Implementation of NGSCheckMate in the Princess Máxima Centre

Date: 29-11-2021

Written By:	Cyriel Huijer
Student Number:	5914671
Supervisor:	Dr. Hindrik Kerstens
First Examiner:	Dr. Hindrik Kerstens
Second Examiner:	Dr. Philip Lijnzaad

Table of Contents

Abstract

Nowadays, in a research hospital such as the Princess Máxima Centre (PMC), research and patient treatment is often substantiated on NGS data. Therefore, quality control of patient data is vital to preserve data integrity. However, several steps of the process from patient to genotype are vulnerable to sample swaps. For this reason, NGSCheckMate was produced, a tool which retrospectively checks whether samples are labelled correctly based on a set of 21K SNPs. Nevertheless, running NGSCheckMate utilising the original 21K SNP set was found to be computationally inefficient in the PMC, with runtimes of patient samples adding up to ~68 hours. Moreover, data coming out of the PMC biobank sequencing pipeline was observed not to be compatible with NGSCheckMate as no integration of RNA-Seq with W[GX]S was achieved, even though samples were obtained from the same biomaterial. By selection of SNPs based on variety in minor allele and coverage across RNA-Seq samples, smaller SNP sets were created that maintained and improved performance compared to the original 21K set. Total runtime of NGSCheckMate was decreased from ~68 to ~2 hours. Furthermore, in combination with pre-processing and additional filtering of low-quality files, RNA-Seq integration was improved. In conclusion, this study presents a range of smaller SNP sets that both decrease runtime and improve performance of NGSCheckMate in sample swap detection.

1. Introduction

To date, paediatric cancer is still the leading cause of disease-related child death in the western world, even though the average survival rate has increased from 10% to 80% in the past decades[1]. In the Netherlands, around 600 children are diagnosed with cancer each year[2]. Despite the main cause of death, childhood cancer is considered a rare disease[3]. Eventually, the need for centralisation in order to pool all expertise in one hospital led to the establishment of the Princess Máxima Centre (PMC) in 2014[4]. In 2020, a total of 919 patients were treated in the PMC[5].

In cancer diagnosis/treatment, somatic mutations are investigated by tumour sample sequencing, with blood (normal) samples taken as reference. RNA-Sequencing (RNA-Seq) is used as a diagnostic tool to study gene expression and gene fusion[6]. In *Figure 1A*, the processing of patient samples at the PMC is visualised. From each patient, a normal and tumour sample is taken. While normal samples are only sequenced for the exome, tumour samples are also sequenced for the transcriptome. Only if a patient has given informed consent, additional whole genome sequencing (WGS) is performed besides whole exome sequencing (WXS). Furthermore, a tumoroid can be derived from malignant tissue, on which also W[GX]S and RNA-Seq is performed. Tumoroids are grown to test for potential treatments with, for example, a drug screening assay. After the sequencing, metadata is entered into Trecode, an integrative database that records both clinical and research (meta) data[7]. Once the metadata is entered into Trecode, the sequencing analysis is highly automated. From Trecode, fastq-files are automatically converted to variant call format (VCF) files.

Especially in a clinical environment such as the PMC, quality-control (QC) is essential, because patient treatment may be substantiated on NGS data. Several steps of the process from patient to genotype are vulnerable to sample swaps, for instance during sample collection, by mislabelling of tubes in the lab, or by introduction of typos when registering samples. Additionally, because there are several routes (HiX, GLIMS, Sympathy) from which an individual ID is created for a patient, in rare cases more than one identifier is accidentally created per patient. Moreover, research metadata was managed in an Excel file, which is also thought to be one of the major causes of sample swaps. Therefore, prior to entering patient data into Trecode, the process is error prone as a result of human handling.

Currently, hospitals often use two approaches to preserve data integrity: batch-wise genotype checks and retrospective cohort checks. Several methods have been introduced for matching genotypes of patient samples in order to predict sample swaps, i.e. examining short tandem repeats (STRs)[8] and single nucleotide polymorphisms (SNPs)[9,10]. However, the former is not feasible for sequencing strategies such as WXS and RNA-Seq, since most STRs are in the

non-coding region of the genome and are longer than typical sequencing reads. Therefore, NGSCheckMate was introduced to match patient samples based on highly variable SNPs. *Lee et al.* presented NGSCheckMate, a tool for sample-pairing QC, applicable to multiple data formats[11]. Integration of various sequencing strategies is required for the PMC, as both W[GX]S and RNA-Seq samples are derived from patients. Consequently, NGSCheckMate was included in the pipeline of the PMC to retrospectively check for potential sample swaps.

In practice, successful pairing of WGS and WXS data was achieved with the use of NGSCheckMate, nonetheless, integrating RNA-Seq was unsuccessful for VCF-files coming out of the sequencing pipeline. Moreover, running NGSCheckMate was computationally inefficient, as runtimes may take up multiple hours and sometimes even more than a day (*Figure 1B)*. From *Figure 1B* can be concluded that the number of samples has grown enormously over the past months, with an increase of approximately 500 RNA-Seq samples over the course of seven months, adding up to a total of 2314 samples. The highest number of analyses included in the article of *Lee et al.* is 842[11], whereas the total number of samples in the PMC is considerably higher (~6000), which could eventually cause NGSCheckMate not to be computationally feasible anymore. This is especially the case when different sequencing strategies are combined, as the runtime of all WGS/WXS/RNA-Seq samples (n = 5632) was roughly 68 hours.

The aim of this this study is to decrease the runtime of NGSCheckMate and to improve integration of RNA-Seq together with W[GX]S. Since NGSCheckMate sample match prediction is based on a feature set of 21K SNPs, the approach of the study is to perform a feature set optimisation in order to find SNPs that have genotype calls in all sequencing strategies. By selectively dropping SNPs from the feature set, with a drastically smaller set as a result, we expect to lower the runtime of NGSCheckMate, together with an improved integration of RNA-Seq samples with other library strategies while maintaining sufficient resolution in the sample correlations to reliably detect sample swaps.

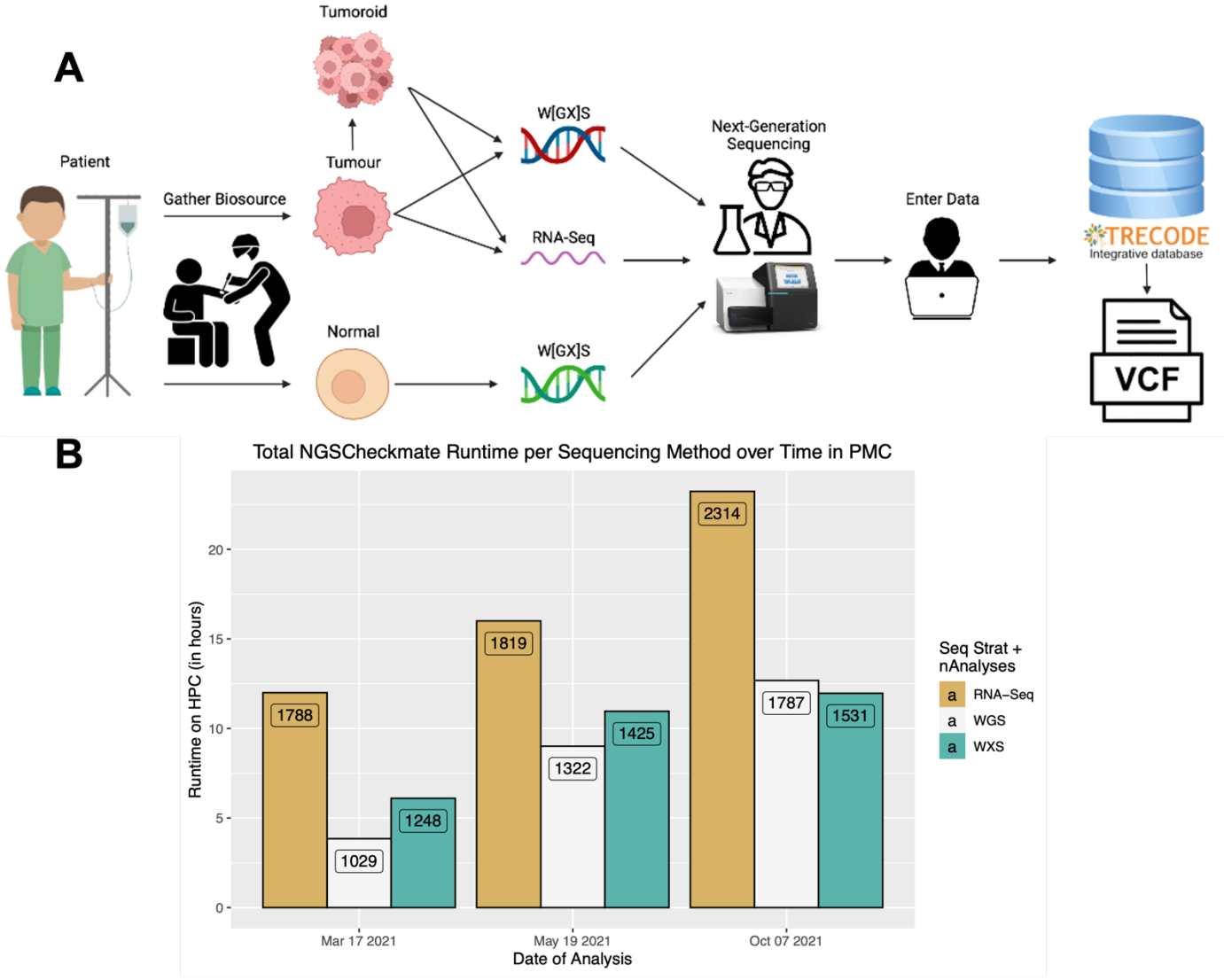

Figure 1: *A) Data process from patient to VCF-file. Normal and tumour tissue is collected, followed by WGS/WXS/RNA-Seq analysis. After NGS, files are entered in Trecode and converted to VCF-files via an automatised pipeline. B) Runtimes and sample input per sequencing strategy of NGSCheckMate.*

2. Methods

2.1. Cohort & data processing

The main dataset consisted of all current and previous patients in our centre whose samples were sequenced before 07-10-2021 and passed the QC-check, which provided us with WGS, WXS, and RNA-Seq data of patients, however, VCF-files only contained the 21K SNP genotypes derived from the article of NGSCheckMate[11] for privacy reasons. During this study, curated subsets were used in order to test for specific parameters, e.g., the rhabdomyosarcoma subset (RMS set, n = 111), the Wilms' tumour subset (n = 130), and the organoid medulloblastoma subset (n = 20). For other subsets reported in this paper, selection criteria are described in the result section.

In accordance with the standardised biobanking protocols of the PMC (Hehir-Kwa, 2021, manuscript under revision), RNA and DNA isolation was performed on tumour tissue, and for its respective normal sample, DNA was isolated from whole blood. More details on sequencing protocols can be found in the paper of Van Belzen et al[12].

Pre-processing of RNA-Seq and W[GX]S was conducted with the standardised pipelines within the PMC, which implement GATK 4.0 best practices for variant calling using a WDL and Cromwell-based workflow[7,13]. QC was performed with Fastqc (version 0.11.5) to calculate the number of sequencing reads. For W[GX]S and RNA-Seq metrics such as insert size and MarkDuplicates, Picard (version 2.20.1) was applied. RNA-Seq reads were aligned using Star (version 2.7.2b) to GRCh38 and Gencode (version 31). BWA mem (version 0.7.13) was used for the alignment of W[GX]S to GRCh38.

2.2. NGSCheckMate

NGSCheckMate is a program that can retrospectively verify the identity of NGS data, which is done by correlating variant allele fractions (VAFs) of samples[11]. The VAF is defined as the ratio of the number of reads supporting a biallelic SNP to the total number of reads on that SNP locus. Two samples are classified as matched based upon the score of the Pearson correlation between all VAFs of the surveyed SNPs for these two samples. Correlation cut-offs were determined on the matched and unmatched correlations of a training data set of germline WGS TCGA samples from stomach cancer patients (n = 40). When the correlation of two samples is closer to the matched correlation of the training dataset, it is considered a match. Furthermore, cut-off values are based on the minimal average sequencing depth of two samples, which accounts for the quality of samples. The SNPs on which genotypes are called were derived from dbSNP, from this database 21,067 exonic SNPs were selected[14]. These SNPs are highly variable across the population and therefore suitable for barcoding

genomes. NGSCheckMate outputs a correlation matrix between all samples that can be used to create a dendrogram in which sample relations become visible. An overview of the workflow of NGSCheckMate is shown in *Figure S1*.

2.3. QCheckMate

QCheckMate is a program that was produced in-house by Hindrik Kerstens (unpublished work) for better interpretation of complex dendrograms. NGSCheckMate software already has a standard script which produces output in the form of a dendrogram, however, these dendrograms are hard to interpret once the sample input starts increasing. The paper only presents dendrograms in the range of approximately 50 samples, on the contrary, in the PMC the sample input can be as high as almost 6,000 samples, a number that is only expected to increase over the years.

Therefore, QCheckMate uses colour coding and symbols in order to clarify the dendrogram and makes it easier to interpret for users. Moreover, it automatically creates informative sample labels based on metadata that is found in Trecode to the samples included in the analysis. In this paper, the sample labels will include biomaterial ID, individual ID, followed by the library strategy and tissue type. An example is shown here: *PM[AGLOR]BMxxxxxx_PMCIDxxxxxx_W[GX]S/RNA-Seq_tumor/tumoroid/normal.* If QCheckMate detects two samples with different individual IDs in the same tree, it adds "***" in front of the label, indicating that a sample swap has occurred. Additionally, when samples from one individual ID are not found in consecutive order, it colour-codes the label of these samples, making it clear that elsewhere in the dendrogram samples from that individual are present. In *Figure 2*, typical output of QCheckMate is visualised.

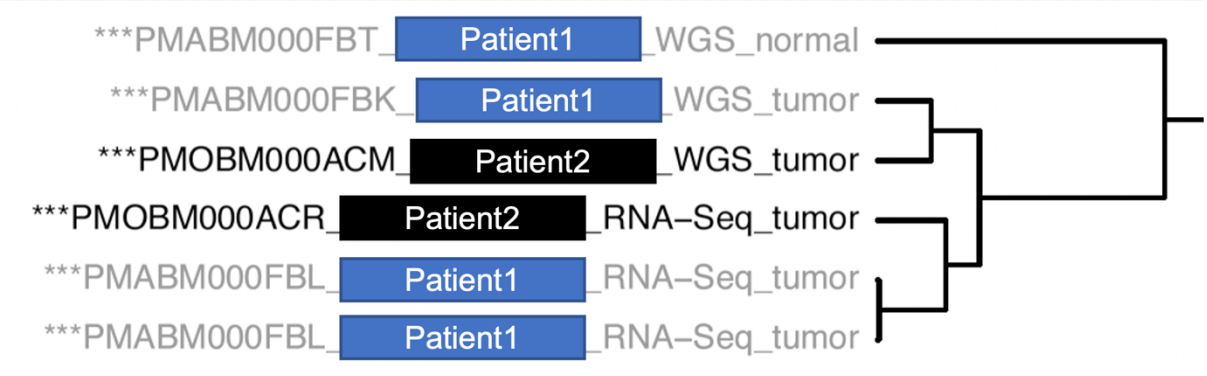

Figure 2: *Typical output of QCheckMate. Two different individual IDs are found in the same tree of the dendrogram, suggesting a potential sample swap. Moreover, the label shows the biomaterial ID, Individual ID, library strategy, and sample type.*

3. Results

3.1. VCF pre-filtering steps necessary for NGSCheckMate

A common problem observed in the PMC is that W[GX]S and RNA-Seq samples coming out of the sequencing pipeline are not matched by NGSCheckMate. For instance, NGSCheckMate predicted WGS and RNA-Seq from one individual to be a mismatch, as shown in the dendrogram in *Figure 3A*. To explain this unexpected result, variant allele fractions were scrutinised for which SNPs were detected in various samples. When comparing the variant allele fractions from NGSCheckMate and in the VCF-files, we noted that if a feature had no reads in a VCF-file, it was still interpreted as 0, which indicates a homozygous allele compared to the reference. Therefore, lines in VCF-files were filtered if the SAMPLE column equals ''./.:0,0:0:0'' or the QUAL column equals ''LowQual'', indicating zero reads and low-quality reads, respectively. To test the effect of the filtering step, both unprocessed and processed VCFs were run in NGSCheckMate. As the input dataset, the RMS subset was utilised, which contains a total of 111 WGS and RNA-Seq samples. Samples in the RMS set are high-quality and sample associations are well-curated.

Figure 3B shows two density plots of how many SNPs were called pre- and post-filtering. The number of SNPs detected for RNA-Seq were considerably lower after the filtering step, whereas the feature detection in WGS samples remained constant, indicating that in RNA-Seq a high number of SNPs were interpreted as homozygous even though no reads were detected for that certain feature. Consequently, filtering VCF-files prior to running NGSCheckMate improved integration of RNA-Seq and WGS samples in the RMS set, as depicted by the right dendrogram in *Figure 3A*. This is clearly visible with, for example, patient Patient1, as without filtering samples from different sequencing strategies were in a different tree, while WGS and RNA-Seq samples were found to have concordant genotypes after filtering lines in the VCF-files.

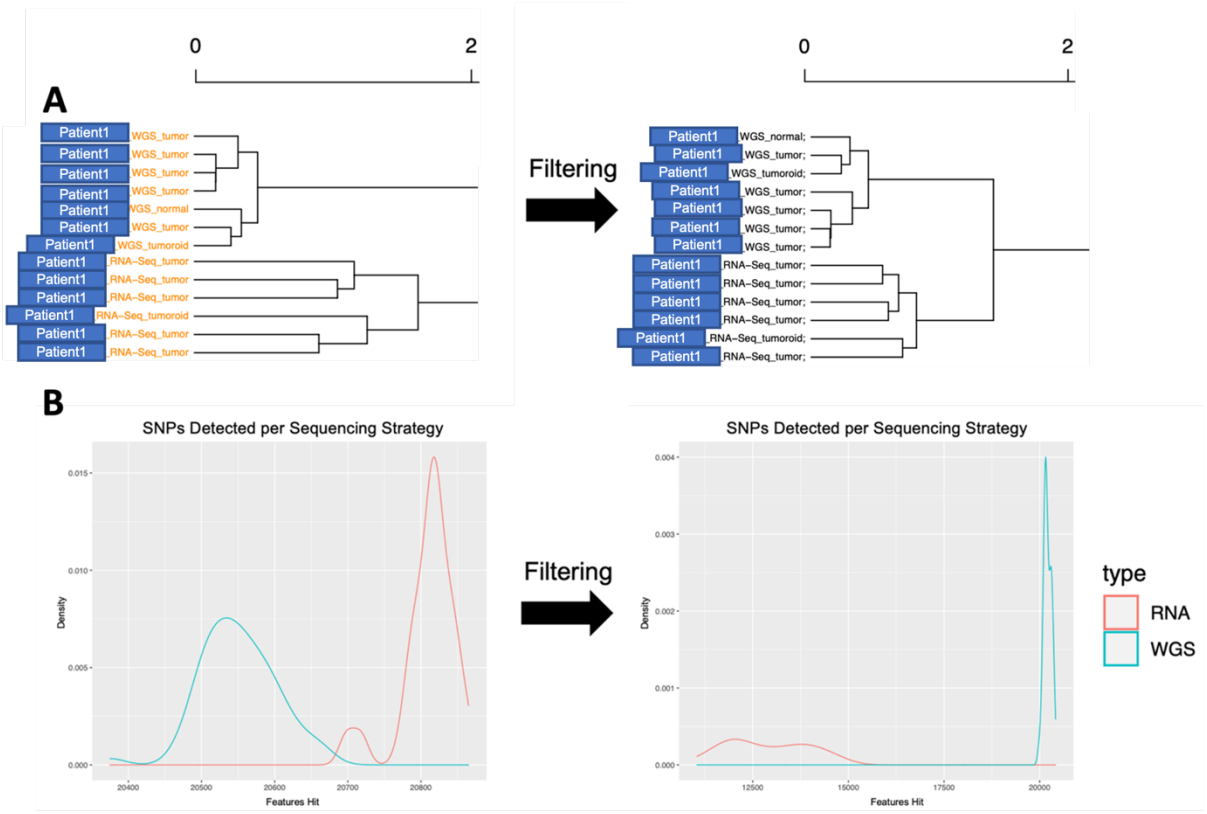

Figure 3: *Importance of filtering zero-reads and low-quality reads in VCF-files. A) Dendrogram on the left: Without filtering, RNA-Seq and WGS samples are classified in different trees. Dendrogram on the right: With filtering parameters applied, clustering of RNA-Seq and WGS samples is improved. Even though full integration of RNA-Seq samples with WGS is not achieved, identification of sample swaps would be feasible with this setup. B) Differences in features hit pre- and post-filtering, again highlighting the importance of filtering, since zero reads are interpreted by NGSCheckMate as 0 (indicating a homozygous allele).*

3.2. Filtering low-quality VCF-files

When running NGSCheckMate with smaller SNP sets, low-quality samples that are included in the analysis may potentially cause errors in the NGSCheckMate code. This is observed when the Pearson correlation cannot be calculated as no SNPs are called in the VCF-file, which causes a division by zero error (*Box 1*). Even though sometimes VCF-files in the database pass the QC-control in the sequencing pipeline, there are still files with too low quality in the dataset.

```
Traceback (most recent call last):
     File “ncm4.py”, line 1467, in <module>
          classifying()
     File “ncm4.py”, line 905, in classifying
          distance = pearson_def(vecA, vecB)
     File “ncm4.py”, line 59, in pearson_def
          return diffprod / math.sqrt(xdiff2 * ydiff2)
ZeroDivisionError: float division by zero
```

Box 1: *NGSCheckMate error when low-quality VCF-files are included in the analysis. ncm4.py is the NGSCheckMate script v4 (modified for the purposes of this paper). In the classifying() function, two vectors of variant allele fraction scores are compared with the Pearson correlation. ZeroDivisionError indicates that the math.sqrt(xdiff2 * ydiff2) equals zero, which causes an error as one cannot divide by 0. Error is caused when no features are detected in the analysis, which is possible when the bed-file includes a limited number of features and none of the SNPs included are detected in a sample.*

To investigate the number of callable genotypes, the number of callable SNPs per VCF-file was plotted in *Figure 4*, where the total number of lines in a VCF-file was plotted in a boxplot. The number of lines was used as a measure for quality, because the zero reads and low-quality reads were filtered out, indicating that in VCF-files with a low number of lines also covered a low number of SNPs. As can be appreciated from the boxplot, the number of lines in the VCF-files for WGS and WXS were comparable, above 20,000, indicating that most of the features of the 21K set were covered with sufficient confidence. Cut-off values were different for RNA-Seq and W[GX]S, as for RNA-Seq the lower tail of outliers was filtered out, whereas for W[GX]S the outliers lower than 19,500 were filtered out. This arbitrary cut-off was applied because VCF-files with a total number of lines larger than 19,500 can still be considered confident as still a lot of features are genotyped with sufficient confidence. Moreover, the higher tail of outliers for RNA-Seq (not found in current analysis) would also be sufficient for genotype-calling. The filtering step was most critical to transcriptome samples, as many samples were filtered out by these cut-offs. After this filtering step, no errors were encountered by NGSCheckMate, indicating that the filtering step was important for exclusion low-quality samples.

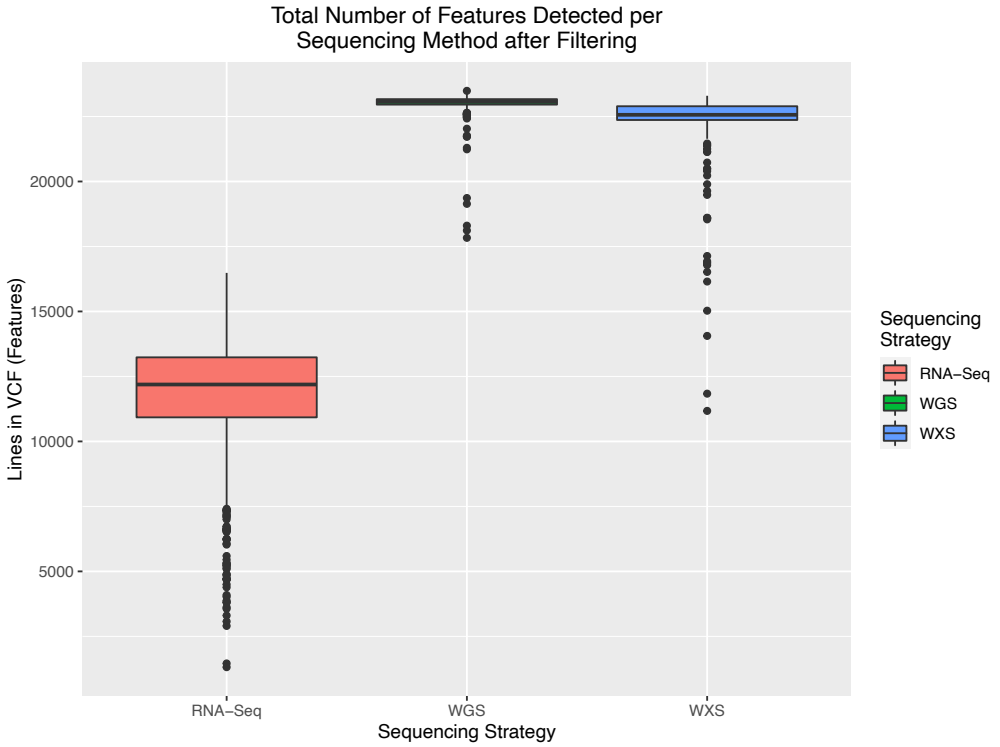

Figure 4: *Boxplot showing the total number of lines in a VCF-file per sequencing type after the filtering steps described in Section 3.2. Number of lines is used as a measure for sample quality, because more lines indicate more SNPs being detected with sufficient confidence.*

3.3. Feature set optimisation

3.3.1. Feature selection

In order to improve the integration of RNA-Seq with W[GX]S and decrease runtimes of NGSCheckMate, a feature set optimisation was conducted. The size of a SNP set is hypothesised to have a strong effect on runtime. Two parameters were selected for assessment whether a feature was included or not: 1) SNP detection across sequencing strategies, and 2) variant allele fraction spread (standard deviation) per feature. To prevent dataset bias on the feature set, instead of the RMS set, a random set with a total of 1000 randomly selected samples was used from all available samples in Trecode, making it a pan-cancer dataset. From the standard NGSCheckMate code[11] (ncm.py via GitHub), alterations were made in the script for getting an output file which lists all variant allele fractions per sample included in the analysis. Inside the script, if a feature is not detected in a VCF, the score is set to "None". Supplemental with this manuscript, the altered script is given (ncm4.py). Changes in the code include the change from 0 to "None", but also the print of output files for the VAFs per sample, SNPs covered per sample, and total number of features. These output files were added to the script for the feature set optimisation.

For the SNP detection parameter, the detected features were counted across samples (n = 1000). Features that had a VAF-score were given the value 1, whereas samples that had "None" were given the value 0. From these values, the sum was calculated to examine the overall detection of a feature (score ranging from 0 – 1000). The second parameter (VAF standard deviation) was obtained by calculating the standard deviation of all VAF scores from a feature. To select for the best performing features, a weighted score was derived from both parameters, according to the following formula:

$$weighted\_score = \frac{detectionScore\_feature}{detectionScore\_max} * \frac{VAFStDev\_feature}{VAFStDev\_max}$$

Equation 1: *Equation for the weighted score of a SNP in the feature set optimization. Where detectionScore_feature is the number of times a feature is observed, detectionScore_max is the maximum detection score across all features, VAFStDev_feature is the standard deviation of all VAFs of a feature, and VAFStDev_max is the maximum standard deviation across all features.*

The calculation of the weighted score for a single SNP is described in further detail in *Equation 1*. These weighted scores across all features are then ranked, consequently, the highest scores are chosen by picking the topN features, where N is the desired feature set size. In *Figure 5A*, an example in shown where the feature set size is 4,000 features. With these selection criteria, the most abundant and most varying minor allele are selected in the feature set. From this rank-based approach, multiple feature sets were created (500, 1K, 1.5K, 2K, 2.5K, 3K, 3.5K, 4K, 4.5K, 5K, 7.5K, 10K, 12.5K, 15K, 17.5K), which were further examined for several parameters.

Remarkably, full integration of RNA-Seq samples with W[GX]S was not possible with the created SNP sets. However, this was achieved by a SNP set that is based on the highly covered genes in RNA-Seq samples, a set consisting out of only 241 SNPs (Ellen van de Geer, unpublished work). A part of the dendrogram including RNA-Seq and WGS samples for the RMS set is visualised in *Figure S2*, showing that integration between WGS and RNA-Seq is possible. From now on, this set will be referred to as the Fingerprint set. To focus more on SNP coverage in RNA-Seq samples, the dataset was filtered to only include RNA-Seq samples, and the weighted analysis was performed with exclusively RNA-Seq samples (n = 410). To only include the most abundant SNPs in the optimisation, a filter step was performed, which selected only SNPs that were present in at least 80% of samples, leaving only 7681 features in the analysis. After filtering, selection of features was conducted with the same formula as described above. For comparison with the included features in the Fingerprint set, an intersect was taken for features that were both included in the Fingerprint set as well as the feature optimisation. As can be observed in *Figure 5B*, most features that are also included in the Fingerprint set are in the upper right corner of the scatter plot, which is in accordance with the selection criteria for which SNPs are selected in the feature set optimisation. The overlap between both the Fingerprint set and the optimised set is in the upper right corner (SNPs with highest variability and highest coverage). This finding indicates that SNPs selected in the Fingerprint set will probably also be selected in our feature set optimisation (dependent on the size of the SNP set).

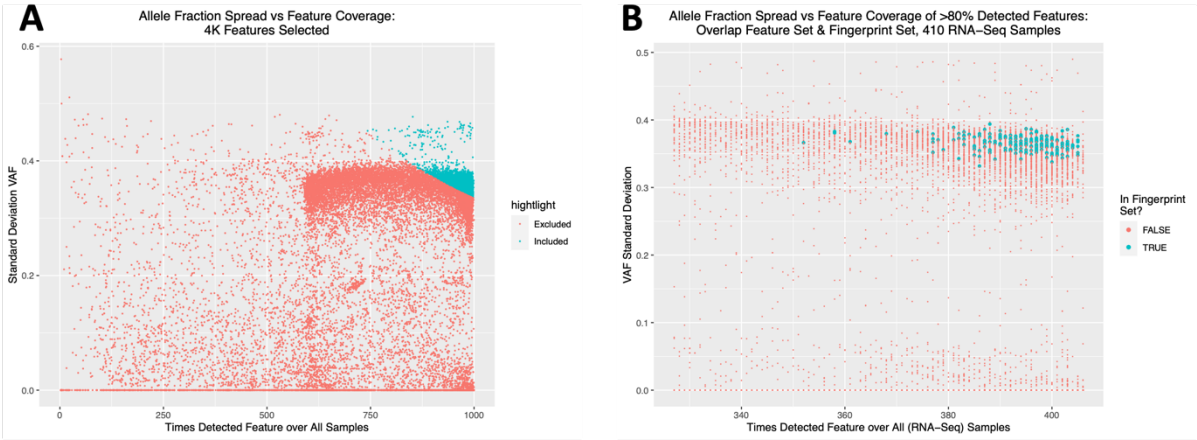

Figure 5: *Feature set optimisation derived from 21K SNP set. A) Weighted selection of 4,000 (4K) SNPs based on parameters variabilty of VAF-scores and coverage in all samples. As can be observed, SNPs in the upper right corner are selected for inclusion in the optimised set. B) Parameters VAF-score variability and coverage in all samples plotted based on coverage in RNA-Seq samples (n = 410, SNP must be covered in at least 80% (328) of all RNA-Seq samples). On top of that, overlap with the Fingerprint set is plotted, with most SNPs in both sets in the upper right corner, showing the resemblance in both selection procedures.*

3.3.2. Feature performance

After creation of multiple feature sets based on their coverage in NGS (and most predominantly RNA-Seq) samples, the sets were tested for certain parameters: The runtime

of NGSCheckMate per feature set and cluster performance of a feature set. Firstly, the runtime of all feature sets was examined, for this the ranked selection on all randomly picked 1000 samples (W[GX]S and RNA-Seq) was used to also include feature sets larger than 7.5K SNPs. Because on heterogeneous HPC clusters runtimes may vary dependent on which compute-node the job is run, each SNP set was run in triplicate. In *Figure 6A*, average runtimes of each feature set are visualised. As anticipated, runtimes increased with larger feature sets, showing that decreasing the feature set size is essential for limiting runtime on the HPC.

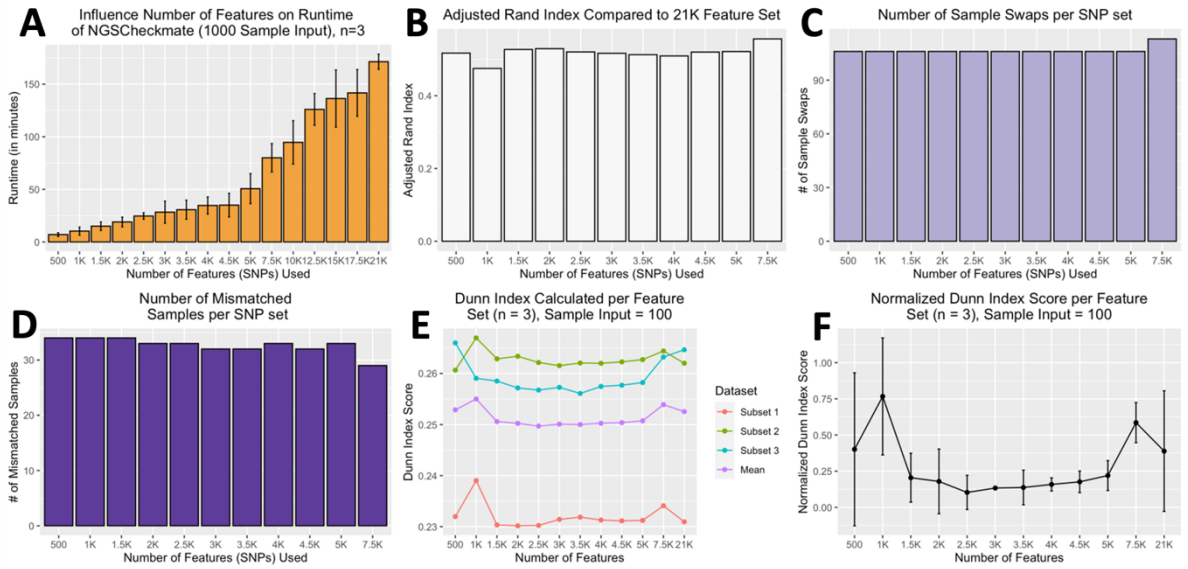

Figure 6: *Checking performance of different SNP sets. A) Influence of the SNP set size on runtime. B) Adjusted Rand Score for all SNP sets selected on RNA-Seq experiment. C) Number of sample swaps calculated per SNP set. D) Number of sample mismatches, indicating a sample of an individual is found elsewhere in the dendrogram than its respective samples. E) Absolute and F) normalised Performance of SNP sets based on Dunn Index. Performance is highly varying for some SNP sets (500, 1K, and 21K), whereas between 2.5K and 5K performance is determined ot be more stabilised.*

Alongside a shorter runtime, clustering performance is essential for implementing a feature set into the pipeline. Therefore, several cluster assessment tools were examined in order to test for performance of different sets. Firstly, the Adjusted Rand Index was calculated for all feature sets, which is a measure for similarity between two data clusterings[15]. Rand score calculations are always done compared to a reference dataset, in this case, we utilised the clustering results of the 21K feature set as the ground truth. Since one of the issues with the 21K feature set is the misclustering of RNA-Seq and W[GX]S samples from single individuals, only W[GX]S samples were included which were known to cluster sufficiently with the use of the 21K feature set. Adjusted Rand indices were constant for the transcriptome-based feature sets up to 7.5K (*Figure 6B)* Even though a score of approximately 0.4 might not be convincing, clustering of W[GX]S samples with the optimised sets showed comparable a constant number of sample swaps and sample mismatches for all RNA-Seq based SNP sets (*Figure 6CD*).
Another method for assessing cluster efficiency is the Dunn index, which considers both the smallest distance between two clusters as well as the largest distance within a cluster.

Between those two values, a ratio is calculated, which is informative to assess how different clusters are from each other. However, an issue to the Dunn index is the low maximum number of samples it can handle, limiting our approach which includes all W[GX]S and RNA-Seq samples (5851 samples from 1781 patients). Therefore, from the whole PMC dataset including all samples, the individuals with most samples were included to limit the number of clusters and detect how well known hierarchical sample relations are displayed. The analysis was performed in triplicate, and optimised sets were compared to the 21K dataset. In *Figure 6E* the performance of all sets is plotted together with the mean. Small differences can be observed between different subsets, particularly between subset 1 and subset 2/3. To account for these differences, data was normalized according to the following formula:

$$x_{normalised} = \frac{x - x_{minimum}}{x_{maximum} - x_{minimum}}$$

Equation 2: *Equation used for normalization of data for Dunn Index calculations. x is the mean Dunn Index value for a SNP set, the $x_{minimum}$ is the lowest mean score across all SNP sets, the $x_{maximum}$ is the highest mean score across all SNP sets.*

The formula scores all values between 0 and 1, where 0 is the lowest Dunn index score, while 1 is the highest Dunn index score. Highest average Dunn Index scores were found for the 500, 1K, 7.5K and 21K set, but with a high standard deviation (*Figure 6F*) An important note is that the lowest and highest Dunn index score were 0.23 and 0.27, respectively, which is a relatively small difference, indicating that the normalised graph might be biased towards and show larger differences than calculated.

Furthermore, an additional method for measuring cluster efficiency is the Silhouette Index, where each cluster is represented by a so-called silhouette. Silhouette plots show which objects fit well in a cluster, as well as the objects that are merely somewhere in between clusters[16]. The average silhouette width is the average width of all silhouettes, ranging from -1 – 1. With an increasing average silhouette width, clusters are better separated from each other. To investigate the performance the WGS/WXS/RNA-Seq and RNA-Seq based SNP sets from *Section 3.3.1.* silhouette widths were calculated for Subset 1, a dataset that was also used for Dunn Index calculations. Silhouette widths were determined at: WGS/WXS/RNA-Seq 500/4K set = 0.96, RNA-Seq 500/4K set = 0.97, original 21K set = 0.95, indicating that the performance of these sets was consistent, and selection based on RNA-Seq had no additional effect. After that, a Silhouette plot was constructed for the RMS set for 500, 4K and 21K RNA-based SNP sets, resulting in average silhouette widths of respectively 0.93, 0.92, and 0.71 (*Figure 7*). Noticeable is that for the 21K feature set the Silhouette plot predicted 21 clusters, while 15 clusters were predicted for the 500 and 4K feature sets. The improved resolution for the smaller sets was also visualised in the dendrograms produced by QCheckMate (RNA-Seq samples in *Figure S3)*.

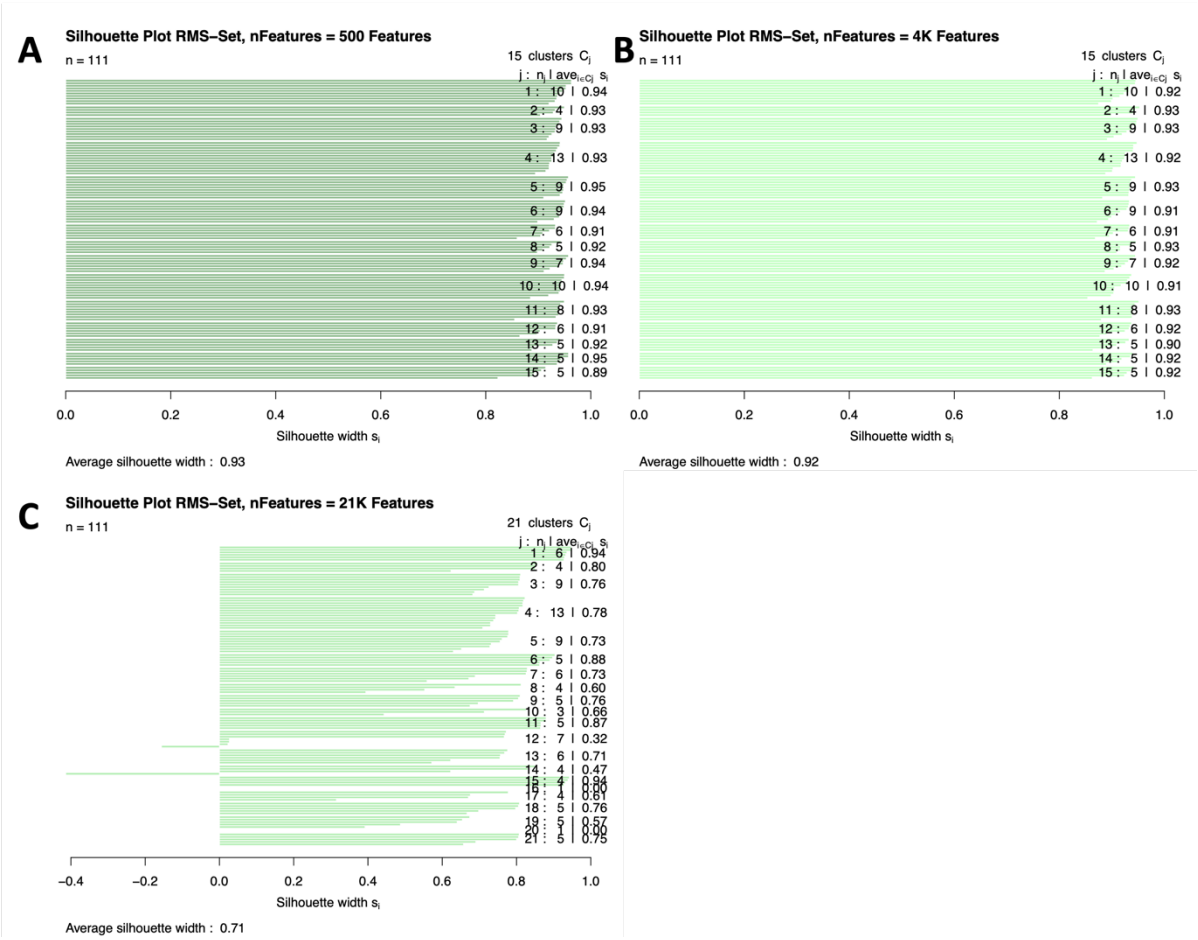

Figure 7: *Silhouette plots of RMS set in combination with different SNP sets: A) 500, B) 4K, and C) 21K. Average silhouette widths were determined at 0.93, 0.92, and 0.71 for 500, 4K, and 21K, respectively. Noticeable is that for the 21K SNP set the silhouette plot predicts 21 clusters, whereas for the other two SNP sets only 15 clusters are predicted, in accordance with the number distinct patient identifiers in the analysis.*

To assess the SNP sets in a more translational method to the clinic, all samples after QC-control were included in the analysis (5769 samples, 1781 PMC IDs). Average silhouette width and number of clusters were calculated for the 500, 1K, and 4K RNA-based and the original 21K SNP set. Average silhouette widths were calculated at respectively 0.85, 0.41, 0.85, and 0.82, showing a large drop in clustering efficiency for the 1K set. Also, other parameters were calculated for the different feature sets, this data is shown in *Table 1*.

Feature Set	*Ave. Sil. Width*	*# Pred. Clusters*	*IDs w/ Swap*	*Trees w/ Swap*
500	0.85	1792	143	76
1K	0.43	3190	121	65
4K	0.85	1739	141	75
21K	0.82	1835	136	73

Table 1: *Several parameters on cluster efficiency were calculated. The 500, 1K, 4K, and 21K feature sets were included. Total number of patient identifiers in analysis is 1781, total number of samples included in the analysis is*

5769. Parameters include average silhouette width, number of clusters predicted by silhouette plot, total number of PMC IDs (patients IDs) with at least one sample swap, and total number of trees with at least one sample swap.

3.4. Potential applications NGSCheckMate

3.4.1. Assessment of sample quality within datasets

Performance of different SNP sets was tested for NGSCheckMate, for this purpose, different datasets were utilised (e.g., performance on RMS set). One of the potential applications of NGSCheckMate was observed when an in-house dataset for Wilms' tumour samples was entered in NGSCheckMate. This dataset was examined, as it might provide additional information since the Wilms' tumour dataset is not as well annotated as, for instance, the RMS dataset. For the Wilms' tumour set, one of the samples was found to be different than other samples from the same patient ID *Figure 8*. The researcher that curated this dataset was informed about this case and was asked to provide information on this sample. Consequently, a clear explanation for this mismatch was found. The WGS sample of this biomaterial was excluded because the data quality was debatable, however, the RNA-Seq sample was still included as it passed the QC check. As a result, it can be concluded that the not only the WGS sample was of insufficient quality, indeed, the biopsy taken from the patient was supposedly of poor quality.

3.4.2. Assessment of association between samples

NGSCheckMate could potentially be used to scrutinise sample associations, for example, to determine the similarity between tumour and normal samples. Furthermore, the homogeneity between tumour and tumoroid could also be studied, providing information how related a tumoroid is to its respective tumour source. For the organoid dataset, a well-described medulloblastoma dataset, no sample swaps were expected. In *Figure 9* can be observed that there is a clear distinction between tumour and normal samples. Furthermore, the researcher who curated this dataset was requested for additional information on individual samples. For patient Patient1, biomaterial PMOBM000AEM was a tumoroid derived from a primary patient sample, with the same biosource as tumour sample with ID PMLBM000AND. This known relationship corresponds with the dendrogram in *Figure 9*, since these samples are close in the tree. This pattern of biomaterials corresponding with neighbouring objects in the dendrogram was also true for Patient2 and Patient3. Only one tumoroid sample of Patient3 had a low correlation with other samples from the same individual (biomaterial PMOBM000AED, WGS), therefore, this sample was inspected in further detail. Copy number plots of this specific individual (Patient3) were checked. As can be observed in *Figure S4*, the gains/losses in copy number of the tumoroid are not in line with its respective tumour sample, likely explaining why a lower correlation is calculated between these two samples.

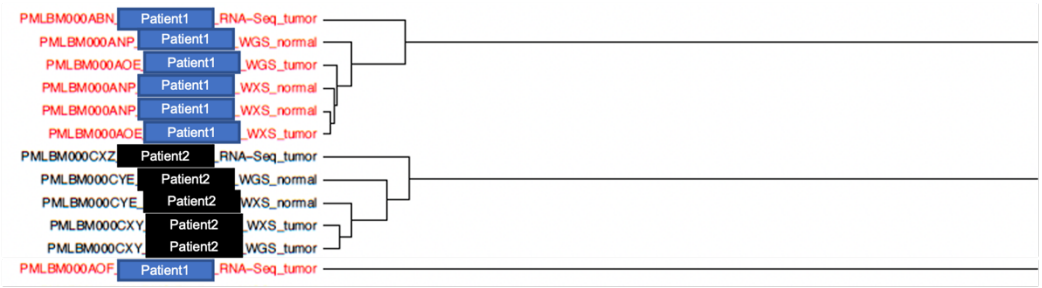

Figure 8: *Part of dendrogram of the Wilms' tumour dataset, pre-processed with the 500 SNP set. Samples in red in the first and the final tree, are clearly separated and not related, even though samples are supposedly derived from the same individual. Original dendrogram was adjusted for purposes of this paper (unrelated samples were cut off the dendrogram for clarification).*

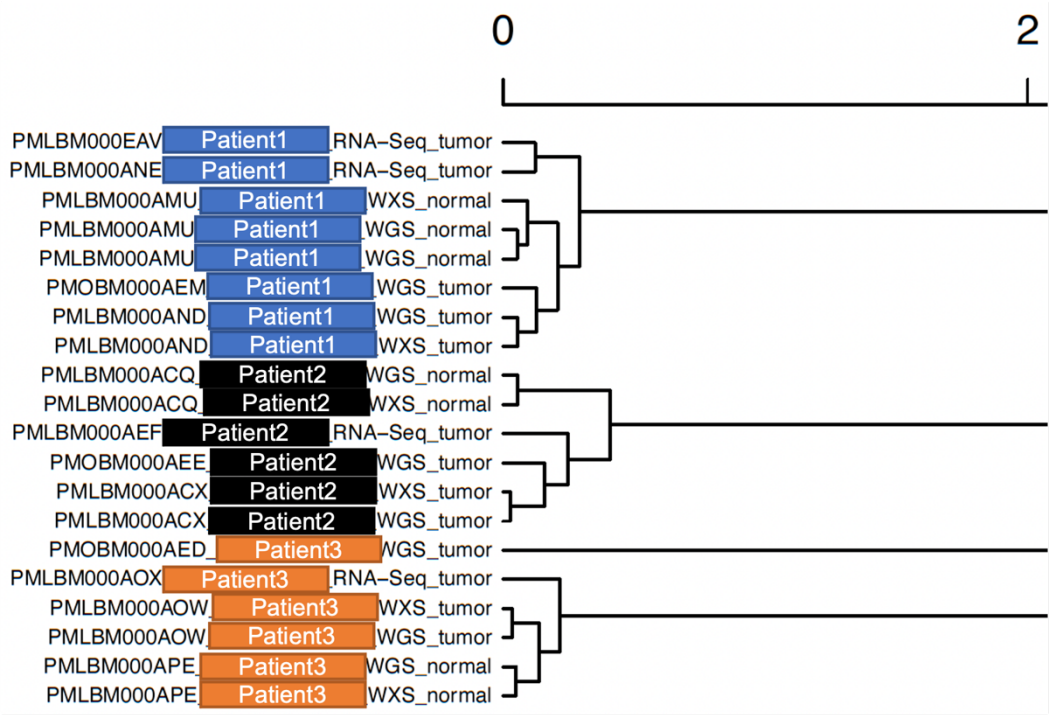

Figure 9: *Dendrogram of organoid dataset, pre-processed with the 500 SNP set. As expected, no sample swaps were found. More importantly, a clear distinction can be observed between tumour and normal samples.*

4. Discussion

Nowadays, sample swaps are still a common phenomenon within hospitals, including the PMC. Given that treatment is often informed by this data, QC-control and data integrity is vital for clinical applications. Besides clinical use, in a research institute such as the PMC, data assurance is crucial for conducting sound research. Therefore, NGSCheckMate was embraced, a program which assesses which samples have concordant genotypes based on a set of SNPs. However, the SNP set in the from the developers is relatively large, namely 21K SNPs. In combination with a low number of samples it is feasible to check for this number of point mutations, however, once the sample input starts increasing, the program approaches computational limits. Consequently, we hypothesised that even with a smaller SNP set, it would still be possible to retrospectively call genotypes, something that was already supported by findings of the Kemmeren Lab with the Fingerprint set. Even though the Fingerprint set showed integration between RNA-Seq and WGS, once larger datasets were run in NGSCheckMate in combination with the SNP set, a loss of resolution was observed. Apart from the long runtime of NGSCheckMate, when entering W[GX]S and RNA-Seq data from the PMC in NGSCheckMate, samples from one individual are not classified into the same tree of a dendrogram, indicating that samples have a low VAF correlation. The approach of this study involved closely inspecting NGSCheckMate code, monitoring in-house data quality, and optimising the SNP set.

Prior to optimising the feature set, our findings suggested that filtering low-quality data was essential for properly clustering genome/exome data together with transcriptome data. For privacy reasons, data coming out of the pipeline in the PMC contains only 21K SNP fingerprints derived from the biobank genomics variant calling results. This format was found not to be compatible with NGSCheckMate, as zero-reads SNPs are also included in the files and were interpreted as homozygous (VAF = 0) by NGSCheckMate. Hence, filtering these features from VCF-files was important for finding concordant genotypes, as RNA-Seq does not cover the entirety of the exome, which led to misinterpretation of zero-reads in NGSCheckMate. After filtering the files that were entered in NGSCheckMate, the number of SNPs that were covered were plotted per sequencing strategy. This plot suggests that most of the RNA-Seq samples cover around 10K – 14K SNPs, which gives are first indication of the size of the optimised SNP set. Nevertheless, it is important to note that different SNPs are called per sample. If the intersect of these SNPs across samples is taken, the SNP set size might decrease. Another aspect that should be accounted for is that these results are derived from the RMS set, which consists of high-quality samples, stipulating that the data quality of other RNA-Seq samples might be worse. Indeed, it was found that some samples were of such poor quality, that NGSCheckMate had issues running once zero-read/low-quality

features were filtered (*Box 1*). Accordingly, quality of files was scrutinised after the initial feature filtering, leading to the finding that even though some samples pass the QC-check, data quality can still be low, shown in *Figure 4*.

Moreover, within the code of NGSCheckMate, standard output of a SNP is 0, suggesting a homozygous allele, however, this does not always hold true. For the feature set optimisation, it was important to distinguish between zero-reads and homozygous features, therefore, alterations in the code were made: In the Python code, the standard VAF value was changed from 0 to “None”. Notably, when selecting SNPs based on the coverage in all sequencing strategies, limited integration between W[GX]S and RNA-Seq samples was found. Moreover, when SNPs were selected were selected only on RNA-Seq, no additional improvement was observed. This result was unexpected because of earlier findings, as most genome/exome samples already most features are covered (*Figure 4*), and therefore, selection criteria can be less stringent on these library strategies. Indeed, the 21K set of the authors of NGSCheckMate is based on exonic SNPs, which explains why also in WXS samples most of the SNPs are covered. On the contrary, transcriptome samples have limited coverage of features, which supports the selection of SNPs based on detection in RNA-Seq. On top of that, the findings described in *Figure 5B* imply that our selection is in line with the findings of previous work on the Fingerprint set, in which SNPs were selected based on their theoretical expression in RNA-Seq samples. SNPs with a high VAF variability and a high coverage are found in the upper right corner of the scatter plot, in this area, most overlapping SNPs from the Fingerprint set are detected. Nevertheless, in high-quality datasets, sufficient clustering of samples and improved integration of multiple library strategies was observed.

Followed by the creation of different size sets, performance was checked. Several methods were used to assess clustering efficiency: the Adjusted Rand Index, Dunn Index, and Silhouette Index. Especially the last method (Silhouette Index) was informative on larger datasets. Results from the Adjusted Rand Index and Dunn index suggest the performance of optimised sets is similar, with the note that the Adjusted Rand Index was only based on W[GX]S samples, and in addition, the score was relative to the 21K set. By visual inspection the dendrograms per set (*data not shown*), a clear distinction between patient samples was visible. However, it is hard to determine whether a high resemblance with the 21K set is positive, since there was no golden truth for clustering all W[GX]S. Possibly, some SNP sets might perform better than the 21K set but have a lower Adjusted Rand Index because of the lower resemblance with the 21K set. Nevertheless, it can be concluded that none of the RNA-Seq based SNP sets were found to be completely different from the 21K, indicating that proper sample clustering was performed. A limitation of utilising the Dunn Index as a parameter for cluster efficiency was the restricted number of samples that could be included in the analysis.

Performances of SNPs were constant, only the 500, 1K, 7.K, and 21K set had a higher performance (with a high standard deviation). Nevertheless, it should be noted that the absolute difference between all scores is small, and it could be argued whether these results are distinct enough to differentiate between the performance of SNP sets.

The Silhouette Index was found to be a suitable measure for testing cluster performance of different SNP sets, as not only the Silhouette Index was calculated, but also the number of clusters was predicted. Firstly, Silhouette plots were created for the RMS set, a dataset which could still be assessed by visual inspection. Remarkably, the resolution of the clusters was lower for the dendrogram created with the 21K set. This finding was confirmed in the Silhouette plots in *Figure 7*, where Silhouette plots of the RNA-based SNP sets had higher Silhouette value. The lower resolution of the 21K set is further emphasised by the predicted number of clusters, namely 21. On the contrary, for the other two SNP sets, the same number of clusters were predicted as the number of patients in the analysis (15), suggesting a sufficient clustering of samples. Ultimately, Silhouette Indices were calculated for a pan-cancer set, which consisted of all samples (W[GX]S + RNA-Seq) available in the dataset. In *Table*, all findings are visualised. From these results, the number of clusters for the 500 set is the closest to the actual number of patient IDs (1792 to 1781, respectively). The 4K set had less clusters predicted than the actual number of patient IDs (1739 to 1781, respectively), which indicates that samples from different patients clustered together. Remarkably, the 1K set was found to have approximately twice as much clusters as patients included in the analysis, an observation which underlines the need for a manual check whether the SNP set performed well after using NGSCheckMate.

Even though the results of *Figure 9* imply that sample associations are visible within dendrograms, this is not always true for all datasets. Visual inspection of dendrograms often showed tree structures in which W[GX]S and RNA-Seq samples were found in the same tree, however, some results were comparable to the dendrogram in the right panel of *Figure 3A*. In this dendrogram there is a clear separation between W[GX]S and RNA-Seq. A possible explanation could be that data quality is crucial for visualising sample associations within a dendrogram, for example, in datasets as the tumoroid and RMS sets. This hypothesis is supported by the fact that even within the large pan-cancer dataset for some patient samples, regardless of the library strategy, clear associations between identical biomaterials are observed. Furthermore, in *Figure 9* and *Figure S4*, clear associations between biomaterials can be observed. Even when a low correlation between a tumour and a tumoroid were calculated, this could be explained by a biological cause. In the case of Patient3, the copy number plots of this specific individual showed that there were considerable differences between the gains/losses of the tumour and its respective tumoroid. This result could be

explained by artifacts of tumoroid selection, suggesting that the tumoroid does not represent its respective tumour sample to a certain extent.

One of the anticipated risks of using a drastically trimmed SNP set for calling genotypes was the loss of resolution, an observation we first encountered with the Fingerprint set. However, our results suggest that this is not the case for feature sets as small as 500 SNPs, this was supported with the use of the Dunn Index and Silhouette Index (*Figure 6EF* & *Figure 7*). For instance, based on the Silhouette plots on the RMS set it can be observed that the average Silhouette width is increased while using a smaller SNP set (*Figure 7*). In line with these findings, it was also concluded that the loss of resolution of the Fingerprint set was not a result of the SNP set size, but rather an issue of data quality (*data not shown*). Also, visual inspection of the dendrograms produced by QCheckmate showed a higher resolution for the 500 and 4K sets. Future studies might provide even smaller SNP sets, since the limits of feature set size have not been reached yet.

In conclusion, the use of a smaller SNP set shows promising results in both decreasing runtime and increasing integration of RNA-Seq with W[GX]S data. In *Figure 10*, the improvements in runtime duration per sequencing strategy are visualised. Differences in sequencing strategies could be explained by the computing-node on which the job was run, but overall, the runtime was considerably decreased for all sequencing strategies. Indeed, runtimes for feature sets smaller than 5K SNPs were found to be short and therefore potential sets to use. Moreover, clustering performance is found to be constant for these feature sets, even for small SNP sets such as the 500 set. Therefore, our results suggest that a SNP set smaller than 5K SNPs could be applied for retrospectively checking for sample swaps in an institute such as the PMC. With such a SNP set, runtimes on the HPC are decreased, multiple library strategies were found in one tree of a dendrogram, and in high-quality datasets integration of RNA-Seq was improved to visualise sample associations.

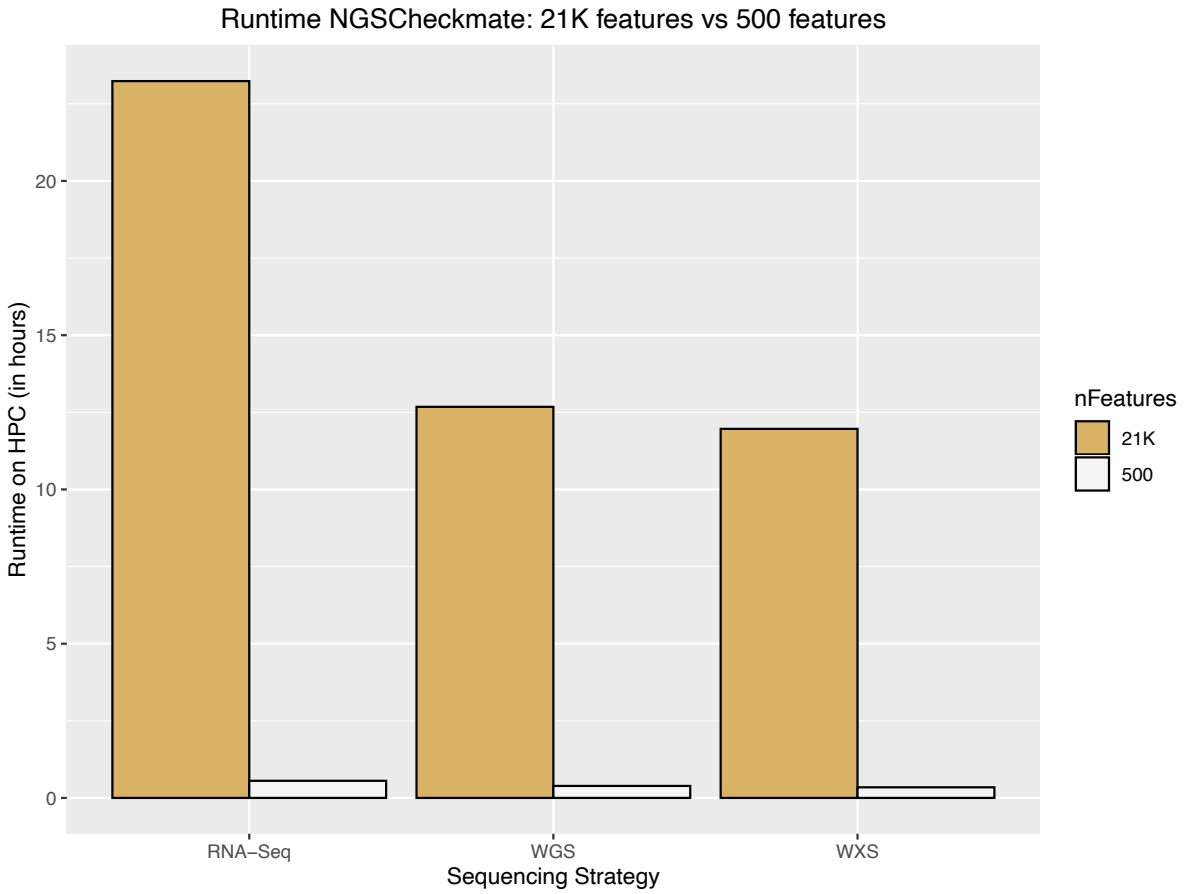

Figure 10: *Differences in NGSCheckMate runtime on the HPC per sequencing strategy (including all samples of that were available for that sequencing strategy on 07/10/21, RNA-Seq = 2252, WGS = 1797, WXS = 1721). With the 500 SNP set, RNA-Seq was ~42x faster, WGS was ~33x faster, and WXS was ~35x faster.*

5. Conclusion

In this study, we present an alternative SNP set that can be utilised in the workflow of NGSCheckMate. Depending on the sequencing strategy, runtimes on the HPC were decreased as much as 42x compared to the original SNP set. Integration of RNA-Seq was limited but improved with the optimised SNP sets. Compared to the original 21K set, resolution of clusters increased when utilising optimised SNP sets. With these optimised feature sets, correct identification of sample swaps was achieved. On top of that, for well-curated and high-quality datasets, strong associations between identical biomaterials were observed, regardless of the sequencing strategy.

Moreover, it was found that pre-processing and additional filtering VCF-files is vital for clustering genome/exome and transcriptome data together. A limitation of retrospectively checking for sample swaps is the durability of a program such as NGSCheckMate, as runtimes keep increasing with a growing amount of data. Nevertheless, this SNP set could be used in at least the coming years, as runtimes are still computationally feasible. In conclusion, SNP sets below 5K show viable runtimes and proper clustering of samples from the PMC, with the smallest SNP set of 500 SNPs tipping the performance, therefore, our findings recommend replacing the original SNP set with our optimised feature set.

6. References

1. Siegel, R. L., Miller, K. D., Fuchs, H. E. & Jemal, A. Cancer Statistics, 2021. *CA. Cancer J. Clin.* **71**, 7–33 (2021).
2. Our story - Prinses Máxima Centrum. https://www.prinsesmaximacentrum.nl/en/about-us/our-story.
3. Eiser, C. Long-term Consequences of Childhood Cancer. *J. Child Psychol. Psychiatry Allied Discip.* **39**, 621–633 (1998).
4. Goudoever, H. van. Concentrating childhood cancer treatment in the Netherlands. *Padiatr. Padol.* **50**, 38 (2015).
5. Kleijer, K. & Litjens, F. *Annual Report 2020 - Prinses Maxima Center*. (2020).
6. Byron, S. A., Van Keuren-Jensen, K. R., Engelthaler, D. M., Carpten, J. D. & Craig, D. W. Translating RNA sequencing into clinical diagnostics: opportunities and challenges. *Nat. Rev. Genet. 2016 175* **17**, 257–271 (2016).
7. Kerstens, H. H. *et al.* Trecode: a FAIR eco-system for the analysis and archiving of omics data in a combined diagnostic and research setting. *bioRxiv* 2020.11.13.363689 (2020) doi:10.1101/2020.11.13.363689.
8. SH, K. & JK, W. Characterization of the standard and recommended CODIS markers. *J. Forensic Sci.* **58 Suppl 1**, (2013).
9. Huang, J., Chen, J., Lathrop, M. & Liang, L. A tool for RNA sequencing sample identity check. *Bioinformatics* **29**, 1463–1464 (2013).
10. Pengelly, R. J. *et al.* A SNP profiling panel for sample tracking in whole-exome sequencing studies. *Genome Med. 2013 59* **5**, 1–7 (2013).
11. Lee, S. *et al.* NGSCheckMate: Software for validating sample identity in Next-generation sequencing studies within and across data types. *Nucleic Acids Res.* **45**, e103 (2017).
12. AEM van Belzen, I. *et al.* Systematic discovery of gene fusions in pediatric cancer by integrating RNA-seq and WGS. *bioRxiv* 2021.08.31.458342 (2021) doi:10.1101/2021.08.31.458342.
13. Van der Auwera, G. A. *et al.* From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr. Protoc. Bioinforma.* **43**, (2013).
14. Sherry, S. T. *et al.* dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* **29**, 308–311 (2001).
15. Santos, J. M. & Embrechts, M. On the Use of the Adjusted Rand Index as a Metric for Evaluating Supervised Classification. *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)* **5769 LNCS**, 175–184 (2009).
16. Rousseeuw, P. J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **20**, 53–65 (1987).

7. Layman's summary

Nowadays, in a research hospital such as the Princess Máxima Centre (PMC), research and patient treatment is often substantiated on Next Generation Sequencing (NGS) data. Therefore, quality control is crucial for preserving data integrity. However, several steps in the process from patient sample to genotype are vulnerable to human errors. For instance, researchers can introduce typos, samples are switched, and wrong filenames are copy pasted, resulting in sample swaps. For prevention of these errors, NGSCheckMate was introduced, a tool which can retrospectively check for sample swaps. NGSCheckMate checks how often a mutation is found on a certain specific part of the genome, such a position is called a single nucleotide polymorphism (SNP). The authors of NGSCheckMate have generated a SNP set which consists out of 21K SNPs.

NGSCheckMate is already implemented in the PMC, but running all samples available in the database added up to ~68 hours and is therefore not computationally efficient. On top of that, it was found that samples coming out of the sequencing pipeline within the PMC biobank were not compatible with NGSCheckMate, as there was limited integration of samples from different sequencing strategies. We hypothesised that by selectively choosing SNPs based on variety of the minor allele and on coverage across samples, the SNP size could be reduced, leading to a smaller runtime with comparable performance. The approach of this study involved closely inspecting NGSCheckMate code, monitoring in-house data quality, and optimising the SNP set.

Our findings suggested that genome and exome samples covered the majority of the SNPs in the original SNP set, nevertheless, the transcriptome samples had a limited coverage of SNPs in the feature set. For this reason, SNPs were selected based on the coverage in RNA-Sequencing samples. This resulted in the creation of multiple subsets of the original 21K set. Several parameters were tested to assess the performance of these subsets, including the Adjusted Rand Index, the Dunn Index, and the Silhouette Score. These different parameters showed that the performance of the smaller subsets was not only comparable to the 21K set, indeed, some subsets showed an increased performance. With the use of these smaller sets, it was found that integration of transcriptome with genome/exome samples was improved. Besides this, runtime of NGSCheckMate was decreased from ~68 to ~2 hours, making it more computationally efficient.

In conclusion, this study presents a range of smaller subsets of the original 21K SNP set. These subsets increase the performance of NGSCheckMate and decrease the runtime, making NGSCheckMate more sustainable to be included in the standard quality control process within the PMC.

Supplemental figures

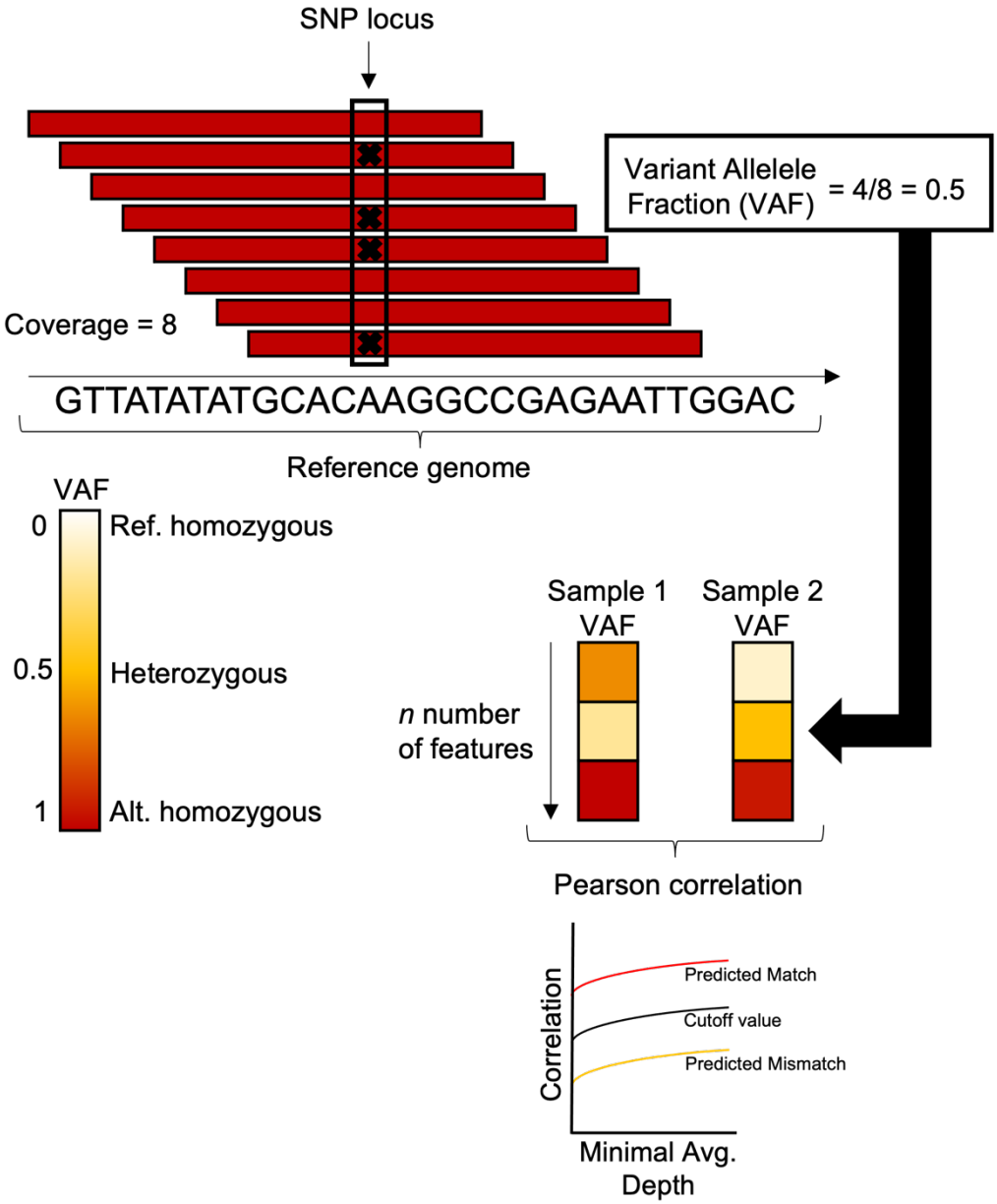

Figure S1: *Overview of workflow NGSCheckMate. NGSCheckMates checks for all SNPs that are included in the SNP set. For each SNP, the VAF is calculated, which is the total number of alternative reads divided by the total number of reads. All SNPs of a sample are then stored in a vector, which can be compared to another sample's vector another sample. Between these sectors, the intersect is taken and the Pearson correlation is calculated. This Pearson correlation value determines whether two samples are a match or not. The cut-off values are based on a TCGA stomach cancer set (n = 40).*

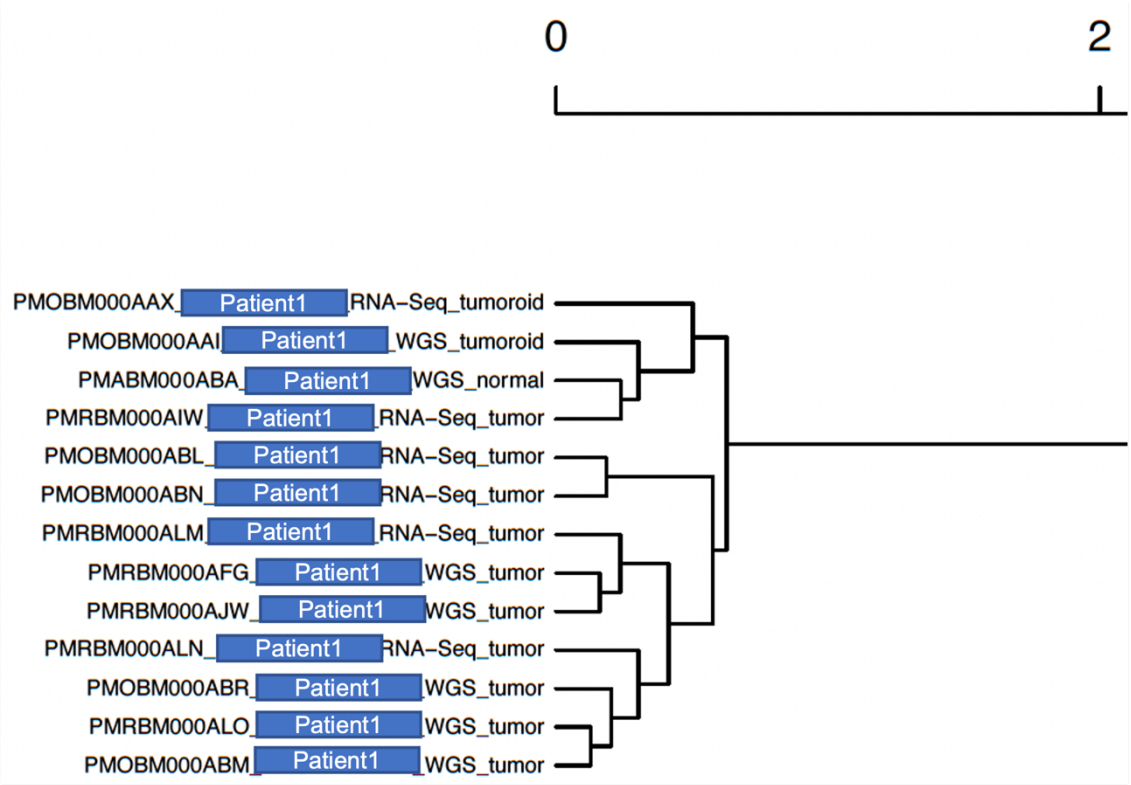

Figure S2: *Dendrogram of RMS set clustered by NGSCheckMate with the use of the Fingerprint set (Ellen van de Geer, unpublished work), consisting out of 241 SNPs. As can be appreciated from the dendrogram, integration of RNA-Seq samples with WGS is achieved.*

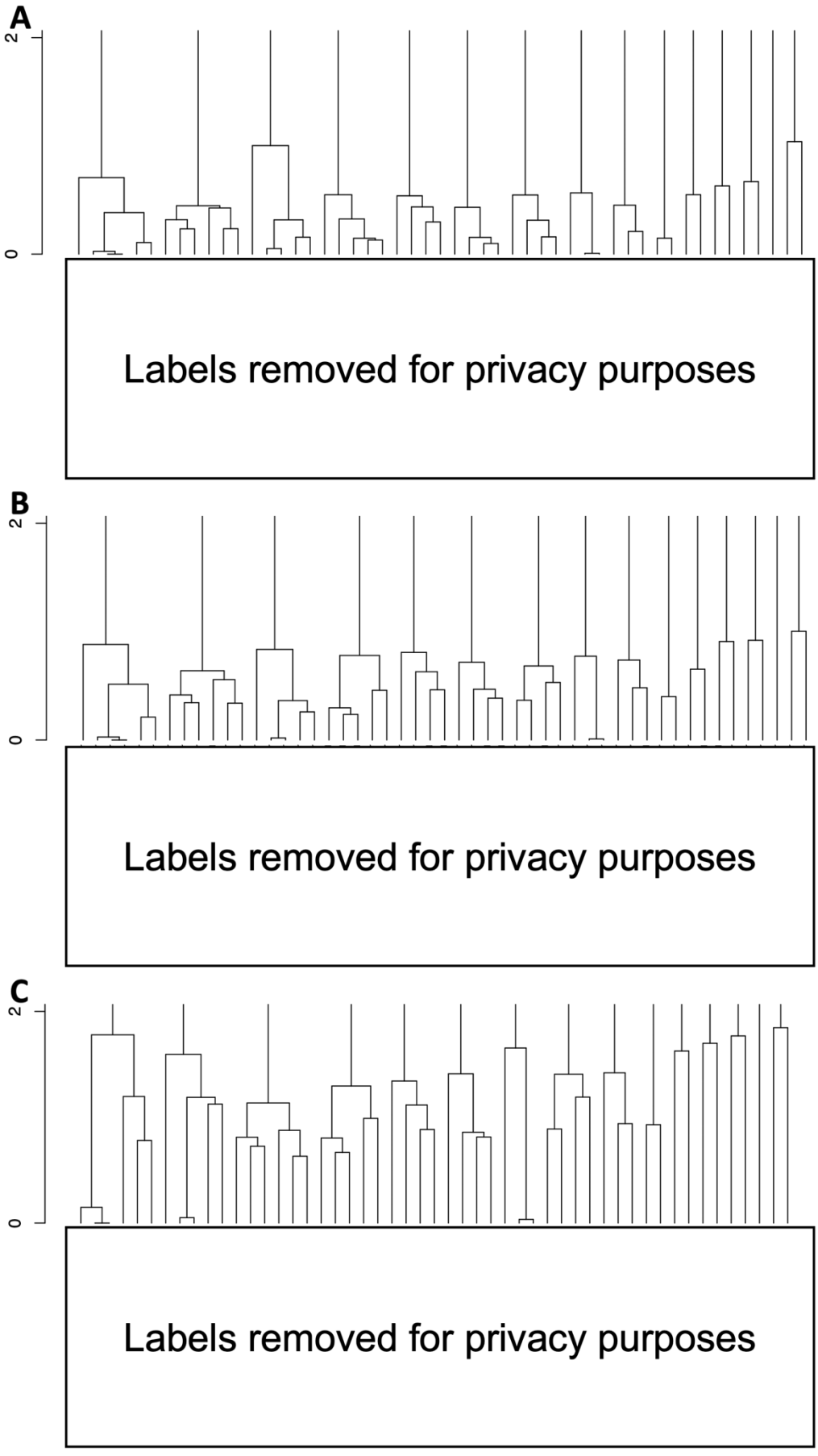
A
2
0
Labels removed for privacy purposes
B
2
0
Labels removed for privacy purposes
C
2
0
Labels removed for privacy purposes

Figure S3: *Dendrogram of RMS RNA-Seq data. Only RNA-Seq data is taken to limit data input. Increased resolution can be observed for the A) 500 and B) 4K optimised compared to the C) original 21K set. When additional WGS samples are added to the dendrogram, the same resolutions are observed (WGS/RNA-Seq data can be requested). Labels were removed for privacy purposes, nevertheless, the loss in resolution is still visible.*

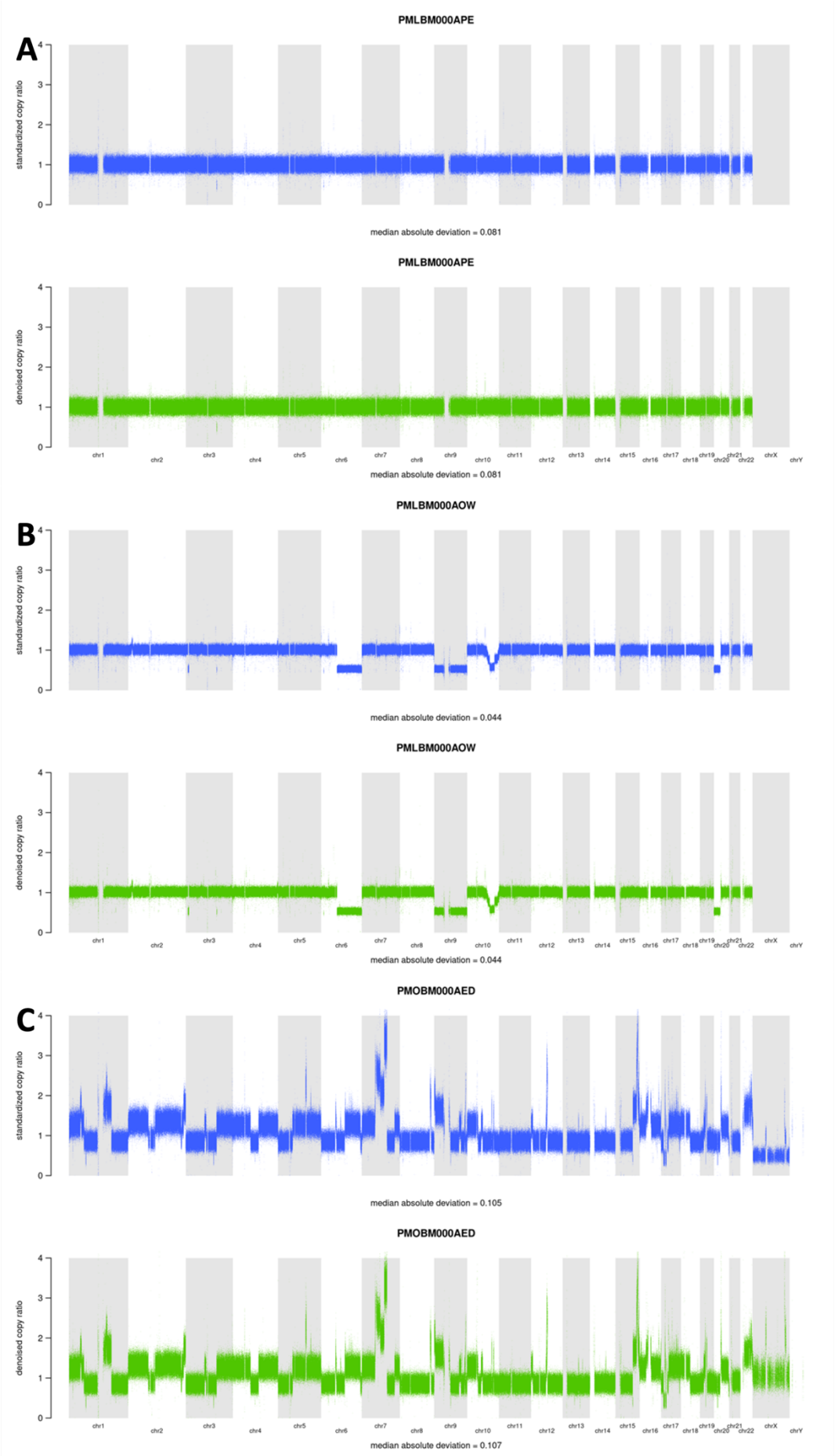
A
PMLBM000APE
standardized copy ratio
median absolute deviation = 0.081
PMLBM000APE
denoised copy ratio
median absolute deviation = 0.081
B
PMLBM000AOW
standardized copy ratio
median absolute deviation = 0.044
PMLBM000AOW
denoised copy ratio
median absolute deviation = 0.044
C
PMOBM000AED
standardized copy ratio
median absolute deviation = 0.105
PMOBM000AED
denoised copy ratio
median absolute deviation = 0.107
chr1 chr2 chr3 chr4 chr5 chr6 chr7 chr8 chr9 chr10 chr11 chr12 chr13 chr14 chr15 chr16 chr17 chr18 chr19 chr20 chr21 chr22 chrX chrY

Figure S4: *Copy number plots of Patient3. A) PMLBM000APE is the normal sample, B) PMLBM000AOW is the tumour sample, and. C) PMOBM000AED is the tumoroid. As can be observed in the plots, the gains/losses of the tumoroid sample do not resemble its respective tumour sample.*