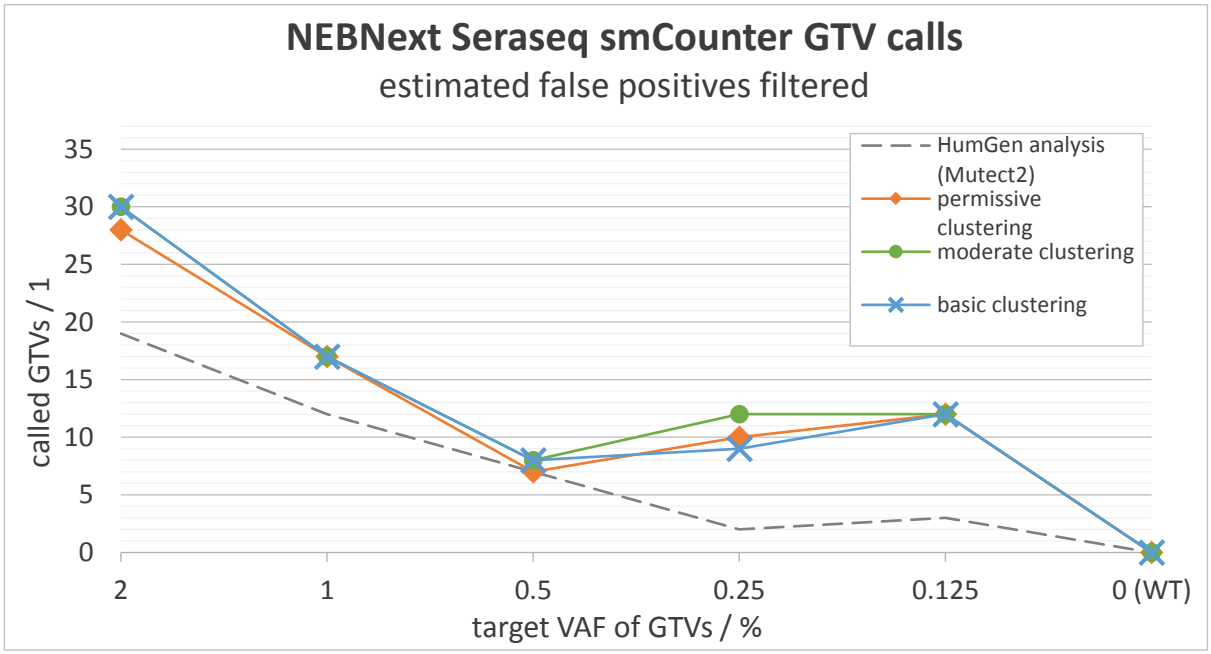


Figure 3.32: GTV call decay over GTV VAF of the NEBNext Seraseq analyses. *Three advanced clustering settings were used. The grey dashed data series is an analysis result created at the D & R Institute of Human Genetics using the Mutect variant caller. This analysis was validated by an employee of the D & R Institute of Human Genetics. The estimated false positives filter was used on reanalysis results. The chart was created with Microsoft Excel.*

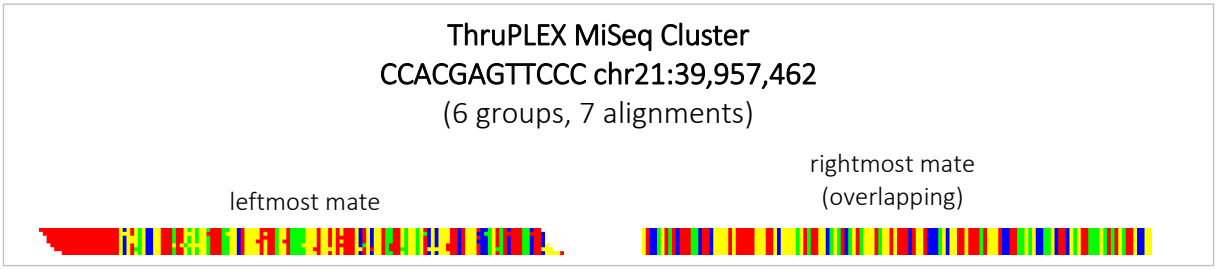


Figure 3.33: exemplary ThruPLEX cluster taken from the MiSeq analysis. *Second in pair reads formed leftmost alignments for this cluster. Mates overlap for the last 18 bases. The initial poly-T sequence led to 6 different mapping positions. Pair orientation was F1R2. The depicted mate clusters are outward facing. Alignment copy numbers are very low. Cluster representations were created in Python.*

Basic clustering was the only setting that showed monotonically decreasing GTV call numbers without a contra intuitive increase (figure 3.31). Therefore, the basic clustering result is viewed as the only trustworthy result. In most cases, the QIAGEN analysis yielded more GTV calls than the false positives filtered reanalysis results.

The same GTV correction tendency for different GTV VAF levels was observed for the NEBNext Seraseq reanalysis (compare figures 3.27 and 3.32). GTV recall differences between clustering settings were still small. The contra intuitive rise in GTV calls at low VAFs was smaller compared to the SOC filter results.

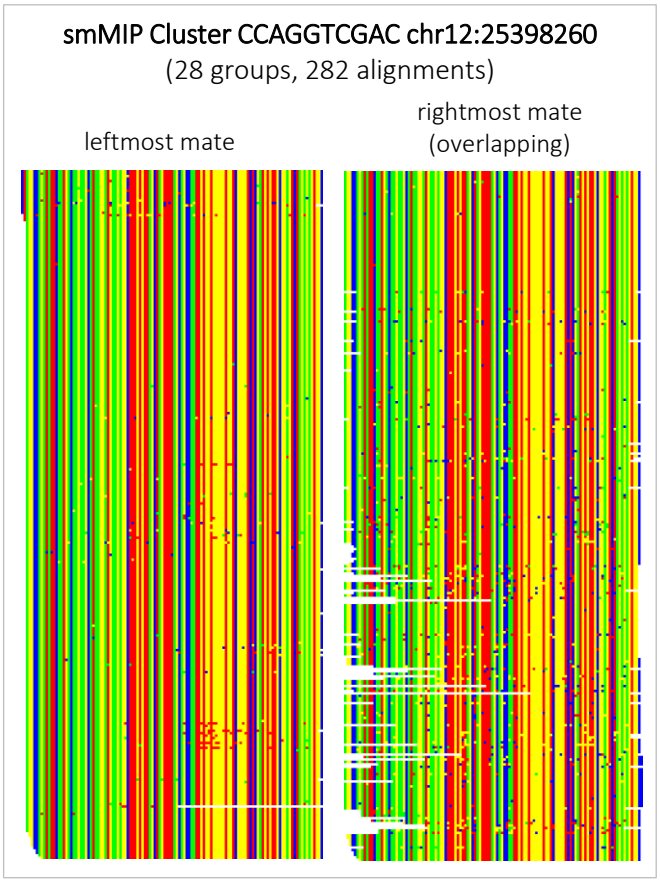


Figure 3.34: exemplary smMIP cluster. *Mates overlap. Rightmost mates are mostly R2 reads. Pair orientation is F1R2 Second in pair reads are lacking the full-length sequence more frequently than first in pair reads. Clusters of smMIP data sets exhibit high base noise. This is especially pronounced in R2 sub-clusters. Cluster representations were created in Python.*

Cluster noise was visualized by created insertion and soft clip-free consensus matrices. Examples are displayed in figures 3.33, 3.34, 3.36, and 3.35 for ThruPLEX, smMIP, QIASeq, and NEBNext respectively. Base colouring was: A=yellow, T=red, C=green, G=blue, deletion=grey, N=black.

Horizontal accumulations of alternative bases indicate high-frequency base error template molecules. Mate sub-clusters which frequently exhibited a higher error rate and an increased

frequency of prematurely terminated sequencing reactions were usually formed by second in pair mates.

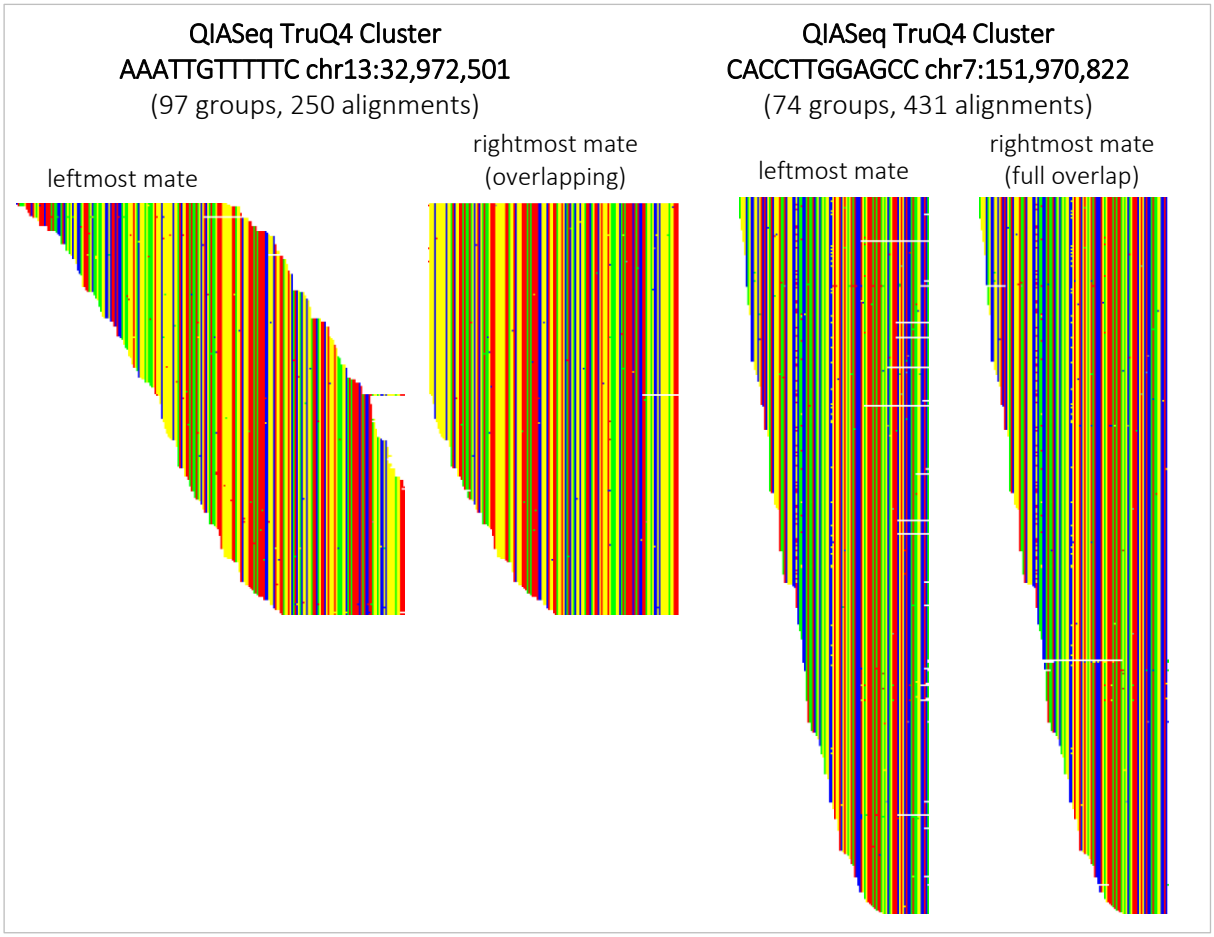


Figure 3.36: exemplary clusters for the QIASeq TruQ4 tagging protocol. *For the cluster on chromosome 13, leftmost alignments are second in pair alignments. Pair orientation is F2R1. The mate sub-cluster overlap ranges from approximately 50% to full overlap (bottom alignments). The cluster on chromosome 7 has fully overlapping mate sub-clusters. A template molecule length decay is visible for the QIASeq clusters. Cluster representations were created in Python.*

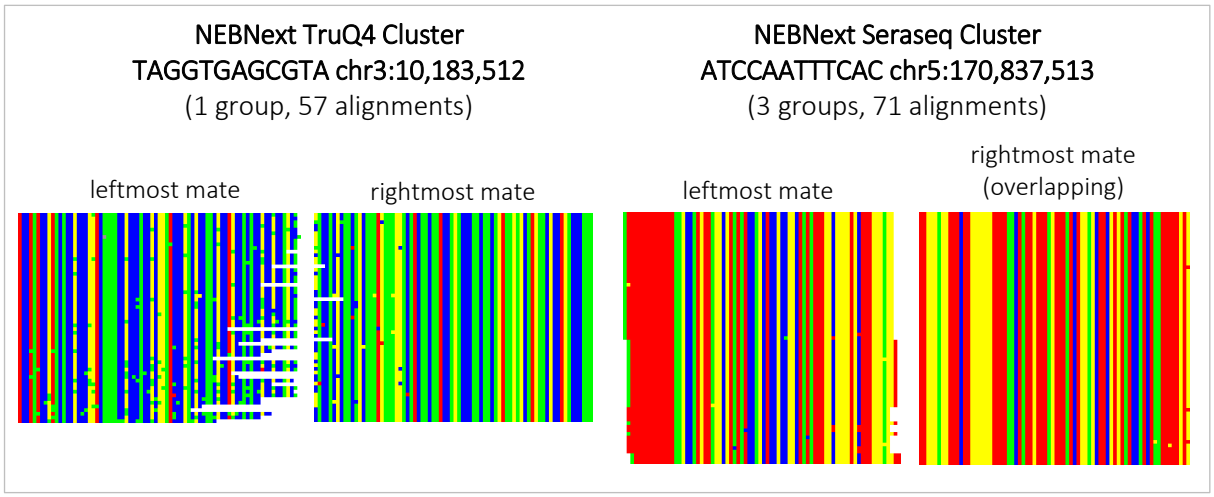


Figure 3.35: exemplary clusters for NEBNext TruQ4 and Seraseq dilution series. *Leftmost in pair alignments are formed by second in pair reads. Pair orientation is F2R1 for both clusters. The cluster on chromosome 5 shows small leftmost mapping position deviations due to the poly-T sequence. Cluster representations were created in Python.*

4. Discussion

4.1. Data Quality

The sequencing data used in this thesis were of sufficient quality for variant calling validation experiments as FastQC quality checks showed. Variability of per base and per tile sequence quality may be attributed to high loading conditions. Another source of variability might be the limited diversity of UMI-tagged libraries because of extensive amplification. This was especially pronounced in smMIP data sets. Per base sequence quality, per sequence quality scores, and per tile sequence quality give a rough picture of the loading situation of the flow cell. High loading conditions manifested as light blue to red tiles in the per-tile sequence quality map. Furthermore, the per base sequence quality was found to decay stronger towards read ends in these cases. This circumstance partially justifies read clipping in sequence duplication statistics calculation by FastQC to obtain a better estimate of the true duplication situation.

The quality drop observed for all QIASeq Seraseq R2 read data in conjunction with low sequence quality per tile indicates a transient, locally confined problem on the flow cell during the sequencing run. As evident from figure 3.1 panel B, the low-quality tiles were located on one edge and both sides of the flow cell. This supports the notion that either bubbles going through the flow cell or debris inside the flow cell lane may have caused the severe quality drop in R2 reads [48]. UMIs were extracted from R1 FASTQ files for QIASeq data sets. Therefore, the steep initial quality decline in Seraseq R2 data did not affect sequencing of UMIs.

Although the transient problem introduced a multitude of sequencing errors in R2 data, validation results for the dilution series confirmed the robustness of the analysis. An increase of variant calls due to sequencing errors is expected for the QIASeq Seraseq dilution series though. The GC-content distribution differences between R1 and R2 data for QIASeq Seraseq data sets (*i.e.* R2 distributions resemble smoothed versions of R1 distributions) indicate that the transient problem did introduce a considerable amount of random-base sequencing errors (figure 3.3). Therefore, the QIASeq tagging procedure might be able to produce less noisy results in case of normal, complication-free sequencing runs.

Notable contamination with Illumina universal adapter was only detected in the NEBNext 10 ng data set of approximately 5%. The per sequence GC-content showed two modes. One mode most likely originated from the adapter contamination.

Inferring quality statistics of the sequencing library's original state from statistics provided by FastQC (*e.g.* GC-content, overrepresented sequences) may be misleading to even non-informative, because of the excessive amplification, limited library diversity, and amplicon structures. The

categories involving sequence duplication levels yielded only a rough estimate of the amplification situation in the data set because input files were down-sampled to the top 100k sequences for respective modules by FastQC. Moreover, sequences longer than 75 bases were clipped to 50 bases which removed up to two thirds of the information present in QIASeq, ThruPLEX, and smMIP reads. Highly different sequence duplication levels for R1 and R2 data are thought to have resulted from capture strategies which utilized a single primer for region targeting. Whether DNA molecules are technical duplicates or biological ones, reads starting at the primer terminus are highly similar in either case. Thus, reported duplication level of mates containing a sequence which was targeted by a primer were highly inflated. On the other template end, length altering replication errors may decrease sequence duplication values issued by FastQC. These were predominantly present in QIASeq data sets. Comparing sequence duplication level distributions of R1 and R2 read data created by FastQC for NEBNext and QIASeq Seraseq data sets, QIASeq R1 and R2 data duplication levels were clearly different from each other while R1 and R2 sequence duplication level distributions overlapped in the 50 up to 5k region for NEBNext. This supports the idea that sequence length altering replication errors might affect the sequence duplication estimation by FastQC.

Differences of reported read sequence and UMI sequence duplication levels also show the limited usability of the FastQC sequence duplication values for amplified UMI libraries. While inherently shorter UMI sequences (which are normally positioned at the sequencing start of the tagged DNA molecule) accumulate fewer errors during PCR, estimated read sequence duplication is affected by bias introduced by the utilized capturing approach in addition to the higher number of accumulated errors. Therefore, UMI sequence duplication values were regarded as most trustworthy for UMI data.

Comparing library diversities resulting from 10 ng and 100 ng DNA input material of NEBNext TruQ4 experiments, a reduced R1 and R2 sequence diversity was well visible in the per-base sequence duplication plot for the 10 ng input material sample (see figure 3.7). UMI sequences of the NEBNext wildtype data set are an extreme example of sequence similarity. Based on the colours, the UMI sequence 'AGCGATAG' can be read from the plot for this data set. UMI sequences of all non-wildtype datasets showed a higher guanine proportion of approximately 40% compared to read sequences which resulted in a darker grey tone.

Another quality check module with limited usability for UMI data was the overrepresented sequences category. This category tended to be filled with strongly amplified sequences since a fraction of one tenth of a percent of the whole data set suffices for a sequence to be reported. In

case of low contamination levels or high amplification levels, this may prevent sequences of adapter contaminants to appear in the overrepresented sequences category.

For targeted amplicon sequencing, the per sequence GC-content plots give an idea of the impact of sequence errors on the data set as the example of QIASeq Seraseq dilution series showed. GC-content distributions were smooth and less extreme for R2 data which contained a high amount of sequencing errors compared to R1 data (figure 3.1, A).

4.2. Impurities and Contaminations

In general, no contaminations originating from non-human organisms could be identified for any of the data sets. Screening results like obtained for NEBNext Seraseq data sets with a tolerable part of non-aligning reads below 3.5% were within expected limits. Aside from identified adapter contaminations in the NEBNext TruQ4 data sets, the majority of reads not aligning against the human reference genome most likely stemmed from polymerase errors introduced during amplification combined with sequencing errors. Screening results showed more R2 reads not mapping to the human genome than R1 reads which is in accordance with lower R2 quality compared to R1. In this thesis, this circumstance was commonly observed for Illumina paired-end sequencing data and is most likely caused by the lower sequencing accuracy after bridging the forward read for sequencing of the reverse sequence (R2). Lower R2 quality is therefore not due to some sort of contamination or impurity.

In case of smMIP data sets, the amount of reads not aligning against the human genome was unusually high. This could indicate a systematic contamination of all smMIP samples or purification associated problems that occurred upstream to the tagging step (most likely molecular inversion probe synthesis) and resulted in contaminated tagging reagent. The possibility of severe sequence degradation might have also resulted in a considerable amount of reads not aligning against the human reference genome. Since no high-quality hits with high query coverage were found for non-aligning reads in the BLAST search, the possibility of systematic contamination is considered to be highly unlikely.

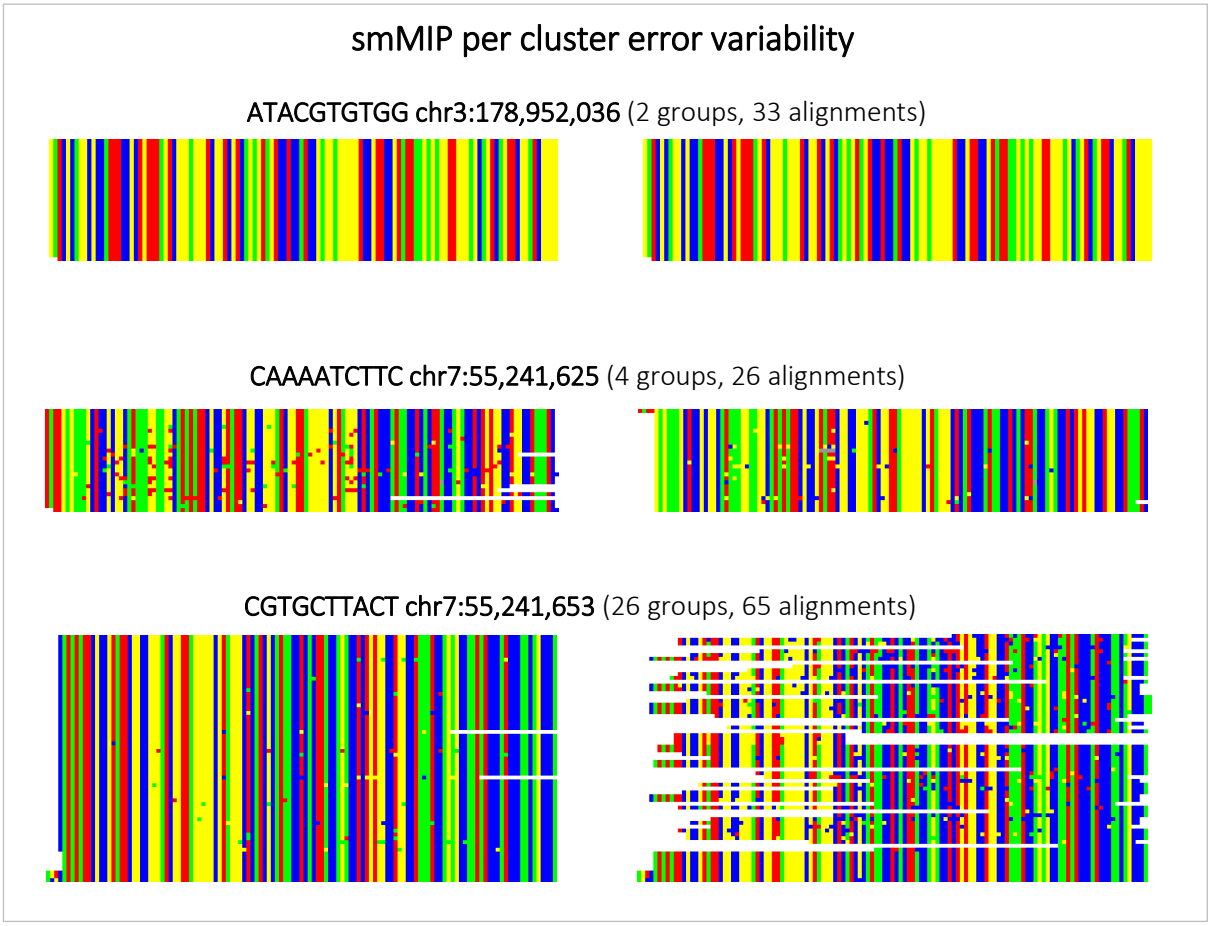


Figure 4.1: varying error levels of smMIP clusters. *Leftmost sub-clusters are positioned on the left, rightmost on the right. Sub-clusters overlap for all cases. Base errors and strand termination artefacts are present at different frequencies in different cluster. Error frequencies and artefact frequencies inside clusters also vary. Plots were created in Python.*

The relatively high base error rate in smMIP clusters supports the notion that the decrease in reads aligning to the human genome originated from accumulated base and sequencing errors (see figure 4.1). This observation of suboptimal read quality and/or purification difficulties alone can be used to advise against the use of in-house produced tagging material. From an economical point of view, sequencing a substantial number of templates which are unusable in downstream analysis would render the whole procedure cost-ineffective compared to experiments utilizing other tagging protocols. In the best case where observed low quality would result from non-systematic, transiently occurring interference or human error, the varying quality would render the whole diagnostic procedure unreliable and, thus, unfavourable. The clinical usability would not be given because of inconsistent results after repeated testing.

Aside from polymerase and sequencing error induced reduction of reads aligning against the human genome, ending up with approximately 1% of unmappable reads from human plasma samples (*e.g.* NEBNext TruQ4 dilution series) were considered normal because of cfDNA from human viruses and other micro-organisms contributing to cfDNA through various mechanisms [88, p. 1].

Nevertheless, the origins of only a minor portion of non-aligning reads could be identified via BLAST search. Identified cloning vectors and expression vectors are expected to be remnants from reagent or reference DNA material production.

4.3. Profiling

After the exploratory analysis, it was concluded that the UMI-tools software should not be integrated into the final analysis pipeline without a suitable countermeasure that limits execution time of the UMI error correction step. The problem of high UMI-tools run times when dealing with deep sequenced amplicon regions in particular was successfully circumvented by subsampling alignment groups. This countermeasure also sped up subsequent computations like variant calling.

Disregarding profiling results of reanalyses which were clearly affected by thread suspending, it can be stated that reanalyses took around three times longer than corresponding minimal exploratory analyses. This was mostly due to the thorough read and alignment processing which was carried out mainly using Python code. The clustering task was especially time consuming. Nevertheless, the overall better variant calling performance of the reanalysis compared to the exploratory analysis justifies longer analysis durations at least to some extent. While most execution times (*e.g.* NEBNext TruQ4 dilution series) were acceptable, a more performant code would be preferable still. Redesigning the pipeline towards parallelized and contig-wise data processing would supposedly shorten the overall analysis duration and reduce main memory requirements (currently about 300 GB) by a factor of 4 to 8 depending on the capabilities of the executing machine.

4.4. Shortcomings of the Exploratory Analysis

The enforced grouping of reads entirely based on UMI sequences deviated from the clustering method used in the smCounter paper. This substantial simplification was thought to be acceptable for a minimal approach at first. The observed overall poor to at most moderate GTV recall for Mutect and smCounter led to a different view on read grouping. One cause of the poor variant calling performance was the compromise of N-masking bases at which no clear consensus base could be found (*e.g.* two bases were observed equally often). This was thought to be valid because of the supposedly rare occurrence of this case. Nevertheless, this posed a problem in groups formed only by two reads which occurred more frequently than expected (see figures 3.8 and 3.10]). In the subgroup of consensus reads computed from groups of two, base substitution errors introduced undefined bases at locations potentially carrying a variant signal. Indel errors, which are known to be rare for Illumina platforms, had a more severe impact on consensus formation. They

led to N-masking of large parts of the consensus sequences, starting at the position of the individual indel and ranging to the last overlapping position between read sequences of the group. These two effects led to removal of base information due to somewhat frequent single base errors and supposedly true SNVs and less frequent indel errors.

Furthermore, the naïve read grouping approach utilized in the exploratory analysis turned out to be not suitable for tagging procedures that produce non-trivial amplicon structures. The negative impact was particularly pronounced for data sets like QIASeq TruQ4 where mapping positions of grouped reads were frequently distributed over a short stretch of reference genome positions. Errors introduced by inaccurate alignment annotation from the highly error-prone consensus analysis branch of the exploratory analysis to the non-consensus branch posed a systematic problem. This had severe effects on variant calling with smCounter which relied on point-accurate mapping position annotations.

Extensive consensus sequence alteration by erroneously assuming identical starting positions of grouped reads led to nonsense consensus sequences which often turned out not to align against the human genome anymore. This was one of the major effects negatively influencing the consensus analysis branch and, thus, Mutect variant calling performance.

Mostly due to the distortions induced by the naïve grouping approach, Mutect still performed better than smCounter. This indicates that errors introduced by erroneous alignment annotation were more severe than the influence of N-masked bases and nonsense consensus sequences combined.

In summary, exploratory analysis results obtained with the naïve approach which also did not make use of UMI error correction and proper identity clustering were unsatisfactory in terms of GTV recall, detection limit, and VAF accuracy. Due to *in silico*-induced distortions which disrupted smCounter variant calling, NEBNext data sets in combination with the Mutect caller performed best in the exploratory analysis. This indicates that the NEBNext protocol produced a negligible number of in-group mapping position deviations and had the smallest amount of groups of size two. This is concordant with reported UMI artefact corrections in the reanalysis (table 3.11) and the group size distribution (figure 3.8).

4.5. UMI Group Size Distributions

In theory, a sequence abundance plot like the read group rank size plots created in this thesis would show a constant duplicate count across all groups for perfect PCR amplification. Deviations from this ideal plateau shape are thought to be caused by polymerase errors inside the UMI sequence,

PCR stochasticity and GC amplification bias. The latter would result in reduced amplification of GC-rich templates. PCR stochasticity during the first few cycles of amplification and GC bias to a lesser extent were shown to result in a shoulder-like region where sequence abundance drops steeply. Polymerase errors create new sequences which are predominantly abundant at low copy numbers after amplification. These errors result in pronounced tail regions [87, p. 14].

The observed pronounced tail regions in all read group and alignment cluster size rank plots indicate PCR errors being the primary source of UMI errors, especially in smMIP and ThruPLEX data sets where no shoulder regions were observed. Even after UMI error correction and permissive alignment clustering, the majority of clusters remained to be of size 1 or 2. The persistence of tail regions in all cases illustrates the limitations of applied error correction approaches and the utilized extended UMI artefact model.

After comparing the total amount of bases inside targeted regions for smMIP and QIASeq TruQ4 data sets, it is evident, that a much smaller number of read groups should have been observed for smMIP than for QIASeq. Although the ratio of QIASeq over smMIP read pairs was 2, QIASeq reads covered approximately 2,000-times more bases. This means that the expected read group diversity should be much higher for QIASeq data sets than for smMIP because of the approximately equal amount of input reference DNA material used for both tagging procedures. The observed ratio of UMI group numbers of QIASeq over smMIP seems to be close to the ratio for read pairs which might be due to a predominantly observed UMI group size of 1. Moreover, the highest observed amplification for smMIP UMI groups was approximately 6-times higher than the maximum observed amplification for QIASeq UMI groups which also indicates that the theoretical ideal plateau should be higher and shorter for smMIP than for QIASeq. These considerations in addition to the observed variability in cluster sequence quality as already discussed in subsection 4.2 lead to the conclusion that smMIP results were strongly affected by a relatively high UMI error rate. This renders the use of in-house synthesized smMIP material barely usable.

The amount of sequence amplification observed for the ThruPLEX MiSeq data set fell short of expectations. Low achieved duplication levels of at most 40 duplicates quickly dropped to 3 duplicates after the first ninth of ranked UMI groups. This renders the tagging procedure almost unusable for the pursued noise reduction approach of finding consensus bases across duplicates based on the majority rule. Due to the similar sequence abundance shapes of smMIP and ThruPLEX, usability considerations based on the dominant tail region for smMIP also apply to ThruPLEX

4.6. Extended UMI Artefact Model and Corrective Measures

Ideally, PCR amplification should yield groups with identical template length and highly similar UMI sequences which would be easy to reduce. Also, UMIs found at any given position of different UMI groups should be dissimilar, *i.e.* two UMIs should have intermediate to high Hamming distance. Nevertheless, high UMI similarity contradicting UMI space occupation and in-group template length and leftmost mapping position variability were frequently observed in the data. Hence, the leftmost mapping position-based single base substitution error correction within UMI sequences by UMI-tools had proven to be insufficient in the exploratory analysis.

Based on this insight, an extended error model was developed for the reanalysis pipeline to correct for a larger variety of errors that might affect the alignment reduction. This should lead to a more accurate representation of the original state of the amplified library. Though being an extension to the model used by the UMI-tools software, the extended model was kept simple and convenient. The extended error model relies on three assumptions concerning PCR and sequencing errors in 5'-regions of alignments which interfere with UMI grouping and subsequent alignment reduction:

1. Base substitution errors occur
2. Small length altering errors occur
3. Errors are infrequent

If error sources other than substitutions are disregarded, single substitution error correction as performed by UMI-tools would be sufficient for obtaining the original state of the amplified library. Since the UMI-tools correction was found to also reduce the number of called GTVs in the exploratory analysis, an additional template length criterion was implemented to improve distinguishing supposedly true UMI errors from similarities occurring by chance. The size of the standard deviation-based window was limited to a maximum of nine bases and a minimum size of three bases around the mean mate mapping position.

For this criterion to work properly, only certain read pair orientations are allowed. In Illumina paired-end sequencing, one read should map to the plus strand in 5' → 3' orientation while the other read of the pair should map to the minus strand in the same orientation relative to the minus strand. The reason for this is the elongation of synthesized DNA towards its 3' end. In most cases, paired-end reads should therefore 'point' towards each other, a characteristic which is also called 'inward facing'. Possible causes for an alignment pair to exhibit an unexpected orientation are repetitive sequences, assembly flaws (which can be mostly ruled out for the current versions of the human reference genome), and reads mapping to palindromic regions or inversions. The latter may occur in tumour cells as a result of extensive mutation or in healthy individuals as structural germ line variants. In all other occasions where paired-end reads are found inward facing, the template

length similarity criterion was expected to work. This was observed to be the case for the clear majority of aligned read pairs.

On top of single base UMI errors, infrequent small indel errors predominantly occurring in homopolymer stretches were expected to cause only short five-prime mapping position deviations. This depends on terminal sequences of the duplicated template molecule and the base diversity inside the UMI sequence. Identity clustering was employed as corrective measure for all artefacts that alter template length but do not induce UMI sequence shift artefacts.

Indels may also occur in or adjacent to UMI sequences which would cause a part of or the whole affected UMI sequence to be shifted relative to its original location. These errors may occur due to the assumption that had to be made in data processing: the actual UMI sequence is always located at the same position inside reads. In theory, shift errors could also occur due to ligation of a degraded sequencing primer to the tagged template molecule or because of degeneration of the initial spacer or the UMI sequence itself (in absence of a spacer). The first type of error is thought to occur rarely since sequencing primers are highly stable and - if at all - tend to break randomly in case they had undergone too many cycles of freezing and unfreezing which is expected to prevent hybridization. The second type of error might occur due to insertion errors either occurring in the initial spacer, or at the start of the UMI sequence. For example, *Taq* polymerase exhibits a 3'-end adenylation activity. Using this polymerase in PCR might alter synthesized molecules accordingly. To correct for shift artefacts that affect the entire UMI sequence, advanced clustering was employed.

Another artefact in the context of altering synthesized molecule length is premature synthesis termination. Termination events are thought to happen more frequently in later rounds of replication because of reaction conditions changing towards inadequate concentrations of deoxy nucleotide triphosphates (dNTPs). This may be caused by unequal base distribution of the template material which also leads to increasing amplification bias with every PCR cycle. Termination may happen during replication of the forward or backward strand which would cause synthesized molecules to shrink from both ends compared to the original template. In case of 5'-end located UMI sequences and reverse strand termination, the bioinformatical UMI extraction frame would be shifted according to the number of missing terminal bases in the reverse strand and propagated to the forward strand after another PCR cycle. This would result in artificial UMIs with likely high Hamming distance and small edit distance up to completely new UMIs depending on the number of missing bases. Besides PCR and sequencing errors, minimal contamination of the sample with nucleases may cause non-uniform length degradation across all templates.

For tagging protocols like NEBNext, which showed high stability of template molecules inside clusters, premature synthesis termination was thought to induce only a negligible number of artefacts. Nuclease contamination were ruled out because of the high standards for wet lab work and quality assurance that are in place at the D & R Institute of Human Genetics. Furthermore, no nuclease contamination was reported for any of the conducted experiments.

Template switching errors result in chimera formation and are not regarded by the extended error model, since they are resolved in the mapping step which finds the best alignment for one of the subsequences of the chimera. The rest of the chimeric sequence is hard-clipped and, thus, omitted. Therefore, chimera formation acts as an information loss increasing factor for both origin molecules contributing to a chimeric molecule. Furthermore, the created artificial molecule ultimately leads to a new alignment cluster. Since shorter sequences may be mapping ambiguously which reduces the quality of the alignment, the minimum quality criterion of the alignment filter may remove an extensively hard-clipped alignment.

The infrequent errors assumption dictates that simultaneous occurrences of base errors, indel errors, and template length altering errors would be extremely rare during synthesis of a single UMI-relevant region of a template molecule. Following the assumptions made in the extended error model, a whole history of errors occurring throughout multiple replication cycles of the origin molecule and its copies should be easily identifiable using the specificity-increased UMI-tools approach in combination with clustering. The resulting UMI group size rank plot would show a distribution which could almost entirely be characterized by a plateau region. Only a small initial peak, a much less pronounced shoulder region and a tail region of negligible length should be visible after successful alignment reduction.

4.7. UMI Error Correction

The necessity of accounting for UMI errors was recognized after investigating the UMI group sizes distribution and the observed Hamming distance of UMI sequences per position. To achieve the expected distribution calculated by UMI-tools would mean that the utilized error correction approach successfully described and corrected for all UMI sequence errors. Moreover, the computation of the ideal edit distance distribution would have been based on the correct assumptions in such a case.

Especially the percentile and the adjacency approach to a lesser extent were inferior to the directional-adjacency and cluster approaches as evident from the results displayed in figure 3.13. Nevertheless, all approaches failed to entirely remove UMI errors. All approaches provided by

UMI-tools showed a second mode at the lower end of the edit distance scale which deviated greatly from the supposedly ideal distribution. Thus, the error model employed by UMI-tools regarding the underlying mechanisms of UMI sequence alteration was regarded as insufficient for obtaining the original library state.

Though the variant call reduction after applying UMI-tools was remarkable compared to the old approach which did not implement UMI error correction, the decrease in GTV recall was not satisfactory. The rigorous UMI error correction by UMI-tools may have resulted in an over correction of UMI sequence similarity due to supposedly missing specificity in determining sequence similarity. The directional-adjacency approach only considered UMI group characteristics involving the UMI sequence and the group size but did not include any measures describing alignment pairs inside the UMI group.

Obtaining unsatisfactory results after including the UMI-tools error correction into the exploratory analysis was realized as an opportunity to reconsider the range of possible UMI sequence artefacts for correction. Specificity increase by considering the mate mapping position had only a minor effect. About 2% of regrouping suggestions created by UMI-tools were rejected for each data set. Another filter criterion checking for similarities of indel positions based on CIGAR strings of absorbing and absorbed UMI group was tested. This criterion did not enhance UMI error correction specificity and, thus, was omitted.

UMI artefacts defined by the extended error model and corrected by clustering had a far greater impact. Small leftmost mapping position deviations were frequently observed for data sets created with the QIASeq protocol. However, the tail regions of UMI group size rank plots were still dominant after clustering for most analysed data sets. This indicates that either the extended error model failed to describe other major error sources, or that the employed error correction approaches were insufficient in reducing substitution errors in UMI sequences. The latter case is deemed more likely because of the causes of tail regions in sequence abundance plots that were described by Kebschull and Zador [87, p. 11].

As an alternative to UMI-tools, a software called UMI-reducer was available for carrying out UMI network-based deduplication. The authors defined duplicates as reads sharing the same UMI sequence and mapping position. This definition was already obtained to be insufficient for error correction with UMI-tools. Also, because an implementation of substitution error correction was missing in this software [89, p. 3f], and the fact that duplicate removal would rule out consensus formation, this software was not investigated further for variant calling validation.

4.8. Variant Caller Performance

The Mutect and the smCounter variant caller were compared based on the performance measures: GTV recall, VAF calling accuracy, detection limit, portion of tumour-classifiable variants, proportion of untraceable background calls, total amount of variant calls, and the fraction of unclassifiable variants.

For variant calling validation, the portion of non-GTV calls was regarded as an ambiguous mixture of calls solely originating from uncleared molecular noise, and in case of the TruQ4 dilution series, heterozygous/homozygous germline variants from the wildtype control material. These ambiguous variants constituted the larger portion of calls for all data sets.

For the discussion of GTV recall, only the results from the reanalysis shall be considered due to the systematic error in the alignment annotation of the exploratory analysis which led to extremely low numbers GTV detections for smCounter.

As evident from GTV recall results (table 3.8), smCounter showed superior performance compared to Mutect for all tagging procedures. The differences were highest for smMIP and lowest for NEBNext.

In terms of VAF accuracy, which was assessed mainly based on GTV calls, smCounter performed better than Mutect, especially for low VAF calls. This is evident from GTV VAF deviation plots (see figure 3.23).

The proportion of background variant calls (figure 3.21) indicates a higher level of uncleared noise for smCounter than for Mutect. Moreover, the majority of variants were called below 5% VAF for the NEBNext Seraseq data set which also indicates that smCounter might be prone to calling variants on uncleared noise. This points towards the necessity of pairing sensitive callers like smCounter with effective variant call filters. However, the number of traceable variants was stable across data sets of the Seraseq dilution series indicating valid calls. Error-prone clusters were omitted for the Mutect analysis which likely reduced the portion of untraceable background calls for this dilution series. Therefore, the comparison of these calling results could be misleading (figure 3.21).

Another aspect that needs to be considered is the lower detection limit of smCounter. Mutect might be unable to detect variants below a certain VAF (see figures 3.22 and 3.23). Mutect called the GTVs with an expected VAF below 5% at increasingly higher VAFs for decreasing expected VAF. The blue GTV VAF LOESS curve for smCounter showed this tendency only to a much smaller extent.

The increase or decrease of a regression slope computed for every variant from the variant calling results of a dilution series was used to classify variants according to their supposed origin. The percentage of variants categorized as originating from the tumour portion (negative slope with decreasing tumour content per data set) should be close to 100% in a scenario where only few heterozygote (~50% VAF) or homozygote (~100% VAF) germline variants are present. A flat regression slope, in contrast, does not allow for a clear classification of a variant's putative source. These variants most likely originate from systematically incorporated base errors by error-prone amplification in a sequence dependent manner and/or sequence dependent sequencing errors and, thus, should not change by a considerable amount for different tumour fractions. Ideally, this portion of unclassifiable variants should be zero. Alignment reduction through consensus formation using the majority rule cannot not remove these artefacts which might explain the persisting observation of these variants for all dilution series in the reanalysis.

Tables 3.8 and 3.9 list the putative tumour fraction and the portion of variants which could not be classified with respect to their putative source (*i.e.* flat regression slope). The percentage of unclassifiable variants was lower for Mutect in the NEBNext TruQ4 dilution series and higher for the smMIP dilution series. It is difficult to interpret these results since for the NEBNext TruQ4 dilution series, error prone clusters were omitted which likely removed low VAF variants exhibiting a flat linear regression slope. In total, percentages of unclassifiable variants varied between 15% and 25% of traceable variants. This percentage depends on the definition of 'flat' for regression slopes though.

The smCounter caller exhibited higher tumour percentages of source-classifiable GTV variants and lower tumour percentages for ambiguous calls compared to Mutect. This might be in part due to the higher VAF accuracy of smCounter and due to the lower detection limit, which might have allowed for the detection of more variants from the tumour material.

The high amount of variant calls for smCounter may in part also be because of the lower detection limit but also due to uncleared molecular noise. Therefore, the observed high number of variant calls at low VAF emphasises the need of potent noise suppression and application of variant filters after calling.

In summary, the smCounter caller performed better than the Mutect caller in detecting GTVs and determining their VAFs. In terms of being impacted by uncleared molecular noise, Mutect was more robust than smCounter.

4.9. Tagging Protocol Performance

Comparison of validation results is tricky since the enriched regions differ between tagging protocols. Therefore, the recall result for NEBNext was regarded as more representative than the smMIP result in the exploratory analysis. Seraseq dilution series were most representative for the underlying tagging protocol performances due to the high number of available GTVs compared to data sets available in the exploratory analysis.

Due to the low duplication numbers per template and the inferior region targeting efficiency, the ThruPLEX protocol was found to be unfavourable for low VAF variant calling and massively parallel sequencing since approximately one third of all reads remain unused. Also, the extraordinary high number of unique variants called with smCounter indicating a large portion of uncleared noise for the MiSeq sample shows that the achieved duplication was insufficient.

In case of smMIP, the results for lowest called VAFs indicate that the ultra-deep amplicons could be very useful for low allelic frequency variant calling. Nevertheless, concordance between replicates was low. As discussed in subsection 4.2, the portion of reads not aligning against the human genome was unfavourably high which could be explained in part by the high variability in accumulated errors per cluster. After alignment reduction, group size rank plots indicated that the majority of UMI groups, which were formed by a single alignment pair, could not be reassigned back to the group they originated from. These groups are thought to be the cause of the low variant calling performance. When comparing the estimates of errors per smMIP UMI sequence to estimates for QIASeq and NEBNext, it becomes obvious that sequencing errors could not be the cause of UMI groups being irreducible. NEBNext exhibited the highest rate of estimated sequencing errors per UMI yet exhibited the best reducibility. This also applies to ThruPLEX results which exhibited an error rate per UMI similar to smMIP. Therefore, it can be supposed that errors causing strong UMI sequence alterations which led to pronounced tail regions in rank plots were introduced during PCR amplification. The amount of UMI errors might be lower, if inversion probes would be available from an independent manufacturer which can guarantee for the quality and stability of the product. In any case, smMIP should not be used in conjunction with Mutect variant calling due to the caller's alignment handling. Examples are local rearrangement which takes the targeting approach resulting in narrow amplicons ad absurdum, and default downsampling per position to a maximum coverage of 1,000 reads which should not be changed as stated by the staff of the broad institute [90]. UMI coverage was observed to reach values above 3,000 which means that Mutect eliminates information from well covered sites. Moreover, downsampling renders detection of variants below 0.1% VAF impossible.

NEBNext exhibited stable clusters and high performance in the exploratory analysis. Cluster stability was demonstrated by the minor influence of different clustering settings on GTV recall and number of corrected artefacts (see table 3.10). However, for GTVs of the Seraseq dilution series below 1% VAF, the number of GTV calls decreased in an unfavourable manner (figures 3.27 and 3.32). This may have been caused by either loss of product during the elaborate and timing sensitive wet lab procedure, as employees of the D & R Institute of Human Genetics stated, or due to a limited capturing and/or tagging efficiency. The true cause for the unfavourably low GTV recall below 1% VAF could not be determined *in silico* though. Thus, NEBNext should not be used for the detection of variants with a VAF below 1%.

QIASeq clusters exhibited a lower stability by continuously decreasing in size from the end distal to the capture primer. This manifested in QIASeq GTV recall results and detection limits being sensible to different clustering settings. Furthermore, the absolute number as well as the types of different artefacts corrected during clustering was higher compared to NEBNext results. Nevertheless, QIASeq performed better for lower GTV VAFs. This is evident from the detection limits (figure 3.25) and the 50% GTV observation VAF threshold (table 3.10). Based on the usability threshold at which 50% of GTVs were theoretically detected, QIASeq should only be used for detection of variants down to a VAF of 0.25% for basic and moderate clustering. In contrast to NEBNext results, a monotonously declining characteristic was observed for false positives filtered GTV detection results of QIASeq combined with basic clustering.

The optimal combination of caller and tagging procedure was found to be smCounter and QIASeq which is emphasized by the high percentage of correctly classified GTVs (table 3.9).

4.10. Clustering

A paper from Peng *et al.* (2015) described an alignment reduction approach using multiple rounds of UMI-based clustering [81, p. 11]. UMI groups are divided into a group of supposedly true UMIs and an ambiguous group based on group size. UMIs of the ambiguous group are merged to the largest supposedly true UMI group with an UMI sequence within an edit distance of one. The allowed edit distance is incremented for each round of clustering.

In this thesis, an attempt was made to circumvent the necessity of carrying out multiple rounds of clustering to increase computational performance. The achieved computation times and the main memory usage, though being acceptable for most data sets, could still be improved. The remaining VAF variability might be due to not dividing UMI groups in absorbing and erroneous groups for incorporation. Basically, all UMI groups other than the largest one were regarded as being

potentially formed by an erroneous UMI sequence. The measure of dividing UMI groups into two sets would likely also speed up clustering. Furthermore, subsampling disabled the correction of early UMI errors for identity clusters with original size of 32 and larger. Therefore, after splitting up clusters in two groups, the large cluster portion should also be checked for error events based on the original size.

During code development, using an overly permissive window was observed to highly inflated mate mapping position variability which reduced the resulting advanced clustering specificity. The values chosen for mate mapping position windows were observed to subsequently reduce the formation of large 'divergent' clusters. Divergent clusters are characterized by high variation of mapping positions at the end distal to the capture primer with only few duplicates occupying individual mapping positions (in case a single primer was used for capture per region). Occurrence of divergent clusters were associated with runtime problems in clustering.

Based on the results obtained for different clustering settings, a suggestion can be formulated based on the decreased number of GTV calls to not use the QIASeq protocol in combination with permissive clustering for detection of variants below 0.5% VAF. For deactivated advanced clustering and mild clustering settings, variants down to a VAF of 0.25% may be only mildly affected by the low VAF supressing clustering effect. Variants below these thresholds should be omitted due to the influence of uncleared noise on the variant calling process. In contrast to QIASeq, NEBNext should only be used for detection of variants as low as 1% VAF in all cases.

VAF distributions of ambiguous calls were affected more by clustering than GTV VAF distributions for both the NEBNext and the QIASeq tagging protocols (see figures 3.28 and 3.29). This supports the notion that a considerable portion of ambiguous MOCs stems from insufficiently suppressed noise. Exceptions to this are the results obtained with permissive clustering for QIASeq Seraseq data sets below 0.5% expected GTV VAF. Permissive clustering also greatly reduced the number of GTV calls for these data sets. This indicates an overly permissive use of clustering which resulted in a reduced GTV signal. Furthermore, it cannot be ruled out that a major portion of GTV calls below 0.5% VAF were caused by noise. This was also supported by the contra intuitive increases in GTV call numbers at low GTV VAFs for both tagging protocols (see figures 3.26 and 3.27).

A major limitation of the employed clustering approach is clearly the use of loci- and sequence-independent parameter values. This represents an oversimplification of the underlying situation in the sequencing data. For example, the identity cluster window size was found to be too stringent in some cases. In Seraseq analyses with deactivated advanced clustering, several larger clusters with identical UMI sequence were found in close proximity to each other. In the QIASeq 2% Seraseq

data set, 17 larger clusters (*i.e.* at least three UMI groups per cluster and 100 alignments) mapping to chromosome four exhibited identical UMIs with leftmost mapping positions being distributed over 164 bases. The largest cluster contained 111 UMI groups with 3,320 alignments (subsampled count). The largest observed leftmost mapping position distance between two incompletely clustered identity groups was 47 bases. Therefore, it is recommended to choose the identity window size in a more permissive way in contrast to the advanced clustering window. This identity clustering deficiency is thought to have contributed to GTV VAF deviations by either inflating the number of alternative allele observations in cases where incompletely clustered UMI groups carried an alternative allele, or to decreasing the GTV VAF in the opposite case.

To improve alignment reduction by clustering, a sequence- and cluster-specific model for choosing parameters also regarding the UMI of the absorbing cluster (*i.e.* larger advanced clustering window for UMIs containing larger homopolymer stretches) would be beneficial. Training such a model would require the availability of several wildtype data sets created with the same wet lab procedure as the investigated data set as was done for another background polishing approach [33, p. 21f]. This kind of data was not available for this thesis. The results of training a neuronal network would be specific to the employed sampling method, targeting strategy, molecular tagging protocol, and sequencing platform and chemistry though. This means that changing a parameter of the setup may render the established model obsolete. Moreover, creating a well generalizing model requires extremely expensive sequencing of large control groups in the size of several hundreds. Thus, establishing a noise filter using deep learning seems to be hard to establish despite its benefits which are out of the question.

4.11. Error Prone Cluster Handling

Removing ambiguous information from the analysis (*e.g.* omitting error-prone clusters) resulted in most cases in a severe drop of GTV recall. Therefore, the removal of information per default cannot be recommended.

The NEBNext dilution series combined with Mutect variant calling was the only case where removing error prone clusters turned out to be beneficial. After UMI error correction, about 60% of UMI groups of the NEBNext 100% TruQ4 data set were composed of 3 or more alignment pairs. The majority of clusters not being error-prone is thought to be the reason why omitting error-prone clusters resulted in a higher GTV recall than including them for this dilution series. For other data sets and dilution series, after error correction, still 80% of NEBNext Seraseq and 55% of QIASeq Seraseq clusters were error prone. This also explains in part the large number of variant calls per kilo base for these data sets. Moreover, the GTV calling performance being higher

with error-prone clusters being included can be reasoned by incomplete separation of noisy alignments and GTV allele supporting alignments after clustering.

It can be concluded that for every tagging procedure there may be a threshold for the percentage of error-prone cluster removal which turns out to be beneficial in terms of GTV recall. To this end, at most 40% of clusters being error prone can be suggested as a requirement for omitting error-prone clusters to exerting a positive effect on variant calling performance at least for the NEBNext protocol.

4.12. False Positives Estimation

As the investigation of overall allelic frequency distributions of dilution series datasets showed, the vast majority of variant calls were due to single UMI-supported alternative allele observations and, thus, exhibited small VAFs below 5% (see figures 3.30 and 3.22). Therefore, the assumption was made that molecular noise and base calling errors mostly affect the lower end of the observed allelic frequency scale. A further assumption was made, that the expected noise level can be estimated from the lowest allelic frequency occurring in a dataset. This assumes a peak-like distribution of noise which could not be observed for investigated data sets. Noise distributions usually followed a right-tailed gaussian-like curve slightly extending into higher allelic frequencies overlapping the signal distribution and, thus, was not limited to the lowest observed variant allele frequency.

A portion of additional randomness is added on top of the molecular noise distribution by base calling errors. This portion was completely disregarded in earlier considerations because applied alignment filters and subsampling of sequencing error-sorted alignments was thought to correct the majority of base calling-induced errors. In practice, both the equally distributed errors assumption and the sequence independent errors assumption are violated though. It is well known that DNA polymerases exhibit a sequence-dependent error probability influenced by GC-content with especially low replication accuracy for short repeats and homopolymer stretches resulting in single nucleotide polymorphisms or altered sequence lengths. Thus, base incorporation errors are non-random. It has also been described that the frequency of alternative base observations, which depends on the utilized DNA polymerase [91, p. 5f] [92, p. 4], is not equally distributed over all possible alternative bases with transitions being more frequent than transversions [93, p. 4]. For the sake of simplicity and due to limitations dictated by the experiment design, the equally distributed errors assumption for the three outcome scenarios was still adopted.

As expected, the estimated false positives (EFP) filter led to a more conservative estimate of the true GTV recall for higher GTV VAF data sets and a more liberal estimate for lower GTV VAF

data sets. The EFP filter allowed for portions of SOCs and DOCs in case of low signal to noise level relations which yielded lower GTV recall results compared to filtering all SOCs for basic and moderate clustering. Exceptions were the 0.25% GTV VAF data sets which showed a contra intuitive increase in GTV detections for the moderate and permissive clustering settings. The supposed reason for this were missing dual observation alternative calls and the small number of single observation alternative calls in the control data set for both clustering settings. This circumstance led to an overall smaller estimated number of false positive single- and dual UMI-supported variant calls. This illustrates the limitations of the utilized false positive GTV call estimation for validation experiments which use small truth sets.

5. Conclusion

In this thesis, bioinformatic ways of harnessing molecular barcoding protocols for the detection of somatic mutations were investigated. Ultimately, a software solution in form of a variant calling pipeline was implemented and its parameters optimized. This pipeline is suitable for analysing any paired-end read data created from liquid biopsy samples that were prepared with any UMI protocol. The provided solution was demonstrated to be applicable for detection of variants exhibiting a VAF as low as 0.25%. The lowest detected VAFs were 0.125% for the NEBNext and the QIASeq tagging protocols.

The usability of four barcoding protocols regarding the described variant calling application was assessed. The ThruPLEX and smMIP protocols were found to have only a limited usability for the pursued variant calling approach.

Aspects crucial to solving the inverse problem of reducing an amplified library to its original state were identified. It could be shown that applying simple rules of thumb like suggested in the supplements of the smCounter paper [36, p. 3] would lead to poor reduction of the amplified library. The presented clustering approach adapted better to the given data regardless of the library construction and yielded optimized library reduction per covered base. Earlier findings from Peng *et al.* (2015) were experimentally confirmed and used to rule out sequencing errors as the major cause for uncleared noise ultimately leading to false positive variant calls. Errors introduced during library amplification were found to describe the clear majority of noise that affected variant calling.

Light was shed on the large portion of ambiguous variant calls by investigation the extent of uncleared base noise in UMI sequences and its contribution to false positive variant calls. This unfortunately large portion could be reduced by fine tuning the developed clustering approach. Estimates of false positive variant calls were used to define limits of uncertainty for the NEBNext and QIASeq tagging protocols in the context of different alignment clustering settings.

The power of visualization was demonstrated by incorporating depictions of alignment clusters prior to consensus formation in error source and UMI artefact search. The inhomogeneous distribution of base errors and sequencing errors inside alignment clusters were hypothesized to be a result of changing reaction conditions becoming suboptimal during later cycles of the library amplification. Methods to regard UMI artefacts during alignment reduction were also implemented considering these visualization results. Further research concerning PCR-induced base errors, sequencing artefacts, and also UMI artefacts may be conducted on the basis of cluster visualizations using image processing.

The performance of two variant callers which are capable of tumour-only variant calling was investigated. It was found that smCounter exhibited an approximately 10-fold lower detection limit than Mutect in tumour-only mode. Combining the QIASeq protocol with the smCounter variant caller resulted in the highest GTV recall and the best results for GTV VAFs below 1%. Limitations of the Mutect caller in somatic mutation detection of variants exhibiting low VAFs were pointed out.

An empirical threshold for omitting error prone clusters could be defined for the NEBNext protocol to obtain an increased variant calling performance. Despite attempts to extend the UMI error model used in the UMI-tools publication [68, p. 1f], in most cases, a considerable amount of information remained in error-prone clusters which also could not be omitted without simultaneously reducing GTV calling performance. As a result, the amount of ambiguous variant calls remained high. These findings emphasize the importance of using effective noise filtering or suppressing strategies when pursuing the detection or monitoring of low VAF mutations.

A future goal should be to extend the usability of the implemented UMI analysis pipeline by increasing its performance, especially in terms of noise suppression and computational efficiency. Runtime and memory usage should be reduced considerably by implementing parallelization in conjunction with contig-wise data segmentation. The quality of variant calling results needs to be enhanced by increasing sensitivity through addition of a UMI-aware variant filter. Such a filter could make use of information gained from cluster consensus formation combined with the quality value emitted for each variant call by smCounter.

References

[1] I. H. G. S. Consortium, "Initial sequencing and analysis of the human genome," *Nature,* pp. 860-921, 01 February 2001.

[2] K. Wetterstrand, "DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP)," National Institute of Health, 25 April 2018. [Online]. Available: www.genome.gov/sequencingcostsdata. [Accessed 02 April 2019].

[3] A. Wilmes, A. Limonciel, L. Aschauer, K. Moenks, C. Bielow, M. Leonard, J. Hamon, D. Carpi, S. Ruzek, A. Handler, O. Schmal, K. Herrgen, P. Bellwon, C. Burek and G. Truisi, "Application of integrated transcriptomic, proteomic and metabolomic profiling for the delineation of mechanisms of drug induced cell stress," *Journal of Proteomics,* pp. 180-194, 21 February 2013.

[4] McKusick-Nathans Institute of Genetic Medicine, John Hopkins University (Baltimore, MD), "Online Mendelian Inheritance in Man, OMIM®," McKusick-Nathans Institute of Genetic Medicine, John Hopkins University (Baltimore, MD), April 2019. [Online]. Available: https://omim.org. [Accessed 15 April 2019].

[5] K. Karczewski, L. Francioli, G. Tiao, B. Cummings, J. Alföldi, Q. Wang, R. Collins, K. Laricchia, A. Ganna, D. Birnbaum, L. Gauthier, H. Brand, M. Solomonson and N. Watts, "Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes," *bioRxiv,* p. 531210, 1930 January 2019.

[6] T. 1. G. P. Consortium, "A global reference for human genetic variation," *Nature,* vol. 526, p. 68–74, 30 September 2015.

[7] P. Sudmant, T. Rausch, E. Gardner, R. Handsaker, A. Abyzov, J. Huddleston, Y. Zhang, K. Ye, G. Jun, M. Hsi-Yang Fritz, M. Konkel, A. Malhotra, A. Stütz, X. Shi and F. Paolo Casale, "An integrated map of structural variation in 2,504 human genomes," *Nature,* vol. 526, p. 75–81, 01 October 2015.

[8] Y. Xue, Y. Chen, Q. Ayub, N. Huang, E. Ball, M. Mort, A. Phillips, K. Shaw, P. Stenson, D. Cooper, C. Tyler-Smith and 1. G. P. Consortium, "Deleterious- and disease-allele prevalence in healthy individuals: insights from current predictions, mutation databases, and population-

scale resequencing," *American journal of human genetics,* vol. 91, no. 6, p. 1022–1032, 07 December 2012.

[9] D. Lane, "p53, guardian of the genome," *Nature,* vol. 358, no. 6381, pp. 15-16, 02 July 1992.

[10] Office for National Statistics, "Cancer survival by stage (experimental statistics): adults diagnosed 2012, 2013 and 2014 and followed up to 2015," 10 June 2016. [Online]. Available: https://www.ons.gov.uk/. [Accessed 03 April 2019].

[11] R. Gulati, H. Cheng, P. Lange, P. Nelson and R. Etzioni, "Screening men at increased risk for prostate cancer diagnosis: Model estimates of benefits and harms," *Cancer Epidemiology and Prevention Biomarkers,* pp. 222-227, 14 October 2016.

[12] A. Tsodikov, A. Mariotto, E. Wever, E. Feuer, G. Draisma, H. de Koning, R. Gulati and R. Etzioni, "Lead Time and Overdiagnosis in Prostate-Specific Antigen Screening: Importance of Methods and Context," *Journal of the National Cancer Institute,* vol. 101, no. 6, pp. 374-383, 18 March 2009.

[13] E. Heitzer, I. Haque, C. Roberts and M. Speicher, "Current and future perspectives of liquid biopsies in genomics-driven oncology," *Nature Reviews Genetics,* pp. 71-88, 01 February 2019.

[14] M. Alieva, J. van Rheenen and M. Broekman, "Potential impact of invasive surgical procedures on primary tumor growth and metastasis," *Clinical & Experimental Metastasis,* vol. 35, no. 4, pp. 319-331, 08 May 2018.

[15] E. Gormally, P. Vineis, G. Matullo, F. Veglia, E. Caboux, E. Le Roux, M. Peluso, S. Garte, S. Guarrera, A. Munnia, L. Airoldi, H. Autrup, C. Malaveille and A. Dunning, "TP53 and KRAS2 Mutations in Plasma DNA of Healthy Subjects and Subsequent Cancer Occurrence: A Prospective Study," *Cancer Research,* vol. 66, no. 13, pp. 6871-6876, 01 July 2006.

[16] J. Belic, R. Graf, T. Bauernhofer, Y. Cherkas, P. Ulz, J. Waldispuehl-Geigl, S. Perakis, M. Gormley, J. Patel, W. Li, J. Geigl, D. Smirnov, E. Heitzer, M. Gross and M. Speicher, "Genomic alterations in plasma DNA from patients with metastasized prostate cancer receiving abiraterone or enzalutamide," *International Journal of Cancer,* vol. 143, no. 5, pp. 1236-1248, 25 March 2018.

[17] A. Bardelli and K. Pantel, "Liquid Biopsies, What We Do Not Know (Yet)," *Cancer Cell,* vol. 31, no. 2, pp. 172-179, 13 February 2017.

[18] M. Stroun, J. Lyautey, C. Lederrey, A. Olson-Sand and P. Anker, "About the possible origin and mechanism of circulating DNA: Apoptosis and active DNA release," *Clinica Chimica Acta,* vol. 313, no. 1, pp. 139-142, 01 November 2001.

[19] P. Laktinov, S. Tamkovich, E. Rykova, O. Bryzgunova, A. Starkov, N. Kuznetsova and V. Vlassov, "Cell-Surface-Bound Nucleic Acids: Free and Cell-Surface-Bound Nucleic Acids in Blood of Healthy Donors and Breast Cancer Patients," *Annals of the New York Academy of Sciences,* vol. 1022, no. 1, pp. 221-227, 08 July 2009.

[20] E. Morozkin, P. Laktionov, E. Rykova and V. Vlassov, "Extracellular Nucleic Acids in Cultures of Long-Term Cultivated Eukaryotic Cells," *Annals of the New York Academy of Sciences,* vol. 1022, no. 1, pp. 244-249, 08 July 2009.

[21] C. Bettegowda, M. Sausen, R. Leary, I. Kinde, Y. Wang, N. Agrawal, B. Bartlett, H. Wang, B. Luber, R. Alani, E. Antonarakis, N. Azad, A. Bardelli, H. Brem, J. Cameron and C. Lee, "Detection of Circulating Tumor DNA in Early- and Late-Stage Human Malignancies," *Science Translational Medicine,* vol. 224, no. 6, p. 224ra24, 19 February 2014.

[22] Y.-C. Yang, D. Wang, L. Jin, H.-W. Yao, J.-H. Zhang, J. Wang, X.-M. Zhao, C.-Y. Shen, W. Chen, X.-L. Wang, R. Shi, S.-Y. Chen and Z.-T. Zhang, "Circulating tumor DNA detectable in early- and late-stage colorectal cancer patients," *Bioscience Reports,* vol. 38, no. 4, p. BSR20180322, 31 July 2018.

[23] K.-L. Spindler, N. Pallisgaard, I. Vogelius and A. Jakobsen, "Quantitative Cell-Free DNA, KRAS, and BRAF Mutations in Plasma from Patients with Metastatic Colorectal Cancer during Treatment with Cetuximab and Irinotecan," *Clinical Cancer Research,* vol. 18, no. 4, pp. 1177-1185, 06 January 2012.

[24] C. Fiala and E. Diamandis, "Utility of circulating tumor DNA in cancer diagnostics with emphasis on early detection," *BMC medicine,* vol. 16, no. 1, p. 166, 02 October 2018.

[25] J. Cohen, L. Li, Y. Wang, C. Thoburn, B. Afsari, L. Danilova, C. Douville, A. Javed, F. Wong, A. Mattox, R. Hruban, C. Wolfgang, M. Goggins, M. Da Molin and T.-L. Wang, "Detection and localization of surgically resectable cancers with a multi-analyte blood test," *Science,* vol. 359, no. 6378, pp. 926-930, 23 February 2018.

[26] R. Ptashkin, D. Mandelker, C. Coombs, K. Bolton, Z. Yelskaya, D. Hyman, D. Solit, J. Baselga, M. Arcila, M. Ladanyi, L. Zhang, R. Levine, M. Berger and A. Zehir, "Prevalence of

Clonal Hematopoiesis Mutations in Tumor-Only Clinical Genomic Profiling of Solid Tumors," *JAMA oncology,* vol. 4, no. 11, pp. 1589-1593, 01 November 2018.

[27] D. Sefrioui, F. Blanchard, E. Toure, P. Basile, L. Beaussire, C. Dolfus, A. Perdrix, M. Paresy, M. Antonietti, I. Iwanicki-Caron, R. Alhameedi, S. Lecleire and A. Gangloff, "Diagnostic value of CA19.9, circulating tumour DNA and circulating tumour cells in patients with solid pancreatic tumours," *British Journal of Cancer,* vol. 117, no. 7, pp. 1017-1025, 26 September 2017.

[28] Y. Wang, L. Li, C. Douville, J. Cohen, T.-T. Yen, I. Kinde, K. Sundfelt, S. Kjaer, R. Hruban, I.-M. Shih, T.-L. Wang, R. Kurman, S. Springer, J. Ptak, M. Popoli and J. Schaefer, "Evaluation of liquid from the Papanicolaou test and other liquid biopsies for the detection of endometrial and ovarian cancers," *Science Translational Medicine,* vol. 10, no. 433, p. eaap8793, 21 March 2018.

[29] Y. Korshunova, R. Maloney, N. Lakey, R. Citek, B. Bacher, A. Budiman, J. Ordway, R. McCombie, J. Leon, J. Jeddeloh and J. McPherson, "Massively parallel bisulphite pyrosequencing reveals the molecular complexity of breast cancer-associated cytosine-methylation patterns obtained from tissue and serum DNA," *Genome Research,* pp. 19-29, 01 Jannuary 2008.

[30] M. Widschwendter, I. Evans, A. Jones, S. Ghazali, D. Reisel, A. Ryan, A. Gentry-Maharaj, M. Zikan, D. Cibula, J. Eichner, M. Alunni-Fabbroni, J. Koch, W. Janni and T. Paprotka, "Methylation patterns in serum DNA for early identification of disseminated breast cancer," *Genome Medicine,* vol. 9, no. 1, p. 115, 22 December 2017.

[31] M. Widschwendter, M. Zikan, B. Wahl, H. Lempiäinen, T. Paprotka, I. Evans, A. Jones, S. Ghazali, D. Reisel, J. Eichner, T. Rujan, Z. Yang, A. Teschendorff, A. Rya and D. Cibula, "The potential of circulating tumor DNA methylation analysis for the early detection and management of ovarian cancer," *Genome Medicine,* vol. 9, no. 1, p. 116, 22 December 2017.

[32] J. Moss, J. Magenheim, D. Neiman, H. Zemmour, N. Loyfer, A. Korach, Y. Samet, M. Maoz, H. Druid, P. Arner, K.-Y. Fu, E. Kiss, K. Spalding, G. Landesberg, A. Zick and A. Grinshpun, "Comprehensive human cell-type methylation atlas reveals origins of circulating cell-free DNA in health and disease," *Nature Communications,* vol. 9, no. 1, p. 5068, 29 November 2018.

[33] A. Newman, A. Lovejoy, D. Klass, D. Kurtz, J. Chabon, F. Scherer, H. Stehr, C. Liu, S. Bratman, C. Say, L. Zhou, J. Carter, R. West, G. Sledge, J. Shrager, B. Loo and J. Neal,

"Integrated digital error suppression for improved detection of circulating tumour DNA," *Nature Biotechnology,* vol. 34, no. 5, pp. 547-555, 28 March 2016.

[34] A. Ståhlberg, P. Krzyzanowski, M. Egyud, S. Filges, L. Stein and T. Godfrey, "Simple multiplexed PCR-based barcoding of DNA for ultrasensitive mutation detection by next-generation sequencing," *Nature Protocols,* pp. 664-682, 02 March 2017.

[35] F. Lan, J. Haliburton, A. Yuan and A. Abate, "Droplet barcoding for massively parallel single-molecule deep sequencing," *Nature Communications,* vol. 7, p. 11784, 29 June 2016.

[36] C. Xu, M. Ranjbar, Z. Wu, J. DiCarlo and Y. Wang, "Detecting very low allele fraction variants using targeted DNA sequencing and a novel molecular barcode-aware variant caller," *BMC Genomics,* pp. Suppl:1-11, 03 January 2017.

[37] C. Quince, U. Ijaz, W. Sloan, M. Schirmer, N. Hall and R. D'Amore, "Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform," *Nucleic Acids Research,* vol. 43, no. 6, p. e37, 13 January 2015.

[38] Broad Institute, "Best Practices for Variant Calling with the GATK," Broad Institute, 19 March 2015. [Online]. Available: https://www.broadinstitute.org/partnerships/education/broade/best-practices-variant-calling-gatk-1. [Accessed 26 December 2018].

[39] New England BioLabs, "NEBNext Technical Guide for Illumina," November 2017. [Online]. Available: https://www.neb.com/-/media/nebus/files/brochures/nebnextillumina.pdf. [Accessed 28 December 2018].

[40] J. Hiatt, C. Pritchard, S. Salipante, B. O'Roak and J. Shendure, "Single molecule molecular inversion probes for targeted, high-accuracy detection of low-frequency variation," *Genome Research,* pp. 843-854, 23 October 2013.

[41] Horizon Discovery, "cfDNA Reference Standards Frequently Asked Questions," Horizon Discovery Group plc, n.d. [Online]. Available: https://www.horizondiscovery.com/resources/scientific-literature/faq/cfdna. [Accessed 26 December 2018].

[42] P. Ulz, J. Belic, G. Ricarda, M. Auer, I. Lafer, J. Geigl and E. Heitzer, "Whole-genome plasma sequencing reveals focal amplifications as a driving force in metastatic prostate cancer," *Nature Communications,* p. 7:12008, 22 June 2016.

[43] Takara Bio, "ThruPLEX DNA-seq Kit User Manual," n.d. [Online]. Available: https://www.takarabio.com/assets/documents/User%20Manual/ThruPLEX%20Tag-seq%20Kit%20User%20Manual_022818.pdf. [Accessed 28 December 2018].

[44] C. Shannon, "A mathematical theory of communication," *The Bell System Technical Journal,* vol. 27, no. 3, pp. 379-423, 01 July 1948.

[45] QIAGEN, "QIAGEN Targeted DNA Panel Handbook," May 2017. [Online]. Available: https://www.qiagen.com/at/resources/download.aspx?id=8907edbe-a462-4883-ae1b-2759657e7fd0&lang=en. [Accessed 28 December 2018].

[46] Illumina, Inc., "Illumina website," Illumina, Inc., 09 April 2019. [Online]. Available: www.illumina.com. [Accessed 15 April 2019].

[47] G. van Rossum and F. Drake, "Python Reference Manual," Python Software Foundation, Wilmington, USA, 18 March 2019. [Online]. Available: https://docs.python.org/2/. [Accessed 15 April 2019].

[48] S. Andrews, "FastQC, Babraham Bioinformatics," Babraham Institute, Cambridge, UK, 04 October 2018. [Online]. Available: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/. [Accessed 01 April 2017].

[49] S. Wingett, "FastQ Screen, Babraham Bioinformatics," Babraham Bioinformatics, Camebridge, UK, 12 September 2018. [Online]. Available: https://www.bioinformatics.babraham.ac.uk/projects/fastq_screen/. [Accessed 15 April 2019].

[50] P. Ewels, "MultiQC website," Science for Life Laboratory, Solna, Sweden, 21 December 2018. [Online]. Available: https://multiqc.info/. [Accessed 3 October 2018].

[51] R Core Team, "R Foundation for Statistical Computing," R Foundation for Statistical Computing, 26 March 2019. [Online]. Available: https://www.R-project.org. [Accessed 15 April 2019].

[52] H. Wickham, ggplot2: Elegant Graphics for Data Analysis, New York: Springer-Verlag New York, 2016.

[53] J. Hunter, "Matplotlib: A 2D Graphics Environment," *Computing in Science & Engineering,* vol. 9, no. 3, pp. 90-95, 18 June 2007.

[54] S. Morton, "natsort Documentation," Seth M. Morton, 04 February 2019. [Online]. Available: https://natsort.readthedocs.io. [Accessed 30 December 2018].

[55] T. Oliphant, A guide to NumPy, USA: Trelgol Publishing, 2006.

[56] W. McKinney, "Data Structures for Statistical Computing in Python," in *Proceedings of the 9th Python in Science Conference*, Austin, texas, 2010.

[57] A. Heger and K. Jacobs, "Documentation - pysam: htslib interface for python," Andreas Heger, Kevin Jacobs et al., 27 July 2018. [Online]. Available: https://pysam.readthedocs.io. [Accessed 30 December 2018].

[58] J. Casbon and @jdoughertyii, "Documentation - PyVCF - A Variant Call Format Parser for Python," James Casbon, 08 February 2012. [Online]. Available: https://pyvcf.readthedocs.io. [Accessed 30 December 2018].

[59] T. Oliphant, "Python for Scientific Computing," *Computing in Science & Engineering,* vol. 9, no. 3, pp. 10-20, 18 June 2007.

[60] M. Waskom, "Documentation - seaborn: statistical data visualization," Michael Waskom, July 2018. [Online]. Available: https://seaborn.pydata.org. [Accessed 30 December 2018].

[61] H. Li and R. Durbin, "Fast and accurate short read alignment with Burrows–Wheeler transform," *Bioinformatics,* vol. 26, no. 5, pp. 589-595, 15 January 2010.

[62] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin and 1. G. P. D. P. Subgroup, "The Sequence Alignment/Map format and SAMtools," *Bioinformatics,* no. 25, pp. 2078-2079, 15 August 2009.

[63] D. Barnett, E. Garrison, A. Quinlan, M. Strömberg and G. Marth, "BamTools: a C++ API and toolkit for analyzing and managing BAM files," *Bioinformatics,* vol. 27, no. 12, pp. 1691-1692, 14 April 2011.

[64] Broad Institute, "Picard toolkit website," Broad Institute, n.d. [Online]. Available: http://broadinstitute.github.io/picard/. [Accessed 4 November 2018].

[65] A. Quinlan and I. Hall, "BEDTools: a flexible suite of utilities for comparing genomic features," *Bioinformatics,* vol. 26, no. 6, pp. 841-842, 28 January 2010.

[66] A. Quinlan and N. Kindlon, "BEDTools website," Aaron Quinlan and Neil Kindlon, 23 March 2019. [Online]. Available: https://bedtools.readthedocs.io. [Accessed 15 April 2019].

[67] M. Marcel, "Cutadapt removes adapter sequences from high-throughput sequencing reads," *EMBnet.journal,* vol. 17, no. 1, pp. 10-12, 1 July 2011.

[68] T. Smith, A. Heger and I. Sudbery, "UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy," *Genome Research,* pp. 491-499, 18 January 2017.

[69] K. Cibulskis, M. Lawrence, S. Carter, A. Sivachenko, D. Jaffe, C. Sougnez, S. Gabriel, M. Meyerson, E. Lander and G. Getz, "Sensitive detection of somatic point mutaions in impure and heterogenious cancer samples," *Computational Biology,* vol. 31, no. 3, pp. 213-219, 10 February 2013.

[70] C. Xu, M. Ranjbar, Z. WU, J. DiCarlo and Y. Wang, "Detecting very low allele fraction variants using targeted DNA sequencing and a novel molecular barcode-aware variant caller," *BMC Genomics,* p. 18:5, 3 January 2017.

[71] Genome Research Limited, "HTSlib," Genome Research Limited, Hinxton, United Kingdom, n.d. [Online]. Available: http://www.htslib.org/. [Accessed 30 December 2018].

[72] International Human Genome Consortium, "Initial sequencing and analysis of the human genome," *Nature,* pp. 860-921, 15 February 2001.

[73] J. Kent, C. Sugnet, T. Furey, K. Roskin, T. Pringle, A. Zahler and D. Haussler, "The Human Genome Browser at UCSC," *Genome research,* vol. 12, no. 6, pp. 996-1006, 06 May 2002.

[74] NCBI Resource Coordinators, "Database resources of the National Center for Biotechnology Information," *Nucleic Acids Research,* vol. 44, pp. D7-D19, 28 November 2015.

[75] JetBrains, "PyCharm IDE," JetBrains s.r.o., Prague, Czech Republic, 27 March 2019. [Online]. Available: https://www.jetbrains.com/pycharm/. [Accessed 15 April 2019].

[76] GitHub, Inc., "GitHub," GitHub, Inc., San Francisco, USA, November 2018. [Online]. Available: https://github.com. [Accessed 30 December 2018].

[77] Software Freedom Conservancy, "Git," Software Freedom Conservancy, New York, USA, n.d. [Online]. Available: https://git-scm.com. [Accessed 30 December 2018].

[78] T. Kosse, "FileZilla website," Tim Kosse and Team, 13 April 2019. [Online]. Available: https://filezilla-project.org/. [Accessed 15 April 2019].

[79] J. Robinson, H. Thorvaldsdóttir, W. Winckler, M. Guttman, E. Lander, G. Getz and J. Mesirov, "Integrative genomics viewer," *Nature Biotechnology,* vol. 29, pp. 24-26, 10 January 2011.

[80] H. Thorvaldsdóttir, J. Robinson and J. Mesirov, "Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration," *Briefings in Bioinformatics,* vol. 14, no. 2, pp. 178-192, 1 March 2013.

[81] Q. Peng, R. Satya, M. Lewis, P. Randad and Y. Wang, "Reducing amplification artifacts in high multiplex amplicon seqeuncing by using molecular barcodes," *BMC Genomics,* p. 589, 7 August 2015.

[82] L. Heng, "Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM," *arXiv.org,* p. arXiv:1303.3997, 26 May 2013.

[83] R. Edgar and H. Flyvbjerg, "Error filtering, pair assembly and error correction for next-generation sequencing reads," *Bioinformatics,* pp. 3476-3482, 2 July 2015.

[84] A. Bolger, M. Lohse and B. Usadel, "Trimmomatic: a flexible trimmer for Illumina sequence data," *Bioinformatics,* pp. 2114-2120, 01 April 2014.

[85] S. Altschul, T. Madden, A. Schäffer, J. Zhang, Z. Zhang, W. Miller and D. Lipman, "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs," *Nucleic Acids Research,* pp. 3389-402, 1 September 1997.

[86] Illumina, "Illumina Adapter Sequences Document," 22 February 2019. [Online]. Available: https://support.illumina.com/downloads/illumina-adapter-sequences-document-1000000002694.html. [Accessed 15 April 2019].

[87] J. Kebschull and A. Zador, "Sources of PCR-induced distortions in high-throughput sequencing data sets," *Nucleic Acids Research,* vol. 43, no. 21, p. e143, 17 July 2015.

[88] S. Quake, M. Kowarsky, J. Camunas-Soler, M. Kertesz, W. Koh, W. Pan, L. Martin, N. Neff, J. Okamoto, R. Wong, S. Kharbanda, Y. El-Sayed, Y. Blumenfeld, N. Wolfe and G. Shaw, "Numerous uncharacterized and highly divergent microbes which colonize humans are revealed by circulating cell-free DNA," *Proceedings of the National Academy of Science,* pp. 9623-9628, 05 September 2017.

[89] S. Mangul, S. Driesche, L. Martin, K. Martin and E. Eskin, "UMI-Reducer: Collapsing duplicate sequencing reads via Unique Molecular Identifiers," *bioRxiv (preprint),* pp. 1-10, 25 January 2017.

[90] BROAD Institute, "GATK forum," BROAD Institute, June 2017. [Online]. Available: https://gatkforums.broadinstitute.org/gatk/discussion/9721/mutect2-gets-different-results-when-i-change-the-downsample-level. [Accessed 03 March 2019].

[91] D. Shagin, I. Shagina, A. Zaretsky, E. Barsova, I. Kelmanson, S. Lukyanov, D. Chudakov and M. Shugay, "A high-throughput assay for quantitative measurement of PCR errors," *Scientific Reports,* vol. 7, p. 2718, 2 June 2017.

[92] V. Potapov and J. Ong, "Examining Sources of Error in PCR by Single-Molecule Seqeuncing," *PloS one,* vol. 12, no. 1, p. e0169774, 6 January 2017.

[93] P. McInerney, P. Adams and M. Hadi, "Error Rate Comparison during Polymerase Chain reaction by DNA Polymerase," *Molecular Biology International,* p. 287430, 17 August 2014.

[94] P. dos Santos Sant'Ana, A. Araujo, C. Pimenta, M. Ker Bezerra, S. Borges, V. Neto, J. Dias, A. Lopes, D. Rios and M. de Barros Pinheiro, "Clinical and laboratory profile of patients with sickle cell anemia," *Revista Brasileira de Hematologia e Hemoterapia,* vol. 39, no. 1, pp. 40-45, 19 October 2016.

[95] S. Tomlins, D. Rhodes, S. Perner, S. Dhanasekaran, R. Mehra, X.-W. Sun, S. Varambally, X. Cao, J. Tchinda, R. Kuefer, C. Lee, J. Montie, R. Shah, K. Pienta and M. Rubin, "Recurrent Fusion of TMPRSS2 and ETS Transcription Factor Genes in Prostate Cancer," *Science,* vol. 310, no. 5748, pp. 644-648, 28 October 2005.

[96] C. Mungall, J. NguyenXuan, N. Dunn, N. Washington, S. Carbon, S. Lewis, B. Laraway, D. Keith, E. Foster, J. Gourdine, J. McMurry, K. Shefchek, M. Engelstad, M. Brush and N.

Vasilevsky, "The Monarch Initiative: an integrative data and analytic platform connecting phenotypes to genotypes across species," *Nucleic Acids Research,* vol. 45, no. D1, pp. D712-D722, 29 November 2016.

[97] R. Palmirotta, D. Lovero, P. Cafforio, C. Felici, F. Mannavola, E. Pellè, D. Quaresmini, M. Tucci and F. Silvestris, "Liquid biopsy of cancer: a multimodal diagnostic tool in clinical oncology," *Therapeutic Advances in Medical Oncology,* p. 1758835918794630, 06 September 2018.

[98] M. Neumann, S. Bender, T. Krahn and T. Schlange, "Progress, ctDNA and CTCs in Liquid Biopsy - Current Status and Where We Need to," *Computational and Structural Biotechnology Journal,* pp. 190-195, 07 July 2018.

List of Figures

Figure **Page**

List of Tables

Table **Page**