Bioinformatic strategies for the analysis of genomic aberrations detected by targeted NGS panels with clinical application

Jakub Hynst[1,2,3], Veronika Navrkalova[1,2], Karol Pal[1,4] and Sarka Pospisilova[1,2,3]

[1] Center of Molecular Medicine, Central European Institute of Technology, Masaryk University, Brno, Czech Republic
[2] Department of Internal Medicine-Hematology and Oncology, Faculty of Medicine and University Hospital Brno, Masaryk University, Brno, Czech Republic
[3] Department of Medical Genetics and Genomics, Faculty of Medicine and University Hospital Brno, Masaryk University, Brno, Czech Republic
[4] Department of Hematology, University Hospital Schleswig-Holstein, Kiel, Germany

ABSTRACT

Molecular profiling of tumor samples has acquired importance in cancer research, but currently also plays an important role in the clinical management of cancer patients. Rapid identification of genomic aberrations improves diagnosis, prognosis and effective therapy selection. This can be attributed mainly to the development of next-generation sequencing (NGS) methods, especially targeted DNA panels. Such panels enable a relatively inexpensive and rapid analysis of various aberrations with clinical impact specific to particular diagnoses. In this review, we discuss the experimental approaches and bioinformatic strategies available for the development of an NGS panel for a reliable analysis of selected biomarkers. Compliance with defined analytical steps is crucial to ensure accurate and reproducible results. In addition, a careful validation procedure has to be performed before the application of NGS targeted assays in routine clinical practice. With more focus on bioinformatics, we emphasize the need for thorough pipeline validation and management in relation to the particular experimental setting as an integral part of the NGS method establishment. A robust and reproducible bioinformatic analysis running on powerful machines is essential for proper detection of genomic variants in clinical settings since distinguishing between experimental noise and real biological variants is fundamental. This review summarizes state-of-the-art bioinformatic solutions for careful detection of the SNV/Indels and CNVs for targeted sequencing resulting in translation of sequencing data into clinically relevant information. Finally, we share our experience with the development of a custom targeted NGS panel for an integrated analysis of biomarkers in lymphoproliferative disorders.

Subjects Bioinformatics, Genomics, Hematology, Oncology
Keywords Bioinformatic analysis, SNV/indel, CNV, Clinical application, Molecular markers, NGS, Targeted panels

Submitted 9 September 2020
Accepted 13 January 2021
Published 31 March 2021

Corresponding author
Sarka Pospisilova, sarka.pospisilova@ceitec.muni.cz

Academic editor
Kumari Sonal Choudhary

Additional Information and Declarations can be found on page 17

DOI **10.7717/peerj.10897**

How to cite this article Hynst J, Navrkalova V, Pal K, Pospisilova S. 2021. Bioinformatic strategies for the analysis of genomic aberrations detected by targeted NGS panels with clinical application. **PeerJ 9:e10897 DOI 10.7717/peerj.10897**

INTRODUCTION

Recent progress and growing application use of next-generation sequencing (NGS) technology, alongside the reduction of costs, has revealed new prospects in the field of personalized medicine. Researchers and clinical laboratories worldwide are implementing NGS to identify defects in the cancer genome to improve patient stratification and treatment. These genomic aberrations are represented by single nucleotide variants (SNVs), small insertions and deletions (Indels), copy number variants (CNVs) and structural variants (SVs), which accumulate in the genome during tumor development. Some of them are present at the time of diagnosis, while others occur as a consequence of clonal evolution during the disease course (*Wang et al., 2014*; *Landau et al., 2015*). Over the past years, it has been demonstrated that NGS is a unique tool for identifying new genomic variants (*Armaou et al., 2009*; *Ascierto et al., 2012*; *Lindsley et al., 2015*; *Zoi & Cross, 2015*), which can serve as important diagnostic and prognostic markers in various cancer types. In the field of hematooncology, rapid adoption of NGS had an enormous influence on the understanding of the genetic landscape, clonal evolution and prognostic impact of new molecular markers during the disease course (*Landau et al., 2015*; *Pastore et al., 2015*; *Nadeu et al., 2016*; *Papaemmanuil et al., 2016*; *Dubois et al., 2016*).

An expanding catalogue of molecular markers with clinical importance can be analyzed by targeted DNA sequencing of relevant regions in a fast, effective, and accessible manner (*Paasinen-Sohns et al., 2017*). While targeted sequencing enables the detection of various alterations in recurrently affected genomic regions, alternative NGS techniques, such as whole-exome sequencing (WES) or whole-genome sequencing, can be used to identify additional disease-related markers outside the areas of targeted assays (especially genome-wide CNVs and SVs). Nevertheless, a higher cost and an overwhelming amount of produced data (*Metzker, 2010*) requiring extensive and time-consuming bioinformatic analysis limit their usage in routine diagnostics (*Kuo et al., 2017*). The use of targeted NGS panels has indeed proven to be a financially feasible approach, providing benefits in cancer patient management (*Hamblin et al., 2017*). Moreover, larger targeted panels (>1 Mb) with the adjustments of design and cutoff values (*Allgäuer et al., 2018*; *Heydt et al., 2020*) can nowadays substitute the use of WES for the estimation of tumor mutation burden, which serves as a surrogate predictive marker in cancer (*Rizvi et al., 2015*). In general, a variety of commercial cancer-specific panels is available and widely used to detect genomic changes in different cancer types (*Nikiforova et al., 2018*; *Steward et al., 2019*). However, an off-the-shelf approach may not always be appropriate for clinical use as these panels may include several genes without established clinical importance (e.g., genes investigated in clinical trials or research studies) or may lack other genes of interest. A trend towards customization of targeted panels to fulfil the needs of individual laboratories is evident. The decision whether to use a commercial NGS panel or whether to put effort and labor into the design, validation and development of a respective bioinformatic pipeline strongly depends on the clinical utility, available bioinformatic support and financial and time capabilities of individual laboratories.

The implementation of a reliable bioinformatic pipeline developed with respect to the experimental approach is a major challenge for many clinical laboratories. Although some targeted panels have been published together with their tailored pipelines (*Kluk et al., 2016*; *Soukupova et al., 2018*), additional validation is necessary for each pipeline to confirm or adjust panel specifics including the accuracy and sensitivity. The process of tailored pipeline development starts with comprehensive literature and software survey followed by an in-depth evaluation of results produced by every single bioinformatic step. Despite several published best practice bioinformatic guidelines (*Van der Auwera et al., 2013*; *Gargis et al., 2015*), the development of a pipeline is still a time-consuming and laborious procedure. The resulting pipeline has to be highly reliable, adjusted to specific laboratory needs and flexible to demands changing over time.

In this review, we discuss the possibilities of a custom targeted NGS panel implementation with a particular focus on bioinformatics. We emphasize state-of-the-art approaches for the identification of genomic aberrations from DNA NGS data to make the process of bioinformatic pipeline development more transparent. Finally, we share our experience with the evaluation of a custom panel designed for a comprehensive analysis of genomic markers in lymphoproliferative disorders with the emphasis on bioinformatic tools assuring accurate results.

REVIEW METHODOLOGY

The motivation behind the compilation of this review was our hands-on experience with the implementation of a custom targeted NGS panel for lymphoproliferative disorders and the scarcity of bioinformatic publications accompanied by practical experience in the development of specific bioinformatic pipelines in clinical use. Relevant and highly impacted articles from 2009 to the present, spanning the field of bioinformatics, cancer genomics and hematooncology, were scrutinized and systematically reviewed.
The bibliography was created using the Zotero citation manager.

EXPERIMENTAL DESIGN REMARKS

During the design of a custom targeted NGS panel, issues such as intended use, panel size with respect to anticipated coverage, clinical validity and utility must be considered. Each laboratory should think over its facilities, time and cost demands, sample turnaround, flexibility and bioinformatic support. It is also essential to determine the spectrum of targeted aberrations before selecting a target enrichment approach and an NGS platform. Generally, targeted NGS gene panels are designed to identify variants such as SNV/Indels and are either limited to only well-described hotspot mutations in clinically relevant genes (especially in routine diagnostics) or include whole coding sequences and splice sites. In addition, larger panels may target selected CNVs and SVs requiring more sophisticated bioinformatic analyses. For the detection of subclonal aberrations, a high sequencing depth is necessary. Crucial recommendations for NGS panel and bioinformatics pipeline validation were published by the Association for Molecular Pathology and College of American pathologists (*Jennings et al., 2017*; *Roy et al., 2018*).

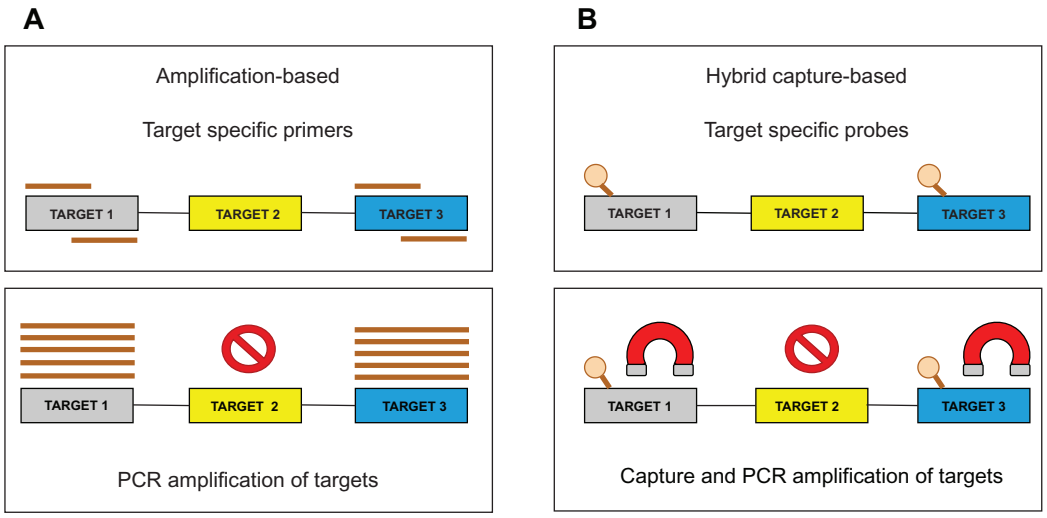

Figure 1 Target enrichment approaches for NGS library preparation. (A) The amplification-based method with the use of the PCR primers. (B) The hybrid capture-based method utilizing target-specific probes. Full-size DOI: 10.7717/peerj.10897/fig-1

Target enrichment is an essential step in NGS testing. For sequencing library preparation, two major approaches are used: (1) hybrid capture-based and (2) amplification-based (Fig. 1). Several studies describe the benefits and drawbacks of both methods (*Sulonen et al., 2011*; *Chilamakuri et al., 2014*; *Gargis et al., 2015*; *Kuo et al., 2017*). The amplification approach is based on the multiplex polymerase chain reaction (PCR) method, which requires short hands-on time, but also precise PCR optimization to produce uniform amplification efficiency across all targets. Moreover, PCR amplification is less efficient in regions with high guanine-cytosine (GC) content or in repetitive regions. Importantly, the identification of larger Indels or chromosomal rearrangements is generally complicated since these aberrations could span over the location of a PCR primer. In the hybrid capture approach, sequence-specific probes are designed to catch DNA fragments of interest. These biotinylated oligonucleotides are significantly longer than PCR primers and can, therefore, tolerate the presence of several mismatches resulting in an effective enrichment process. Generally, it has been shown that the capture-based methods show better performance than amplification-based ones, with respect to sequencing complexity and uniformity (*Samorodnitsky et al., 2015*). Besides, the amount of captured DNA is proportional to the DNA present in a sample allowing CNV detection by a "read depth" approach. The method is also less affected by DNA quality, enabling the analysis of such biological materials as formalin-fixed, paraffin-embedded (FFPE) blocks (*Hung et al., 2018*) or circulating free DNA (cfDNA) (*Rossi et al., 2017*). Several commercial targeted NGS technologies are available in custom design and their comparison performed by *Samorodnitsky et al. (2015)* could serve as an informed decision for individual laboratory applications.

The occurrence of duplicated DNA fragments generated during the amplification step in library preparation represents a common issue in NGS data analysis. It is difficult to determine which sequences originate from genomic DNA and which are a product of PCR

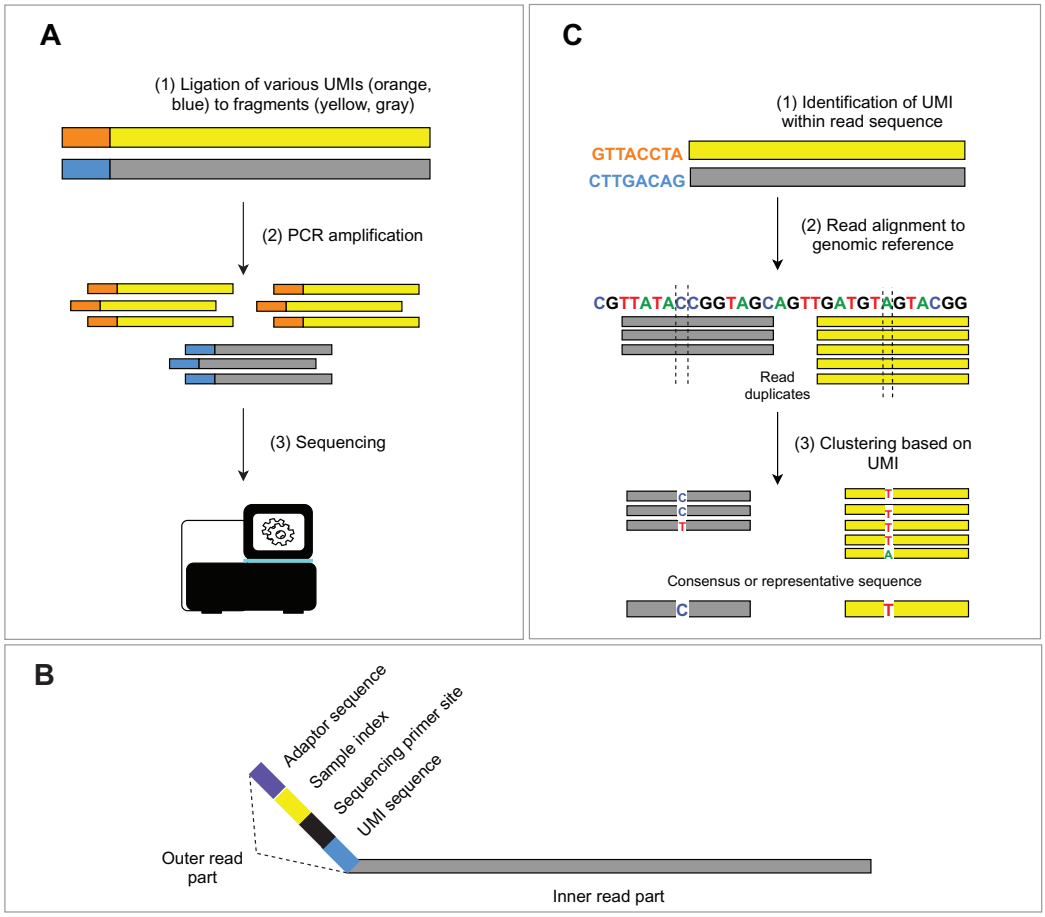

Figure 2 Schema of UMIs experimental and bioinformatic processing. (A) During NGS library preparation, UMIs are ligated to DNA fragments, followed by PCR amplification and sequencing. (B) Structure of the outer part of a read is depicted. (C) Bioinformatic read deduplication and error correction using UMI identification and read clustering. Depending on the selected tool, deduplication results either in a consensus read creation or the selection of the most representative read.

Full-size DOI: 10.7717/peerj.10897/fig-2

amplification. Unique molecular identifiers (UMIs) (usually 8–12 bp long), ligated to genomic fragments before the first PCR amplification, help solve this problem (Fig. 2A). Nonetheless, the incorporation of the UMIs increases the cost and uses up some of the assay capacity. The consequent bioinformatic analysis requires additional solutions to utilize UMIs in PCR duplicate recognition. Subsequent steps of deduplication and the creation of consensus sequences lead to the in silico removal of sequencing duplicates and thus increase the sensitivity of the assay. We discuss bioinformatic solutions for read deduplication with the use of UMIs in the bioinformatic part.

BIOINFORMATIC DATA ANALYSIS

Bioinformatic workflow management

In any NGS experiment, a bioinformatic pipeline needs to fulfil several analytical requirements determined by the assay design. In other words, the analysis has to ensure the

accuracy and reproducibility of results under individual experimental settings, which is especially relevant in routine diagnostics. Both properties are likewise essential when investigating the clonal expansion of aberrations across different time points during the disease course. Proper workflow management helps fulfil these requirements by ensuring software installation and versioning, organization of input and output data, and resource management during the execution of a pipeline.

Software (package) management can be ensured by the Conda package management system. Conda provides a unified method of software installation and concurrently allows the management of individual projects within specific environments, ensuring consistency and stability. The *Bioconda channel* (*Grüning et al., 2018*) specializes in tools used in bioinformatic analysis. The execution of several programs chained through their inputs and outputs is managed by workflow engines such as Snakemake (*Köster & Rahmann, 2018*), Bpipe (*Sadedin, Pope & Oshlack, 2012*), reflow (https://github.com/grailbio/reflow) and nextflow (*Di Tommaso et al., 2017*), which also offers a set of curated pipelines under the nf-core project (*Ewels et al., 2020*). The "chaining" provides efficient administration of input and output data while monitoring used processes and resources. The workflow engines are managed through text-based configuration files.

Alternatively, the open-source Galaxy framework (*Afgan et al., 2018*) provides a user-friendly graphical interface and thus caters to users with little or no prior programming experience. It contains a comprehensive repository of preinstalled software for genomic, proteomic, or metabolomic data processing, which can be easily managed and combined into an analytical pipeline. A public Galaxy server provides computational resources sufficient for experimenting and small analyses with more than 124,000 registered users worldwide.

Yet another approach is represented by the bcbio-nextgen project (*Chapman et al., 2020*), which offers a resource manager (for both tools and data resources) as well as a collection of curated, reusable, highly optimized pipelines. On top of this, pipeline execution is highly customizable through configuration files. This solution shifts a considerable part of the workload on the community behind the development and maintenance of bcbio and is suitable for common applications of NGS.

Finally, one of the most recent yet already widely used technologies are containers such as docker (*Merkel, 2014*) and singularity (*Kurtzer, Sochat & Bauer, 2017*). They allow for a self-contained environment, in which analyses can be run. More importantly, the benefit of encapsulation lies in the ability to distribute a given container, which significantly improves the reproducibility of analyses. Essentially, all current pipeline managers support containers, and containers are in exchange supported by a wide range of underlying architectures, including cloud services mentioned in the following chapter.

Hardware requirements

Appropriate hardware infrastructure is required to perform bioinformatic data analysis. It is important to understand the demands on hardware equipment because insufficient capacity could potentially lead to poor or erroneous results (*Wade, Curtis & Davenport, 2015*). For targeted NGS panel data analysis in diagnostic use, the basic hardware essentials

are: (1) sufficient space to store pipeline inputs and outputs including intermediate or temporary files, (2) enough computational capacity (number of CPUs/GPUs and RAM size) and importantly, (3) security for sensitive data management.

There are several external options, which fulfilled the above-mentioned criteria. Commercial Amazon Web Services' Elastic Compute Cloud (*Amazon, 2020*) provides scalable computational and storage resource to perform a range of bioinformatics analyses. Non-commercial institutional clouds such as European Grid Infrastructure (EGI) (*EGI, 2020*) or US The Open Science Grid (*OSG, 2020*) provide free advance computing services for scientists and research infrastructures. However, it is often advantageous to administrate local hardware for better control over computational resources, data security, and software management. The scale of this computational facility for each lab performing NGS experiments is individual, depending on sample turnaround, the number of ongoing projects and previous experience. In the last chapter, we showcase and further discuss the measurements of the consumption of computational resources of an exemplary analysis of our custom NGS panel. In general, a minimal analytical pipeline, akin to the one presented here, consisting of only the essential steps (read alignment, variant calling and variant annotation) could potentially run on a laptop equipped with at least 8GB of RAM. As these crucial steps undergo the most scrutiny (*Lenis & Senar, 2015*), they are highly optimized for speed and memory efficiency. Despite this possibility, the general trends favor commercial/institutional cloud services and/or large computational servers optionally with dedicated hardware such as GPU's and even FPGA's. Several such GPU implementations for both the alignment and variant calling have been published (*Klus et al., 2012*; *Cardoso et al., 2016*; *Ahmed et al., 2019*; *Ren et al., 2019*).

Sequencing reads preprocessing and alignment

Short, single or paired-end reads are produced by the Illumina sequencers, the most widely used sequencing platform. The read structure (Fig. 2B) consists of the inner part (the actual sequenced DNA fragment, so-called "insert") and the outer part comprising a variable combination of: (1) platform-specific sequences for binding the fragment to a flow cell (adaptor sequences), (2) single or dual indexes for read assignment to an individual sample, (3) sequencing primer sites for paired-end reads and optionally (4) molecular barcodes (i.e., UMIs) used for tagging the individual DNA molecules in a given sample library. Adaptor sequences and indexes are unnecessary for downstream analysis and are removed from the beginning of the read. Trimming is done during the demultiplexing step, where sequencing base calls are transformed into text-based read representation and stored in FASTQ formatted files. The result of the demultiplexing step is usually a pair of FASTQ files (R1 and R2, representing forward and reverse read) per sample. A third FASTQ file (R3) may be generated to store UMI sequences or, optionally, UMIs may be written in the read names of the R1 and R2 files. Finally, parts of the sequencing adaptors may also be present at the end of a read (given a short insert) and can also be removed from the read structure of both read pairs. This step is debatable and may be considered irrelevant as the adaptor sequences will be "soft-clipped" by the aligner. However, several variant callers are "soft-clip aware" (e.g., GATK's HaplotypeCaller

and Mutect2) and will try to realign the sequence around the soft-clipped bases. We, therefore, find adaptor (as well as quality) trimming generally advisable. Recently, more than 30 adaptor trimming tools were published, including the popular Cutadapt (*Martin, 2011*), Trimmomatic (*Bolger, Lohse & Usadel, 2014*) or the fastp (*Chen et al., 2018*) program.

Consequent quality control (QC) of preprocessed reads ensures sequencing data quality and integrity. FASTQ files contain (apart from the actual sequences of reads) read quality information represented by Phred Quality score (Q); base-calling error probability per each base. The FastQC tool can be used to summarize and visualize the Phred Quality score, GC content, sequencing into adaptors, the occurrence of overrepresented sequences and other parameters. Unsatisfactory reads with low Q scores can be trimmed or discarded from further downstream analysis by most of the standard trimming tools.

The preprocessed FASTQ files are subsequently mapped to the reference genome during the alignment step. The most recent human genome assembly is currently available in two, mostly interchangeable, versions; the GRCh38 (*Schneider et al., 2016*) reference (managed by the Genome Reference Consortium) and the hg38 reference (managed by UCSC). An older version of the reference hg19 (GRCh37) (*Church et al., 2011*) is still widely used. In general, the alignment is a fundamental and computationally demanding step, which allows the mapping of a sample onto the reference sequence (i.e., assigning genomic coordinates to each read). This process is error-tolerant, as mismatches between the sequencing reads and the reference may represent genomic variability and, more importantly, real pathogenic variants. The result of the alignment step is stored in a Sequence Alignment Map file or its more compact binary counterpart (BAM file). Within this format, each read contains additional information about genomic alignment coordinates, mapping quality information (MAPQ), splice alignment indicator (Compact Idiosyncratic Gapped Alignment Report, CIGAR) and more. MAPQ is often used to identify low-quality reads, which may adversely affect the identification of gene variants. The CIGAR string is a compressed representation of the alignment of an individual read, encoding which parts of the read match or mismatch the genomic reference and whether there are inserted or removed bases. MAPQ and CIGAR are helpful indicators for proper detection of real genomic variants as well as for determining alignment bias, which arises mainly from sequencing errors or repetitive regions. Freely available software for DNA read alignment such as NovoAlign (*Novocraft, 2020*), Bowtie2 (*Langmead & Salzberg, 2012*), Smalt (*SMALT: Wellcome Sanger Institute, 2020*) and Stampy (*Lunter & Goodson, 2011*) can be used, but they provide non-consistent results in various datasets (*Ruffalo, LaFramboise & Koyutürk, 2011*). Burrows-Wheeler aligner (BWA) is currently the most common software used for mapping (*Li & Durbin, 2009*), showing the best performance across multiple NGS datasets in alignment sensitivity, computational time, or the alignment of reads in repetitive regions (*Thankaswamy-Kosalai, Sen & Nookaew, 2017*). Therefore, BWA represents the preferred alignment tool for the majority of DNA based NGS techniques, including targeted panel assays.

Sample specific BAM files serve as input for the consecutive analytical branches, including the detection of SNV/Indels, CNVs and SVs. Before these analyses, PCR duplicates should be marked and removed from the BAM files to increase the accuracy of testing. This step is essential for the detection of low frequent somatic SNV/Indels and CNVs. The need for duplicates removal is amplified when analyzing fragmented low-quality DNA samples, for example, FFPE. It was shown that the overall duplication level in FFPE samples reached ~50–60% compared to less than 20% for DNA from fresh frozen tissue (*Bewicke-Copley et al., 2019*). To remove PCR duplicates, two approaches can be applied: (1) Picard tool *mark duplicates* (https://broadinstitute.github.io/picard/) is used to either tag or remove duplicated reads according to identical alignment coordinates of sequencing reads, or (2) use of UMIs processed by designated bioinformatics tools described in the following chapter.

UMI processing for in silico read deduplication

The employment of molecular barcodes for the labeling of different DNA fragments is not recent (*Jabara et al., 2011*; *Liang et al., 2014*) and is nowadays implemented in many NGS assays. This advanced experimental approach allows precise identification of PCR duplicates and sequencing error correction in silico. Specifically, dual indexing with UMIs resolves index swaps, increases the sensitivity of variant detection and reduces inaccuracies in read count-based analyses (e.g., gene expression analysis, CNV analysis) (*MacConaill et al., 2018*; *Costello et al., 2018*). Bioinformatic approaches for UMI processing mostly rely on read alignment and consist of three main steps (Fig. 2C): (1) UMIs are identified within the read structure, and corresponding read pairs (forward/reverse) are tagged with the sequence of the UMI, (2) position matched reads (i.e., reads mapping exactly to the identical location on the reference genome) sharing the same UMI are clustered into groups, (3) read clusters are then collapsed into a consensus sequence, or a representative sequence is selected. For example, UMI-tools (*Smith, Heger & Sudbery, 2017*) *dedup* approach uses the highest alignment quality (MAPQ) to choose the most representative read sequence, while the fgbio (*fgbio, 2020*) *CallMolecularConsensusReads* applies a likelihood model to each base of the source read molecule, which finally leads to a consensus read creation. Other commonly used software is gencore (*Chen et al., 2019b*) or Je (*Girardot et al., 2016*). Alternative approaches, like the Calib program (*Orabi et al., 2019*), utilize alignment-free clustering, which is more suitable for high coverage amplicon sequencing. Nevertheless, this approach does not take into account potential substitution errors in UMI sequences.

Evaluation of target enrichment efficiency

Coverage analysis is used to assess the efficiency of the primers or probes, which helps with the optimization of the target enrichment process. It is also essential to evaluate the uniformity of coverage across targets and sequencing runs to ensure maximum result reproducibility. Further, the ability to detect any genomic variants, correctly estimate VAF, and to reduce the false-positive rate at the same time improves with increasing depth and coverage uniformity (*Sims et al., 2014*). The effect of deduplication directly translates into

the decrease in coverage and can be observed by comparing coverage statistics before and after the removal of duplicates.

Should a more thorough examination of coverage, both on- and off-target, be required, bedtools (*Quinlan & Hall, 2010*) *coverage* provide per-base coverage information across the whole genome or specific regions specified within a BED file (*Ensembl., 2020*). Such comprehensive information can be used to get a table of statistics such as min/max read depth, mean or median across defined regions and thus enabling disclosure of probable off-target locations. Statistics regarding the distribution of aligned reads between the targeted and non-targeted regions are an important quality control metric of sequencing assay design. The proportion of on-target reads covering targeted genomic regions enables the assessment of enrichment efficiency. The assignment of roughly >70% of the reads to target regions is considered a good library indicator (*Hung et al., 2018*), but this strongly depends on panel size and overall depth of coverage. A lower amount of the on-target reads can indicate an abundant occurrence of repetitive or homologous sequences in the targeted regions resulting in poor enrichment.

Variant identification algorithms

Bioinformatic approaches for genomic variant identification are aimed to distinguish true biological variants from sequencing background. Dedicated programs are in abundance and, therefore, it is essential to understand underlying algorithms and parameters controlling their behavior to select the correct tool. Independent analytical branches identify different types of aberrations. We will predominantly discuss approaches and software for the detection of somatic and germline SNVs/Indels and CNVs. Furthermore, we will discuss different analytical procedures according to the experimental settings, for example, the availability of matched control (non-tumor or "normal") sample from the same individual. For SVs, specific probes or primers can be designed to enrich and detect those events (*McConnell et al., 2020*). We remark that the common division of CNVs and SVs to the distinct groups is debatable since SVs, in general, include quantitative CNVs (comprising deletions, insertions, and duplications), translocations and/or inversions (*Scherer et al., 2007*). However, alterations that do not change the copy number of the genome have to be detected by specific algorithms (*Guan & Sung, 2016*).

Variant calling of SNV/Indels

The general strategy for variant identification is the calculation of the proportion of non-reference bases in a batch of reads that cover each position in a targeted region. The analysis of germline variants is straightforward since their VAF is about 50% or 100%, and the level of sequencing noise is expected to be of lower frequencies. GATK *HaplotypeCaller* (*McKenna et al., 2010*), MAQ (*Li, Ruan & Durbin, 2008*) or inGAP (*Qi et al., 2010*) represent germline-only variant callers. Nevertheless, some somatic callers can identify both germline and somatic variants such as Strelka2 (*Kim et al., 2018*), Varscan2 (*Koboldt et al., 2013*) or Octopus (*Cooke, Wedge & Lunter, 2018*). Further, we focus on the analysis of the somatic SNV/Indels since they represent a more challenging task compared to the detection of germline variants.

Table 1 A list of commonly used open-source variant callers and their usage possibilities.

Variant caller	Tumor-only mode	Variant type detection	References
CaVEMan	NO	SNV	*Jones et al. (2016)*
DeepSNV	NO	SNV	*Gerstung et al. (2012)*
DeepVariant	YES	SNV, Indel	*Poplin et al. (2018)*
EBCall	NO	SNV, Indel	*Shiraishi et al. (2013)*
FreeBayes	YES	SNV, Indel	*Garrison & Marth (2012)*
HapMuc	YES	SNV, Indel	*Usuyama et al. (2014)*
LocHap	NO	SNV, Indel	*Sengupta et al. (2016)*
LoFreq	YES	SNV, Indel	*Wilm et al. (2012)*
MuSE	NO	SNV	*Fan et al. (2016)*
Mutect	YES	SNV	*Cibulskis et al. (2013)*
Mutect2	YES	SNV, Indel	*Benjamin et al. (2019)*
Octopus	YES	SNV, Indel	*Cooke, Wedge & Lunter (2018)*
Platypus	YES	SNV, Indel	*Rimmer et al. (2014)*
SAMtools	YES	SNV, Indel	*Li et al. (2009)*
SomaticSniper	NO	SNV	*Larson et al. (2012)*
Strelka2	NO	SNV, Indel	*Kim et al. (2018)*
UMI-VarCal	YES	SNV	*Sater et al. (2020)*
VarDict	YES	SNV, Indel	*Lai et al. (2016)*
VarScan2	YES	SNV, Indel	*Koboldt et al. (2012)*

In the case of somatic variant calling, it is necessary to distinguish minor variants from background noise because true somatic variants can occur at low frequencies, especially in samples with low purity or rare tumor subclones. Herein, a statistical evaluation is critical to determine the underlying genotype and to discriminate among variants and artefacts. An ideal scenario for precise somatic variant identification is a paired sample analysis, where tumor and matched normal sample are compared. Variants present in the non-tumor sample are considered germline and excluded from subsequent interpretation. Several studies have evaluated somatic variant calling with default parameters in a tumor/normal setting (*Ellis et al., 2012*; *Parry et al., 2015*; *Chen et al., 2015*). The most popular open-source variant callers (Table 1) used diverse strategies for variant filtering that leads to differently reported variants demonstrating low concordance among various calling algorithms (*Liu et al., 2013*; *Ewing et al., 2015*; *Xu, 2018*). *Cacheiro et al., 2017* concluded that each evaluated caller exhibits a different ability to call SNV/Indels properly and showed discrepancies in the estimation of VAF. In a study by *Bian et al., 2018*, various callers like FreeBayes (*Garrison & Marth, 2012*), VarDict (*Lai et al., 2016*), GATK *MuTect*, GATK *Mutect2* (*Benjamin et al., 2019*) and MuSE (*Fan et al., 2016*) were tested to assess specificity and sensitivity. GATK *MuTect2* shown the best performance according to evaluated metrics, including the highest ratio of true/false positive variants across multiple datasets. (*Chen et al., 2019a*) showed that Strelka2 (*Kim et al., 2018*) identifies variants accurately and outperforms other tools in computational costs. The computational time of analysis is important, especially in routine

diagnostics, where results need to be obtained quickly. Rapid end-to-end analysis of sequencing data contributes to the improvement of cancer patient management, particularly in terms of effective therapy initiation. Combining results of multiple somatic callers to get consensus calls is more time consuming but provides more accurate variant calling. However, taking the intersection or the union of the call sets can lead to a drop in sensitivity or an increase in false-positive variants, respectively (*Callari et al., 2017*).

A matched non-tumor sample is not always easily accessible, especially in routine diagnostics or in retrospective analyses. Fortunately, some callers allow a "tumor only" mode of variant calling (Table 1), albeit with some additional challenges. Such calling indispensably requires the filtering of germline variants according to the information available in public population databases. However, this approach is not comprehensive as each individual could potentially have unknown germline variants, which are not included in the databases and could be wrongly considered somatic (*Jones et al., 2015*). Efficient identification of relevant mutations also strongly depends on the sufficient read depth and exclusion of false-positive calls. Some tools for tumor-only variant calling try to solve the stated issue with machine learning (*Kalatskaya et al., 2017*), but this requires large training datasets, which are not usually available during the implementation of targeted panels. This approach also does not consider the dynamics of tumor clone development in time. The validation of discovered variants is highly recommended not only for in-house developed algorithms but also for all commercial or freely available components (*Roy et al., 2018*).

Variant annotation and prioritization

Most variant callers use Variant Call Format (VCF) files to collect identified variants. VCFs store information such as mutation genotype, variant frequency, or genomic coordinates. Accurate annotation of generated variants is crucial for subsequent interpretation. Generally, the process of variant annotation integrates genomic variants into a functional and clinical context. Recently, several tools such as variant effect predictor (*McLaren et al., 2016*), ANNOVAR (*Wang, Li & Hakonarson, 2010*), SnpEff (*Cingolani et al., 2012*) or GATK VariantAnnotator were developed to provide comprehensive annotations including variant classification, standardized nomenclature, functional prediction, and information from available databases (Table S1). For routine diagnostics, it is essential to adhere to the unified nomenclature proposed by the Human Genome Variation Society (HGVS) (*Claustres et al., 2014*), which is a standard in variant description used for clinical reports. Mutalyzer (*Wildeman et al., 2008*), a web-based software, can be used to retrieve proper HGVS nomenclature for particular variants. However, the variant description itself is not sufficient for the assessment of its functional and clinical impact in the majority of cases. Therefore, multiple clinically relevant annotations are provided by tools collecting information from population databases or reports in published literature, or, in some cases, the effect of a variant may be estimated in silico (Table S2). Such information is permanently updated and is accessible through mentioned stand-alone annotation tools, which can be easily implemented within a bioinformatic

pipeline. Concurrently, the web-based Cancer Genome Interpreter (*Tamborero et al., 2017*) can be used to annotate somatic variants found in tumor samples, and it automatically predicts the possible role of a variant in tumorigenesis and treatment response.

After annotation, a prioritization of variants is performed to select clinically relevant variants to be reported to a clinician. Variant filtering strategies usually include several general steps and also specific steps according to individual panel settings determined by the validation process. Firstly, variants may be filtered by "variant type", where intronic, synonymous, non-coding variants are often of little interest and considered non-pathogenic. Secondly, a VAF cutoff, under which variants can no longer be reliably distinguished from the background noise (assay detection limit), has to be made, all the while considering the absolute depth and the number of alternative reads (i.e., a VAF of 5% at 1,000x coverage is different to VAF of 5% at 20x coverage). Finally, remaining variants must be inspected in the context of polymorphism database (dbSNP), mutation databases (e.g., COSMIC, IARC TP53) and population databases (e.g., 1000 Genomes, GnomAD) to help differentiate between mutations and common polymorphisms. In silico prediction of functional impact should be used only as a supplementary tool for variant interpretation and never as the sole evidence for possible pathogenic impact (*Li et al., 2017*). After prioritization, variants with VAF of about 50% or 100% represent candidates for the investigation of their potential germline origin.

High confident clinically relevant somatic variants may be visualized using Integrative Genomics Viewer (IGV) (*Thorvaldsdottir, Robinson & Mesirov, 2013*). It helps to identify false positive variants, which were not filtered out during the bioinformatic and prioritization steps. Manually identified artifacts are often variants: (1) with low-quality base calls, (2) from the erroneous end of reads, (3) arising from the misleading alignment of repetitive or homologous regions, and (4) with strand bias.

CNV analysis in targeted sequencing

Chromosomal aberrations play an indisputable role in cancer development and generally contribute to human genome variation (*Lupski, 2015*). Targeted identification of deletions and amplifications of specific genomic loci in cancer patients acquired relevance after determining their evident or potential clinical impact (*Concolino et al., 2018*; *Baliakas et al., 2019*; *Yu et al., 2020*). In general, all types of CNVs can be detected by one or more bioinformatic methods (*Zhao et al., 2013*) including: (1) paired-end mapping approach, (2) split read-based approach, (3) read depth-based approach, (4) de novo read assembly of CNV events, or (5) a combination of these approaches. The most widely-used approach is based on split reads and is used to detect and localize breakpoints of any type of CNVs.

In NGS targeted panels, a read depth-based strategy for CNV detection is employed in the majority of the developed software. This approach operates under the basic assumption that the number of DNA copies is proportional to the sequencing read depth in an analyzed genomic locus. Practically, the identification of copy number changes is based on the calculation of read depth variance between the tumor and normal sample within a

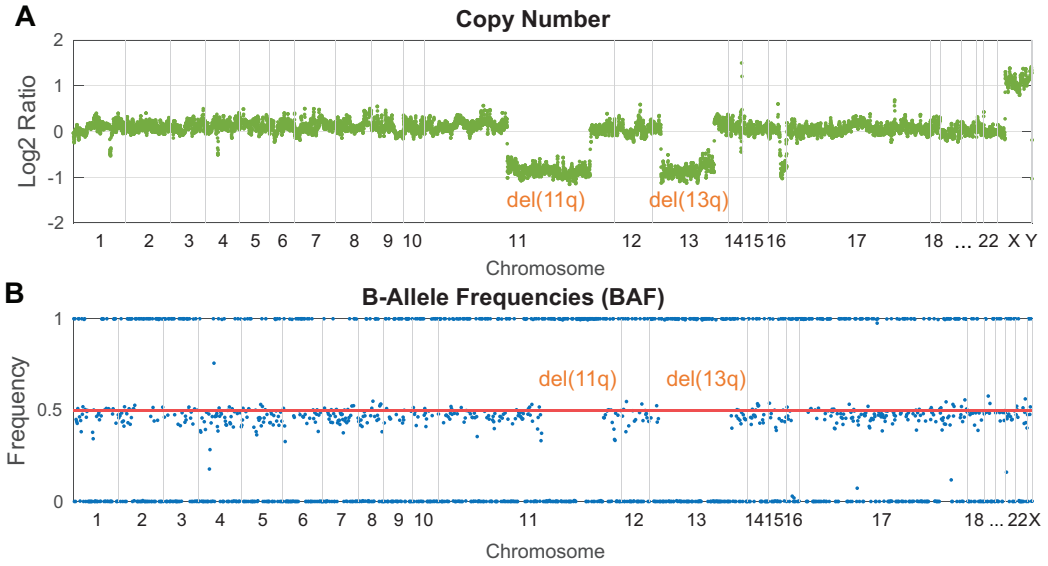

Figure 3 Visualization of clinically relevant CNV markers (del11q and del13p) detected in a peripheral blood sample from a patient with chronic lymphocytic leukemia. (A) Read depth approach, (B) B-allele frequency of analyzed SNPs. Probes along the chromosomes (*X*-axis) are depicted equidistantly. Full-size DOI: 10.7717/peerj.10897/fig-3

given region. Such normalization helps remove biases arising from GC rich regions and non-uniformly covered target regions. However, a comprehensive normalization step is more challenging for samples with fragmented DNA (FFPE samples or cfDNA), where the signal is non-uniformly dispersed. The presence of a heterozygous deletion in the tumor sample leads to halved coverage in the affected region compared to the normal sample, while coverage in the unaffected regions remains approximately the same. Vice versa, a gain of chromosomal material results in increased read depth. A final plot of log2 normalized read depth displays the occurrence of known clinically significant CNV markers. Illustratively, Fig. 3 shows the identification of del(13p) and del(11q) in a patient with chronic lymphocytic leukemia. Conversely, it could be challenging to identify novel potentially relevant disease markers due to probe density and plot resolution given by assay design. Notably, shorter aberrations or minor clones can be barely visible in a noisy background. Circular Binary Segmentation (CBS) segmentation (*Olshen et al., 2004*) is used to translate and join these noisy copy number neutral or aberrant regions to segments, which represent an equal copy number state.

In a tumor-only scenario, the selection of an appropriate normal sample represents an additional task, which dramatically influences the precision of results. Two main approaches can be applied. Firstly, a copy number neutral reference sample sequenced in the same batch with the tumor samples can be used for normalization to ensure coverage uniformity. The most suitable material seems to be commercially available reference genomic DNA provided intentionally for human genome sequencing (*Zook et al., 2016*). Secondly, a "virtual normal sample" can be generated from an overall read depth mean of multiple tumor samples analyzed concurrently in a sequencing run. This statistical

solution is implemented in software such as ExomeDepth (*Plagnol et al., 2012*), ONCOCNV (*Boeva et al., 2014*), CNVkit (*Talevich et al., 2016*) and panelcn.MOPS (*Povysil et al., 2017*). While ExomeDepth was designed to analyze CNVs in whole-exome data, ONCOCNV, panelcn.MOPS and CNVkit allow efficient analysis of targeted panels as well (*Paradiso et al., 2018*). Moreover, CNVPanelizer (*Oliveira & Wolf, 2019*) was explicitly developed to analyze CNVs in targeted panel assays. GATK's best practice provides a pipeline for somatic CNV calling; however, this workflow was recently optimized for WES.

B-Allele Frequency (BAF) analysis constitutes a complementary method to identify copy number aberrations (Fig. 3B) with the benefit of copy neutral loss-of-heterozygosity (cnLOH) detection. BAFs are mostly seen in the context of SNP arrays and represent allelic frequencies of germline SNPs. Their values are expected to be 0 (SNP not present), 0.5 (a heterozygous SNP), 1 (homozygous SNP). In affected regions of tumor DNA, heterozygous SNPs are expected to be "out of phase", shifted symmetrically above and below the heterozygous state. In perfectly pure tumor samples with fully clonal deletion, heterozygous SNPs in the affected region are expected to have BAF of 0 and 1. In reality, tumor samples are often "contaminated" by normal cells (especially in solid tumors) or contain minor subclones, which leads to the typical BAF range from <0 to 0.5> and <0.5 to 1> depending on the degree of contamination and the clonality of a given aberration. In targeted NGS, the informative strength and resolution of this approach are limited by the number of heterozygous SNPs analyzed in the target regions.

A comprehensive analysis of genomic markers by the targeted custom NGS panel-practical experience from Czech laboratory

Finally, we would like to share our practical experience with the implementation of a versatile capture-based NGS panel targeting various molecular markers in lymphoproliferative diseases. We aimed to integrate analyses of several markers with established or potential prognostic and predictive impact in B-cell neoplasms, including gene mutations, chromosomal aberrations, immunoglobulin (IG) or T-cell receptor (TR) rearrangements, and clinically relevant translocations. None of the available commercial or published research panels (*Rodríguez-Vicente, Díaz & Hernández-Rivas, 2013*; *Kluk et al., 2016*; *Hung et al., 2018*; *Kim et al., 2019*) suited our needs, and therefore we decided for a capture-based technology utilizing UMIs and a custom design due to the advantage of tailored options. Finally, our panel with a total capture size of 1.13 Mb included probes for the analysis of: (1) all exons and splice sites of 70 protein-coding genes and all functional genes of IG and TR loci, (2) recurrent deletions 17p, 11q, 13q in the desired resolution (300 kb–1 Mb) and (3) genome-wide CNVs and cnLOH enabled by the evenly spaced backbone of probes. The identification of common translocations in lymphomas, that is t(11;14), t(14;18), t(3;14), was ensured by probes covering the whole IGHJ region.

The validation of experimental and bioinformatic procedures was performed by the sequencing of 63 DNA samples extracted from various types of biological material obtained from 49 patients with diverse lymphoproliferation. The validation sample cohort

was selected to get a representative set of previously identified genetic alterations: (1) 109 SNV/Indels at various VAF (1–100%), (2) 79 CNVs (gains, losses) and cnLOHs of various extent (0.014–137 Mb) and tumor load (15–100%), (3) common translocations and IG/TR rearrangements. Commercial reference gDNA (NA24631; Coriell Institute, Camden, NJ, USA) was used in each sequencing run as a normal sample for CNV analysis and also for the assessment of panel performance (*Hardwick, Deveson & Mercer, 2017*). We validated several parameters according to available guidelines (*Jennings et al., 2017*) including accuracy metrics (positive percentage agreement, PPA; positive predictive value, PPV), sensitivity (limit of detection, LOD), reproducibility and specificity. The intended coverage was approximately 1000x after deduplication to achieve assay sensitivity of at least 5% VAF.

The in-house bioinformatic data analysis workflow consists of two independent branches: (1) a pipeline for detection of SNVs/Indels and CNVs implemented in Snakemake, and (2) a pipeline for the identification of IG/TR gene rearrangements and translocations. All scripts and software used in the first branch of bioinformatic analysis and with all non-default parameters are listed in Table S3. We measured the analysis run time (Fig. S1A) and memory (RAM) usage (Fig. S1B) in each step for ten samples analyzed in one sequencing run. Four CPUs cores with no parallelization were used, and the overall analysis running time was ~78 h. We will not dissect the second analytical branch since it exceeds the scope of this review.

Our panel demonstrated high coverage uniformity throughout all targets and across individual runs and showed the following basic parameters: (1) median coverage 921x after deduplication with 90% of targets >500x, 0.4% of targets <100x, (2) 43% PCR duplicates on average and (3) 27% off-targets reads on average (calculated before deduplication). The panel enables reproducible detection of SNV/Indels with high accuracy (PPA 100%, PPV 100%) and sensitivity when complying with our 5/5 rule for variant prioritization (i.e., VAF >5% and >5 variant reads) established during the validation process. The pre-defined LOD of 5% was corroborated by a dilution experiment. Manual inspection of suspicious variants in IGV was performed to avoid misinterpreting artifacts as true variants. The evaluation of CNV detection confirmed an expected high resolution of 300 kb–1 Mb in recurrently deleted loci of 17p, 11q and 13q arms and over 6 Mb across the whole genome depending on the backbone probe density. Out of all the tested CNVs, 91% were correctly identified. Seven undetected aberrations were below the resolution of the assay (either by extent or by their clonal proportion). The dilution experiment and validation results led us to set the threshold for CNV detection to log2 ratio $\geq \pm 0.2$ and BAF $\geq \pm 0.1$, which corresponds to the presence of at least 20% of cells with a respective chromosomal aberration in a sample. It is advantageous to combine both approaches, read depth and SNP analysis, since altogether, they provide complementary results.

In summary, we successfully implemented a versatile NGS capture-based tool for integrated analysis of molecular markers with research and clinical merit for patients with lymphoid malignancies (publication is under revision).

CONCLUSIONS

The purpose of this review was to provide a guidebook for the development of a robust bioinformatic pipeline for the analysis of clinically relevant molecular markers detected by targeted NGS panel intended for routine use. We present an overview of contemporary bioinformatic approaches for the analysis of genomic aberrations supported by an example of a successfully implemented comprehensive capture-based NGS panel. According to our experience, the most crucial steps for targeted NGS tool development are: (1) appropriate selection of validation cohort comprising plenitude of representative samples and diverse targets, (2) careful optimization and validation of the analytical pipeline based on state-of-the-art bioinformatic approaches to ensure high accuracy of the results and (3) robust software and hardware environment. The whole procedure of specific tool implementation is rather time-consuming but highly rewarding, especially when a custom assay with long-term use needs to be established.

ACKNOWLEDGEMENTS

We acknowledge the Core Facility Genomics CEITEC MU supported by the NCMG research infrastructure (LM2015091 funded by MEYS CR) for generating sequencing data. We thank Francesco Muto for English proofreading.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

This work was supported by projects MH CR AZV NV19-03-00091, MUNI/A/1395/2019, and European Regional Development Fund-Project "A-C-G-T" (MEYS No. CZ.02.1.01/0.0/0.0/16_026/0008448). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the authors:
MH CR AZV NV19-03-00091, MUNI/A/1395/2019 and "A-C-G-T": MEYS No. CZ.02.1.01/0.0/0.0/16_026/0008448.

Competing Interests

The authors declare that they have no competing interests.

Author Contributions

- Jakub Hynst analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the paper, and approved the final draft.
- Veronika Navrkalova conceived and designed the experiments, performed the experiments, analyzed the data, authored or reviewed drafts of the paper, and approved the final draft.
- Karol Pal analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the paper, and approved the final draft.

- Sarka Pospisilova authored or reviewed drafts of the paper, and approved the final draft.

Data Availability

The following information was supplied regarding data availability:

There is no raw data or code because our article is a literature review.

Supplemental Information

Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj.10897#supplemental-information.

REFERENCES

Afgan E, Baker D, Batut B, Van den Beek M, Bouvier D, Cech M, Chilton J, Clements D, Coraor N, Grüning BA, Guerler A, Hillman-Jackson J, Hiltemann S, Jalili V, Rasche H, Soranzo N, Goecks J, Taylor J, Nekrutenko A, Blankenberg D. 2018. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. *Nucleic Acids Research* **46(W1)**:W537–W544 DOI 10.1093/nar/gky379.

Ahmed N, Lévy J, Ren S, Mushtaq H, Bertels K, Al-Ars Z. 2019. GASAL2: a GPU accelerated sequence alignment library for high-throughput NGS data. *BMC Bioinformatics* **20(1)**:520 DOI 10.1186/s12859-019-3086-9.

Allgäuer M, Budczies J, Christopoulos P, Endris V, Lier A, Rempel E, Volckmar A-L, Kirchner M, Von Winterfeld M, Leichsenring J, Neumann O, Fröhling S, Penzel R, Thomas M, Schirmacher P, Stenzinger A. 2018. Implementing tumor mutational burden (TMB) analysis in routine diagnostics—a primer for molecular pathologists and clinicians. *Translational Lung Cancer Research* **7(5)**:703–715 DOI 10.21037/tlcr.2018.08.14.

Armaou S, Pertesi M, Fostira F, Thodi G, Athanasopoulos PS, Kamakari S, Athanasiou A, Gogas H, Yannoukakos D, Fountzilas G, Konstantopoulou I. 2009. Contribution of BRCA1 germ-line mutations to breast cancer in Greece: a hospital-based study of 987 unselected breast cancer cases. *British Journal of Cancer* **101(1)**:32–37 DOI 10.1038/sj.bjc.6605115.

Ascierto PA, Kirkwood JM, Grob J-J, Simeone E, Grimaldi AM, Maio M, Palmieri G, Testori A, Marincola FM, Mozzillo N. 2012. The role of BRAF V600 mutation in melanoma. *Journal of Translational Medicine* **10(1)**:85 DOI 10.1186/1479-5876-10-85.

Amazon. 2020. AWS Free Tier. *Available at https://aws.amazon.com/free/* (accessed 18 December 2020).

Baliakas P, Jeromin S, Iskas M, Puiggros A, Plevova K, Nguyen-Khac F, Davis Z, Rigolin GM, Visentin A, Xochelli A, Delgado J, Baran-Marszak F, Stalika E, Abrisqueta P, Durechova K, Papaioannou G, Eclache V, Dimou M, Iliakis T, Collado R, Doubek M, Calasanz MJ, Ruiz-Xiville N, Moreno C, Jarosova M, Leeksma AC, Panayiotidis P, Podgornik H, Cymbalista F, Anagnostopoulos A, Trentin L, Stavroyianni N, Davi F, Ghia P, Kater AP, Cuneo A, Pospisilova S, Espinet B, Athanasiadou A, Oscier D, Haferlach C, Stamatopoulos K, ERIC, the European Research Initiative on CLL. 2019. Cytogenetic complexity in chronic lymphocytic leukemia: definitions, associations, and clinical impact. *Blood* **133(11)**:1205–1216 DOI 10.1182/blood-2018-09-873083.

Benjamin D, Sato T, Cibulskis K, Getz G, Stewart C, Lichtenstein L. 2019. Calling somatic SNVs and indels with mutect2. *bioRxiv* DOI 10.1101/861054.

Bewicke-Copley F, Arjun Kumar E, Palladino G, Korfi K, Wang J. 2019. Applications and analysis of targeted genomic sequencing in cancer studies. *Computational and Structural Biotechnology Journal* **17(1029–1041)**:1348–1359 DOI 10.1016/j.csbj.2019.10.004.

Bian X, Zhu B, Wang M, Hu Y, Chen Q, Nguyen C, Hicks B, Meerzaman D. 2018. Comparing the performance of selected variant callers using synthetic data and genome segmentation. *BMC Bioinformatics* **19(1)**:429 DOI 10.1186/s12859-018-2440-7.

Boeva V, Popova T, Lienard M, Toffoli S, Kamal M, Le Tourneau C, Gentien D, Servant N, Gestraud P, Rio Frio T, Hupé P, Barillot E, Laes J-F. 2014. Multi-factor data normalization enables the detection of copy number aberrations in amplicon sequencing data. *Bioinformatics* **30(24)**:3443–3450 DOI 10.1093/bioinformatics/btu436.

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30(15)**:2114–2120 DOI 10.1093/bioinformatics/btu170.

Cacheiro P, Ordóñez-Ugalde A, Quintáns B, Piñeiro-Hermida S, Amigo J, García-Murias M, Pascual-Pascual SI, Grandas F, Arpa J, Carracedo A, Sobrido MJ. 2017. Evaluating the calling performance of a rare disease NGS panel for single nucleotide and copy number variants. *Molecular Diagnosis & Therapy* **21(3)**:303–313 DOI 10.1007/s40291-017-0268-x.

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

Cardoso JMP, Fey D, Hannig F, Pionteck T, Schröder-Preikschat W, Teich J. 2016. Architecture of computing systems—ARCS 2016. In: *29th International Conference, Nuremberg, Germany, April 4-7, 2016, Proceedings*. Cham: Springer International Publishing.

Chapman B, Kirchner R, Pantano L, Smet MD, Beltrame L, Khotiainsteva T, Naumenko S, Saveliev V, Guimera RV, Sytchev I, Kern J, Brueffer C, Carrasco G, Giovacchini M, Tang P, Ahdesmaki M, Kanwal S, Porter JJ, Möller S, Le V, Coman A, Svensson V, Bogdang989 M, Edwards M, Hammerbacher M, Pedersen J, Cock B, Apastore P, Turner S. 2020. *bcbio/bcbio-nextgen. Zenodo*. Version 1.2.3. *Available at http://doi.org/10.5281/zenodo.3743344*.

Chen J, Li X, Zhong H, Meng Y, Du H. 2019a. Systematic comparison of germline variant calling pipelines cross multiple next-generation sequencers. *Scientific Reports* **9(1)**:9345 DOI 10.1038/s41598-019-45835-3.

Chen K, Meric-Bernstam F, Zhao H, Zhang Q, Ezzeddine N, Tang L, Qi Y, Mao Y, Chen T, Chong Z, Zhou W, Zheng X, Johnson A, Aldape KD, Routbort MJ, Luthra R, Kopetz S, Davies MA, De Groot J, Moulder S, Vinod R, Farhangfar CJ, Shaw KM, Mendelsohn J, Mills GB, Karina Eterovic A. 2015. Clinical actionability enhanced through deep targeted sequencing of solid tumors. *Clinical Chemistry* **61(3)**:544–553 DOI 10.1373/clinchem.2014.231100.

Chen S, Zhou Y, Chen Y, Gu J. 2018. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34(17)**:i884–i890 DOI 10.1093/bioinformatics/bty560.

Chen S, Zhou Y, Chen Y, Huang T, Liao W, Xu Y, Li Z, Gu J. 2019b. Gencore: an efficient tool to generate consensus reads for error suppressing and duplicate removing of NGS data. *BMC Bioinformatics* **20(S23)**:606 DOI 10.1186/s12859-019-3280-9.

Chilamakuri CSR, Lorenz S, Madoui M-A, Vodák D, Sun J, Hovig E, Myklebost O, Meza-Zepeda LA. 2014. Performance comparison of four exome capture systems for deep sequencing. *BMC Genomics* **15(1)**:449 DOI 10.1186/1471-2164-15-449.

Church DM, Schneider VA, Graves T, Auger K, Cunningham F, Bouk N, Chen H-C, Agarwala R, McLaren WM, Ritchie GRS, Albracht D, Kremitzki M, Rock S, Kotkiewicz H,

Kremitzki C, Wollam A, Trani L, Fulton L, Fulton R, Matthews L, Whitehead S, Chow W, Torrance J, Dunn M, Harden G, Threadgold G, Wood J, Collins J, Heath P, Griffiths G, Pelan S, Grafham D, Eichler EE, Weinstock G, Mardis ER, Wilson RK, Howe K, Flicek P, Hubbard T. 2011. Modernizing reference genome assemblies. *PLOS Biology* **9(7)**:e1001091 DOI 10.1371/journal.pbio.1001091.

Cibulskis K, Lawrence MS, Carter SL, Sivachenko A, Jaffe D, Sougnez C, Gabriel S, Meyerson M, Lander ES, Getz G. 2013. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nature Biotechnology* **31(3)**:213–219 DOI 10.1038/nbt.2514.

Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM. 2012. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly* **6(2)**:80–92 DOI 10.4161/fly.19695.

Claustres M, Kožich V, Dequeker E, Fowler B, Hehir-Kwa JY, Miller K, Oosterwijk C, Peterlin B, van Ravenswaaij-Arts C, Zimmermann U, Zuffardi O, Hastings RJ, Barton DE, European Society of Human Genetics. 2014. Recommendations for reporting results of diagnostic genetic testing (biochemical, cytogenetic and molecular genetic). *European Journal of Human Genetics* **22**:160–170 DOI 10.1038/ejhg.2013.125.

Concolino P, Rizza R, Mignone F, Costella A, Guarino D, Carboni I, Capoluongo E, Santonocito C, Urbani A, Minucci A. 2018. A comprehensive BRCA1/2 NGS pipeline for an immediate Copy Number Variation (CNV) detection in breast and ovarian cancer molecular diagnosis. *Clinica Chimica Acta* **480**:173–179 DOI 10.1016/j.cca.2018.02.012.

Cooke DP, Wedge DC, Lunter G. 2018. A unified haplotype-based method for accurate and comprehensive variant calling. *Bioinformatics* **1**:20 DOI 10.1101/456103.

Costello M, Fleharty M, Abreu J, Farjoun Y, Ferriera S, Holmes L, Granger B, Green L, Howd T, Mason T, Vicente G, Dasilva M, Brodeur W, DeSmet T, Dodge S, Lennon NJ, Gabriel S. 2018. Characterization and remediation of sample index swaps by non-redundant dual indexing on massively parallel sequencing platforms. *BMC Genomics* **19(1)**:332 DOI 10.1186/s12864-018-4703-0.

Di Tommaso P, Chatzou M, Floden EW, Barja PP, Palumbo E, Notredame C. 2017. Nextflow enables reproducible computational workflows. *Nature Biotechnology* **35(4)**:316–319 DOI 10.1038/nbt.3820.

Dubois S, Viailly P-J, Mareschal S, Bohers E, Bertrand P, Ruminy P, Maingonnat C, Jais J-P, Peyrouze P, Figeac M, Molina TJ, Desmots F, Fest T, Haioun C, Lamy T, Copie-Bergman C, Brière J, Petrella T, Canioni D, Fabiani B, Coiffier B, Delarue R, Peyrade F, Bosly A, André M, Ketterer N, Salles G, Tilly H, Leroy K, Jardin F. 2016. Next-generation sequencing in diffuse large B-cell lymphoma highlights molecular divergence and therapeutic opportunities: a LYSA study. *Clinical Cancer Research: An Official Journal of the American Association for Cancer Research* **22(12)**:2919–2928 DOI 10.1158/1078-0432.CCR-15-2305.

EGI. 2020. EGI Advanced Computing Services for Research. *Available at https://www.egi.eu/* (accessed 18 December 2020).

Ellis MJ, Ding L, Shen D, Luo J, Suman VJ, Wallis JW, Van Tine BA, Hoog J, Goiffon RJ, Goldstein TC, Ng S, Lin L, Crowder R, Snider J, Ballman K, Weber J, Chen K, Koboldt DC, Kandoth C, Schierding WS, McMichael JF, Miller CA, Lu C, Harris CC, McLellan MD, Wendl MC, DeSchryver K, Allred DC, Esserman L, Unzeitig G, Margenthaler J, Babiera GV, Marcom PK, Guenther JM, Leitch M, Hunt K, Olson J, Tao Y, Maher CA, Fulton LL, Fulton RS, Harrison M, Oberkfell B, Du F, Demeter R, Vickery TL, Elhammali A,

Piwnica-Worms H, McDonald S, Watson M, Dooling DJ, Ota D, Chang L-W, Bose R, Ley TJ, Piwnica-Worms D, Stuart JM, Wilson RK, Mardis ER. 2012. Whole-genome analysis informs breast cancer response to aromatase inhibition. *Nature* **486(7403)**:353–360 DOI 10.1038/nature11143.

Ensembl. 2020. BED file format. *Available at https://www.ensembl.org/info/website/upload/bed.html* (accessed 18 December 2020).

Ewels PA, Peltzer A, Fillinger S, Patel H, Alneberg J, Wilm A, Garcia MU, Di Tommaso P, Nahnsen S. 2020. The nf-core framework for community-curated bioinformatics pipelines. *Nature Biotechnology* **38(3)**:276–278 DOI 10.1038/s41587-020-0439-x.

Ewing AD, Houlahan KE, Hu Y, Ellrott K, Caloian C, Yamaguchi TN, Bare JC, P'ng C, Waggott D, Sabelnykova VY, Kellen MR, Norman TC, Haussler D, Friend SH, Stolovitzky G, Margolin AA, Stuart JM, Boutros PC, ICGC-TCGA DREAM Somatic Mutation Calling Challenge Participants. 2015. Combining tumor genome simulation with crowdsourcing to benchmark somatic single-nucleotide-variant detection. *Nature Methods* **12(7)**:623–630 DOI 10.1038/nmeth.3407.

Fan Y, Xi L, Hughes DST, Zhang J, Zhang J, Futreal PA, Wheeler DA, Wang W. 2016. MuSE: accounting for tumor heterogeneity using a sample-specific error model improves sensitivity and specificity in mutation calling from sequencing data. *Genome Biology* **17(1)**:178 DOI 10.1186/s13059-016-1029-6.

fgbio. 2020. fgbio: Tools for working with genomic and high throughput sequencing data. *Available at http://fulcrumgenomics.github.io/fgbio/*.

Gargis AS, Kalman L, Bick DP, Da Silva C, Dimmock DP, Funke BH, Gowrisankar S, Hegde MR, Kulkarni S, Mason CE, Nagarajan R, Voelkerding KV, Worthey EA, Aziz N, Barnes J, Bennett SF, Bisht H, Church DM, Dimitrova Z, Gargis SR, Hafez N, Hambuch T, Hyland FCL, Luna RA, MacCannell D, Mann T, McCluskey MR, McDaniel TK, Ganova-Raeva LM, Rehm HL, Reid J, Campo DS, Resnick RB, Ridge PG, Salit ML, Skums P, Wong L-JC, Zehnbauer BA, Zook JM, Lubin IM. 2015. Good laboratory practice for clinical next-generation sequencing informatics pipelines. *Nature Biotechnology* **33(7)**:689–693 DOI 10.1038/nbt.3237.

Garrison E, Marth G. 2012. Haplotype-based variant detection from short-read sequencing. *ArXiv*. *Available at http://arxiv.org/abs/1207.3907*.

Gerstung M, Beisel C, Rechsteiner M, Wild P, Schraml P, Moch H, Beerenwinkel N. 2012. Reliable detection of subclonal single-nucleotide variants in tumour cell populations. *Nature Communications* **3(1)**:811 DOI 10.1038/ncomms1814.

Girardot C, Scholtalbers J, Sauer S, Su S-Y, Furlong EEM. 2016. Je, a versatile suite to handle multiplexed NGS libraries with unique molecular identifiers. *BMC Bioinformatics* **17(1)**:419 DOI 10.1186/s12859-016-1284-2.

Grüning B, Dale R, Sjödin A, Chapman BA, Rowe J, Tomkins-Tinch CH, Valieris R, Köster J, Bioconda Team. 2018. Bioconda: sustainable and comprehensive software distribution for the life sciences. *Nature Methods* **15(7)**:475–476 DOI 10.1038/s41592-018-0046-7.

Guan P, Sung W-K. 2016. Structural variation detection using next-generation sequencing data: a comparative technical review. *Methods* **102(21)**:36–49 DOI 10.1016/j.ymeth.2016.01.020.

Hamblin A, Wordsworth S, Fermont JM, Page S, Kaur K, Camps C, Kaisaki P, Gupta A, Talbot D, Middleton M, Henderson S, Cutts A, Vavoulis DV, Housby N, Tomlinson I, Taylor JC, Schuh A. 2017. Clinical applicability and cost of a 46-gene panel for genomic analysis of solid tumours: retrospective validation and prospective audit in the UK National health service. *PLOS Medicine* **14(2)**:e1002230 DOI 10.1371/journal.pmed.1002230.

Hardwick SA, Deveson IW, Mercer TR. 2017. Reference standards for next-generation sequencing. *Nature Reviews Genetics* **18(8)**:473–484 DOI 10.1038/nrg.2017.44.

Heydt C, Rehker J, Pappesch R, Buhl T, Ball M, Siebolts U, Haak A, Lohneis P, Büttner R, Hillmer AM, Merkelbach-Bruse S. 2020. Analysis of tumor mutational burden: correlation of five large gene panels with whole exome sequencing. *Scientific Reports* **10(1)**:11387 DOI 10.1038/s41598-020-68394-4.

Hung SS, Meissner B, Chavez EA, Ben-Neriah S, Ennishi D, Jones MR, Shulha HP, Chan FC, Boyle M, Kridel R, Gascoyne RD, Mungall AJ, Marra MA, Scott DW, Connors JM, Steidl C. 2018. Assessment of capture and amplicon-based approaches for the development of a targeted next-generation sequencing pipeline to personalize lymphoma management. *Journal of Molecular Diagnostics* **20(2)**:203–214 DOI 10.1016/j.jmoldx.2017.11.010.

Jabara CB, Jones CD, Roach J, Anderson JA, Swanstrom R. 2011. Accurate sampling and deep sequencing of the HIV-1 protease gene using a Primer ID. *Proceedings of the National Academy of Sciences of the United States of America* **108**:20166–20171 DOI 10.1073/pnas.1110064108.

Jennings LJ, Arcila ME, Corless C, Kamel-Reid S, Lubin IM, Pfeifer J, Temple-Smolkin RL, Voelkerding KV, Nikiforova MN. 2017. Guidelines for validation of next-generation sequencing-based oncology panels: a joint consensus recommendation of the association for molecular pathology and college of american pathologists. *Journal of Molecular Diagnostics* **19(3)**:341–365 DOI 10.1016/j.jmoldx.2017.01.011.

Jones S, Anagnostou V, Lytle K, Parpart-Li S, Nesselbush M, Riley DR, Shukla M, Chesnick B, Kadan M, Papp E, Galens KG, Murphy D, Zhang T, Kann L, Sausen M, Angiuoli SV, Diaz LA, Velculescu VE. 2015. Personalized genomic analyses for cancer mutation discovery and interpretation. *Science Translational Medicine* **7(283)**:283ra53 DOI 10.1126/scitranslmed.aaa7161.

Jones D, Raine KM, Davies H, Tarpey PS, Butler AP, Teague JW, Nik-Zainal S, Campbell PJ. 2016. cgpCaVEManWrapper: simple execution of CaVEMan in order to detect somatic single nucleotide variants in NGS data. *Current Protocols in Bioinformatics* **56(1)**:15.10.1–15.10.18 DOI 10.1002/cpbi.20.

Kalatskaya I, Trinh QM, Spears M, McPherson JD, Bartlett JMS, Stein L. 2017. ISOWN: accurate somatic mutation identification in the absence of normal tissue controls. *Genome Medicine* **9(1)**:59 DOI 10.1186/s13073-017-0446-9.

Kim B, Lee H, Kim E, Shin S, Lee S-T, Choi JR. 2019. Clinical utility of targeted NGS panel with comprehensive bioinformatics analysis for patients with acute lymphoblastic leukemia. *Leukemia & Lymphoma* **60(13)**:1–8 DOI 10.1080/10428194.2019.1627538.

Kim S, Scheffler K, Halpern AL, Bekritsky MA, Noh E, Källberg M, Chen X, Kim Y, Beyter D, Krusche P, Saunders CT. 2018. Strelka2: fast and accurate calling of germline and somatic variants. *Nature Methods* **15(8)**:591–594 DOI 10.1038/s41592-018-0051-x.

Kluk MJ, Lindsley RC, Aster JC, Lindeman NI, Szeto D, Hall D, Kuo FC. 2016. Validation and implementation of a custom next-generation sequencing clinical assay for hematologic malignancies. *Journal of Molecular Diagnostics* **18(4)**:507–515 DOI 10.1016/j.jmoldx.2016.02.003.

Klus P, Lam S, Lyberg D, Cheung MS, Pullan G, McFarlane I, Yeo GS, Lam BY. 2012. BarraCUDA—a fast short read sequence aligner using graphics processing units. *BMC Research Notes* **5(1)**:27 DOI 10.1186/1756-0500-5-27.

Koboldt DC, Steinberg KM, Larson DE, Wilson RK, Mardis ER. 2013. The next-generation sequencing revolution and its impact on genomics. *Cell* **155(1)**:27–38 DOI 10.1016/j.cell.2013.09.006.

Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, Miller CA, Mardis ER, Ding L, Wilson RK. 2012. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Research* **22(3)**:568–576 DOI 10.1101/gr.129684.111.

Köster J, Rahmann S. 2018. Snakemake-a scalable bioinformatics workflow engine. *Bioinformatics* **34**:3600 DOI 10.1093/bioinformatics/bty350.

Kuo FC, Mar BG, Lindsley RC, Lindeman NI. 2017. The relative utilities of genome-wide, gene panel, and individual gene sequencing in clinical practice. *Blood* **130(4)**:433–439 DOI 10.1182/blood-2017-03-734533.

Kurtzer GM, Sochat V, Bauer MW. 2017. Singularity: scientific containers for mobility of compute. *PLOS ONE* **12(5)**:e0177459 DOI 10.1371/journal.pone.0177459.

Lai Z, Markovets A, Ahdesmaki M, Chapman B, Hofmann O, McEwen R, Johnson J, Dougherty B, Barrett JC, Dry JR. 2016. VarDict: a novel and versatile variant caller for next-generation sequencing in cancer research. *Nucleic Acids Research* **44(11)**:e108–e108 DOI 10.1093/nar/gkw227.

Landau DA, Tausch E, Taylor-Weiner AN, Stewart C, Reiter JG, Bahlo J, Kluth S, Bozic I, Lawrence M, Böttcher S, Carter SL, Cibulskis K, Mertens D, Sougnez CL, Rosenberg M, Hess JM, Edelmann J, Kless S, Kneba M, Ritgen M, Fink A, Fischer K, Gabriel S, Lander ES, Nowak MA, Döhner H, Hallek M, Neuberg D, Getz G, Stilgenbauer S, Wu CJ. 2015. Mutations driving CLL and their evolution in progression and relapse. *Nature* **526(7574)**:525–530 DOI 10.1038/nature15395.

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9(4)**:357–359 DOI 10.1038/nmeth.1923.

Larson DE, Harris CC, Chen K, Koboldt DC, Abbott TE, Dooling DJ, Ley TJ, Mardis ER, Wilson RK, Ding L. 2012. SomaticSniper: identification of somatic point mutations in whole genome sequencing data. *Bioinformatics* **28(3)**:311–317 DOI 10.1093/bioinformatics/btr665.

Lenis J, Senar MA. 2015. On the performance of BWA on NUMA architectures. In: *2015 IEEE Trustcom/BigDataSE/ISPA*. 236–241.

Li MM, Datto M, Duncavage EJ, Kulkarni S, Lindeman NI, Roy S, Tsimberidou AM, Vnencak-Jones CL, Wolff DJ, Younes A, Nikiforova MN. 2017. Standards and guidelines for the interpretation and reporting of sequence variants in cancer: a joint consensus recommendation of the association for molecular pathology, American Society of Clinical Oncology, and College of American Pathologists. *Journal of Molecular Diagnostics* **19(1)**:4–23 DOI 10.1016/j.jmoldx.2016.10.002.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25(14)**:1754–1760 DOI 10.1093/bioinformatics/btp324.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. 2009. The sequence alignment/map format and SAMtools. *Bioinformatics* **25(16)**:2078–2079 DOI 10.1093/bioinformatics/btp352.

Li H, Ruan J, Durbin R. 2008. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Research* **18(11)**:1851–1858 DOI 10.1101/gr.078212.108.

Liang RH, Mo T, Dong W, Lee GQ, Swenson LC, McCloskey RM, Woods CK, Brumme CJ, Ho CKY, Schinkel J, Joy JB, Harrigan PR, Poon AFY. 2014. Theoretical and experimental assessment of degenerate primer tagging in ultra-deep applications of next-generation sequencing. *Nucleic Acids Research* **42**:e98 DOI 10.1093/nar/gku355.

Lindsley RC, Mar BG, Mazzola E, Grauman PV, Shareef S, Allen SL, Pigneux A, Wetzler M, Stuart RK, Erba HP, Damon LE, Powell BL, Lindeman N, Steensma DP, Wadleigh M, DeAngelo DJ, Neuberg D, Stone RM, Ebert BL. 2015. Acute myeloid leukemia ontogeny is

defined by distinct somatic mutations. *Blood* **125(9)**:1367–1376 DOI 10.1182/blood-2014-11-610543.

Liu X, Han S, Wang Z, Gelernter J, Yang B-Z. 2013. Variant callers for next-generation sequencing data: a comparison study. *PLOS ONE* **8(9)**:e75619 DOI 10.1371/journal.pone.0075619.

Lunter G, Goodson M. 2011. Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome Research* **21(6)**:936–939 DOI 10.1101/gr.111120.110.

Lupski JR. 2015. Structural variation mutagenesis of the human genome: Impact on disease and evolution. *Environmental and Molecular Mutagenesis* **56(5)**:419–436 DOI 10.1002/em.21943.

MacConaill LE, Burns RT, Nag A, Coleman HA, Slevin MK, Giorda K, Light M, Lai K, Jarosz M, McNeill MS, Ducar MD, Meyerson M, Thorner AR. 2018. Unique, dual-indexed sequencing adapters with UMIs effectively eliminate index cross-talk and significantly improve sensitivity of massively parallel sequencing. *BMC Genomics* **19(1)**:30 DOI 10.1186/s12864-017-4428-5.

Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17(1)**:10 DOI 10.14806/ej.17.1.200.

McConnell L, Houghton O, Stewart P, Gazdova J, Srivastava S, Kim C, Catherwood M, Strobl A, Flanagan AM, Oniscu A, Kroeze LI, Groenen P, Taniere P, Salto-Tellez M, Gonzalez D. 2020. A novel next generation sequencing approach to improve sarcoma diagnosis. *Modern Pathology* **33(7)**:1350–1359 DOI 10.1038/s41379-020-0488-1.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA. 2010. The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research* **20(9)**:1297–1303 DOI 10.1101/gr.107524.110.

McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, Flicek P, Cunningham F. 2016. The ensembl variant effect predictor. *Genome Biology* **17(1)**:122 DOI 10.1186/s13059-016-0974-4.

Merkel D. 2014. Docker: lightweight linux containers for consistent development and deployment. *Linux journal* **(239)**:2.

Metzker ML. 2010. Sequencing technologies—the next generation. *Nature Reviews Genetics* **11(1)**:31–46 DOI 10.1038/nrg2626.

Nadeu F, Delgado J, Royo C, Baumann T, Stankovic T, Pinyol M, Jares P, Navarro A, Martín-García D, Beà S, Salaverria I, Oldreive C, Aymerich M, Suárez-Cisneros H, Rozman M, Villamor N, Colomer D, López-Guillermo A, González M, Alcoceba M, Terol MJ, Colado E, Puente XS, López-Otín C, Enjuanes A, Campo E. 2016. Clinical impact of clonal and subclonal TP53, SF3B1, BIRC3, NOTCH1, and ATM mutations in chronic lymphocytic leukemia. *Blood* **127(17)**:2122–2130 DOI 10.1182/blood-2015-07-659144.

Nikiforova MN, Mercurio S, Wald AI, Barbi de Moura M, Callenberg K, Santana-Santos L, Gooding WE, Yip L, Ferris RL, Nikiforov YE. 2018. Analytical performance of the ThyroSeq v3 genomic classifier for cancer diagnosis in thyroid nodules. *Cancer* **124(8)**:1682–1690 DOI 10.1002/cncr.31245.

Novocraft. 2020. NovoAlign. *Available at http://www.novocraft.com/products/novoalign/* (accessed 18 December 2020).

Oliveira C, Wolf T. 2019. *CNVPanelizer: reliable CNV detection in targeted sequencing applications*. R package version 1.18.0. *Available at https://bioconductor.org/packages/release/bioc/html/CNVPanelizer.html*.

Olshen AB, Venkatraman ES, Lucito R, Wigler M. 2004. Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics* **5(4)**:557–572 DOI 10.1093/biostatistics/kxh008.

Orabi B, Erhan E, McConeghy B, Volik SV, Le Bihan S, Bell R, Collins CC, Chauve C, Hach F. 2019. Alignment-free clustering of UMI tagged DNA molecules. *Bioinformatics* **35(11)**:1829–1836 DOI 10.1093/bioinformatics/bty888.

OSG. 2020. Open Science Grid. *Available at http://opensciencegrid.org/about/introduction/* (accessed 18 December 2020).

Paasinen-Sohns A, Koelzer VH, Frank A, Schafroth J, Gisler A, Sachs M, Graber A, Rothschild SI, Wicki A, Cathomas G, Mertz KD. 2017. Single-center experience with a targeted next generation sequencing assay for assessment of relevant somatic alterations in solid tumors. *Neoplasia* **19(3)**:196–206 DOI 10.1016/j.neo.2017.01.003.

Papaemmanuil E, Gerstung M, Bullinger L, Gaidzik VI, Paschka P, Roberts ND, Potter NE, Heuser M, Thol F, Bolli N, Gundem G, Van Loo P, Martincorena I, Ganly P, Mudie L, McLaren S, O'Meara S, Raine K, Jones DR, Teague JW, Butler AP, Greaves MF, Ganser A, Döhner K, Schlenk RF, Döhner H, Campbell PJ. 2016. Genomic classification and prognosis in acute myeloid leukemia. *New England Journal of Medicine* **374(23)**:2209–2221 DOI 10.1056/NEJMoa1516192.

Paradiso V, Garofoli A, Tosti N, Lanzafame M, Perrina V, Quagliata L, Matter MS, Wieland S, Heim MH, Piscuoglio S, Ng CKY, Terracciano LM. 2018. Diagnostic targeted sequencing panel for hepatocellular carcinoma genomic screening. *Journal of Molecular Diagnostics* **20(6)**:836–848 DOI 10.1016/j.jmoldx.2018.07.003.

Parry M, Rose-Zerilli MJJ, Ljungstrom V, Gibson J, Wang J, Walewska R, Parker H, Parker A, Davis Z, Gardiner A, McIver-Brown N, Kalpadakis C, Xochelli A, Anagnostopoulos A, Fazi C, Gonzalez de Castro D, Dearden C, Pratt G, Rosenquist R, Ashton-Key M, Forconi F, Collins A, Ghia P, Matutes E, Pangalis G, Stamatopoulos K, Oscier D, Strefford JC. 2015. Genetics and prognostication in splenic marginal zone lymphoma: revelations from deep sequencing. *Clinical Cancer Research* **21(18)**:4174–4183 DOI 10.1158/1078-0432.CCR-14-2759.

Pastore A, Jurinovic V, Kridel R, Hoster E, Staiger AM, Szczepanowski M, Pott C, Kopp N, Murakami M, Horn H, Leich E, Moccia AA, Mottok A, Sunkavalli A, Van Hummelen P, Ducar M, Ennishi D, Shulha HP, Hother C, Connors JM, Sehn LH, Dreyling M, Neuberg D, Möller P, Feller AC, Hansmann ML, Stein H, Rosenwald A, Ott G, Klapper W, Unterhalt M, Hiddemann W, Gascoyne RD, Weinstock DM, Weigert O. 2015. Integration of gene mutations in risk prognostication for patients receiving first-line immunochemotherapy for follicular lymphoma: a retrospective analysis of a prospective clinical trial and validation in a population-based registry. *Lancet Oncology* **16(9)**:1111–1122 DOI 10.1016/S1470-2045(15)00169-2.

Plagnol V, Curtis J, Epstein M, Mok KY, Stebbings E, Grigoriadou S, Wood NW, Hambleton S, Burns SO, Thrasher AJ, Kumararatne D, Doffinger R, Nejentsev S. 2012. A robust model for read count data in exome sequencing experiments and implications for copy number variant calling. *Bioinformatics* **28(21)**:2747–2754 DOI 10.1093/bioinformatics/bts526.

Poplin R, Chang P-C, Alexander D, Schwartz S, Colthurst T, Ku A, Newburger D, Dijamco J, Nguyen N, Afshar PT, Gross SS, Dorfman L, McLean CY, DePristo MA. 2018. A universal SNP and small-indel variant caller using deep neural networks. *Nature Biotechnology* **36(10)**:983–987 DOI 10.1038/nbt.4235.

Povysil G, Tzika A, Vogt J, Haunschmid V, Messiaen L, Zschocke J, Klambauer G, Hochreiter S, Wimmer K. 2017. panelcn.MOPS: copy-number detection in targeted NGS panel data for clinical diagnostics. *Human Mutation* **38(7)**:889–897 DOI 10.1002/humu.23237.

Qi J, Zhao F, Buboltz A, Schuster SC. 2010. inGAP: an integrated next-generation genome analysis pipeline. *Bioinformatics* **26(1)**:127–129 DOI 10.1093/bioinformatics/btp615.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26(6)**:841–842 DOI 10.1093/bioinformatics/btq033.

Ren S, Ahmed N, Bertels K, Al-Ars Z. 2019. GPU accelerated sequence alignment with traceback for GATK HaplotypeCaller. *BMC Genomics* **20(S2)**:184 DOI 10.1186/s12864-019-5468-9.

Rimmer A, Phan H, Mathieson I, Iqbal Z, Twigg SRF, Wilkie AOM, McVean G, Lunter G, WGS500 Consortium. 2014. Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. *Nature Genetics* **46**:912–918 DOI 10.1038/ng.3036.

Rizvi NA, Hellmann MD, Snyder A, Kvistborg P, Makarov V, Havel JJ, Lee W, Yuan J, Wong P, Ho TS, Miller ML, Rekhtman N, Moreira AL, Ibrahim F, Bruggeman C, Gasmi B, Zappasodi R, Maeda Y, Sander C, Garon EB, Merghoub T, Wolchok JD, Schumacher TN, Chan TA. 2015. Cancer immunology. mutational landscape determines sensitivity to PD-1 blockade in non-small cell lung cancer. *Science* **348(6230)**:124–128 DOI 10.1126/science.aaa1348.

Rodríguez-Vicente AE, Díaz MG, Hernández-Rivas JM. 2013. Chronic lymphocytic leukemia: a clinical and molecular heterogenous disease. *Cancer Genetics* **206(3)**:49–62 DOI 10.1016/j.cancergen.2013.01.003.

Rossi D, Diop F, Spaccarotella E, Monti S, Zanni M, Rasi S, Deambrogi C, Spina V, Bruscaggin A, Favini C, Serra R, Ramponi A, Boldorini R, Foà R, Gaidano G. 2017. Diffuse large B-cell lymphoma genotyping on the liquid biopsy. *Blood* **129(14)**:1947–1957 DOI 10.1182/blood-2016-05-719641.

Roy S, Coldren C, Karunamurthy A, Kip NS, Klee EW, Lincoln SE, Leon A, Pullambhatla M, Temple-Smolkin RL, Voelkerding KV, Wang C, Carter AB. 2018. Standards and guidelines for validating next-generation sequencing bioinformatics pipelines: a joint recommendation of the association for molecular pathology and the college of american pathologists. *Journal of Molecular Diagnostics* **20(1)**:4–27 DOI 10.1016/j.jmoldx.2017.11.003.

Ruffalo M, LaFramboise T, Koyutürk M. 2011. Comparative analysis of algorithms for next-generation sequencing read alignment. *Bioinformatics* **27(20)**:2790–2796 DOI 10.1093/bioinformatics/btr477.

Sadedin SP, Pope B, Oshlack A. 2012. Bpipe: a tool for running and managing bioinformatics pipelines. *Bioinformatics* **28**:1525–1526 DOI 10.1093/bioinformatics/bts167.

Samorodnitsky E, Jewell BM, Hagopian R, Miya J, Wing MR, Lyon E, Damodaran S, Bhatt D, Reeser JW, Datta J, Roychowdhury S. 2015. Evaluation of hybridization capture versus amplicon-based methods for whole-exome sequencing. *Human Mutation* **36(9)**:903–914 DOI 10.1002/humu.22825.

Sater V, Viailly P-J, Lecroq T, Prieur-Gaston É, Bohers É, Viennot M, Ruminy P, Dauchel H, Vera P, Jardin F. 2020. UMI-VarCal: a new UMI-based variant caller that efficiently improves low-frequency variant detection in paired-end sequencing NGS libraries. *Bioinformatics* **36(9)**:2718–2724 DOI 10.1093/bioinformatics/btaa053.

Scherer SW, Lee C, Birney E, Altshuler DM, Eichler EE, Carter NP, Hurles ME, Feuk L. 2007. Challenges and standards in integrating surveys of structural variation. *Nature Genetics* **39(S7)**:S7–15 DOI 10.1038/ng2093.

Schneider VA, Graves-Lindsay T, Howe K, Bouk N, Chen H-C, Kitts PA, Murphy TD, Pruitt KD, Thibaud-Nissen F, Albracht D, Fulton RS, Kremitzki M, Magrini V, Markovic C, McGrath S, Steinberg KM, Auger K, Chow W, Collins J, Harden G, Hubbard T, Pelan S, Simpson JT, Threadgold G, Torrance J, Wood J, Clarke L, Koren S, Boitano M, Li H, Chin C-S, Phillippy AM, Durbin R, Wilson RK, Flicek P, Church DM. 2016. Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genomics* **16**:24 DOI 10.1101/072116.

Sengupta S, Gulukota K, Zhu Y, Ober C, Naughton K, Wentworth-Sheilds W, Ji Y. 2016. Ultra-fast local-haplotype variant calling using paired-end DNA-sequencing data reveals somatic mosaicism in tumor and normal blood samples. *Nucleic Acids Research* **44(3)**:e25–e25 DOI 10.1093/nar/gkv953.

Shiraishi Y, Sato Y, Chiba K, Okuno Y, Nagata Y, Yoshida K, Shiba N, Hayashi Y, Kume H, Homma Y, Sanada M, Ogawa S, Miyano S. 2013. An empirical Bayesian framework for somatic mutation detection from cancer genome sequencing data. *Nucleic Acids Research* **41(7)**:e89–e89 DOI 10.1093/nar/gkt126.

Sims D, Sudbery I, Ilott NE, Heger A, Ponting CP. 2014. Sequencing depth and coverage: key considerations in genomic analyses. *Nature Reviews Genetics* **15(2)**:121–132 DOI 10.1038/nrg3642.

SMALT: Wellcome Sanger Institute. 2020. SMALT. *Available at https://www.sanger.ac.uk/tool/smalt-0/* (accessed 18 December 2020).

Smith T, Heger A, Sudbery I. 2017. UMI-tools: modeling sequencing errors in unique molecular identifiers to improve quantification accuracy. *Genome Research* **27(3)**:491–499 DOI 10.1101/gr.209601.116.

Soukupova J, Zemankova P, Lhotova K, Janatova M, Borecka M, Stolarova L, Lhota F, Foretova L, Machackova E, Stranecky V, Tavandzis S, Kleiblova P, Vocka M, Hartmannova H, Hodanova K, Kmoch S, Kleibl Z. 2018. Validation of CZECANCA (CZEch CAncer paNel for clinical application) for targeted NGS-based analysis of hereditary cancer syndromes. *PLOS ONE* **13(4)**:e0195761 DOI 10.1371/journal.pone.0195761.

Steward DL, Carty SE, Sippel RS, Yang SP, Sosa JA, Sipos JA, Figge JJ, Mandel S, Haugen BR, Burman KD, Baloch ZW, Lloyd RV, Seethala RR, Gooding WE, Chiosea SI, Gomes-Lima C, Ferris RL, Folek JM, Khawaja RA, Kundra P, Loh KS, Marshall CB, Mayson S, McCoy KL, Nga ME, Ngiam KY, Nikiforova MN, Poehls JL, Ringel MD, Yang H, Yip L, Nikiforov YE. 2019. Performance of a multigene genomic classifier in thyroid nodules with indeterminate cytology: a prospective blinded multicenter study. *JAMA Oncology* **5(2)**:204 DOI 10.1001/jamaoncol.2018.4616.

Sulonen A-M, Ellonen P, Almusa H, Lepistö M, Eldfors S, Hannula S, Miettinen T, Tyynismaa H, Salo P, Heckman C, Joensuu H, Raivio T, Suomalainen A, Saarela J. 2011. Comparison of solution-based exome capture methods for next generation sequencing. *Genome Biology* **12(9)**:R94 DOI 10.1186/gb-2011-12-9-r94.

Talevich E, Shain AH, Botton T, Bastian BC. 2016. CNVkit: genome-wide copy number detection and visualization from targeted DNA sequencing. *PLOS Computational Biology* **12(4)**:e1004873 DOI 10.1371/journal.pcbi.1004873.

Tamborero D, Rubio-Perez C, Deu-Pons J, Schroeder MP, Vivancos A, Rovira A, Tusquets I, Albanell J, Rodon J, Tabernero J, De Torres C, Dienstmann R, Gonzalez-Perez A, Lopez-Bigas N. 2017. Cancer genome interpreter annotates the biological and clinical relevance of tumor alterations. *Cancer Biology* **13**:806 DOI 10.1101/140475.

Thankaswamy-Kosalai S, Sen P, Nookaew I. 2017. Evaluation and assessment of read-mapping by multiple next-generation sequencing aligners based on genome-wide characteristics. *Genomics* **109(3–4)**:186–191 DOI 10.1016/j.ygeno.2017.03.001.

Thorvaldsdottir H, Robinson JT, Mesirov JP. 2013. Integrative genomics viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in Bioinformatics* **14(2)**:178–192 DOI 10.1093/bib/bbs017.

Usuyama N, Shiraishi Y, Sato Y, Kume H, Homma Y, Ogawa S, Miyano S, Imoto S. 2014. HapMuC: somatic mutation calling using heterozygous germ line variants near candidate mutations. *Bioinformatics* **30(23)**:3302–3309 DOI 10.1093/bioinformatics/btu537.

Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, Del Angel G, Levy-Moonshine A, Jordan T, Shakir K, Roazen D, Thibault J, Banks E, Garimella KV, Altshuler D, Gabriel S, DePristo MA. 2013. From fastQ data to high confidence variant calls: the genome analysis toolkit best practices pipeline. *Current Protocols in Bioinformatics* **43(1)**:11 10 1–11 10 33 DOI 10.1002/0471250953.bi1110s43.

Wade MJ, Curtis TP, Davenport RJ. 2015. Modelling computational resources for next generation sequencing bioinformatics analysis of 16S rRNA samples. *ArXiv*. *Available at http://arxiv.org/abs/1503.02974*.

Wang K, Li M, Hakonarson H. 2010. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research* **38(16)**:e164 DOI 10.1093/nar/gkq603.

Wang Y, Waters J, Leung ML, Unruh A, Roh W, Shi X, Chen K, Scheet P, Vattathil S, Liang H, Multani A, Zhang H, Zhao R, Michor F, Meric-Bernstam F, Navin NE. 2014. Clonal evolution in breast cancer revealed by single nucleus genome sequencing. *Nature* **512(7513)**:155–160 DOI 10.1038/nature13600.

Wildeman M, Van Ophuizen E, Den Dunnen JT, Taschner PEM. 2008. Improving sequence variant descriptions in mutation databases and literature using the Mutalyzer sequence variation nomenclature checker. *Human Mutation* **29(1)**:6–13 DOI 10.1002/humu.20654.

Wilm A, Aw PPK, Bertrand D, Yeo GHT, Ong SH, Wong CH, Khor CC, Petric R, Hibberd ML, Nagarajan N. 2012. LoFreq: a sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets. *Nucleic Acids Research* **40(22)**:11189–11201 DOI 10.1093/nar/gks918.

Xu C. 2018. A review of somatic single nucleotide variant calling algorithms for next-generation sequencing data. *Computational and Structural Biotechnology Journal* **16(1)**:15–24 DOI 10.1016/j.csbj.2018.01.003.

Yu D, Liu Z, Su C, Han Y, Duan X, Zhang R, Liu X, Yang Y, Xu S. 2020. Copy number variation in plasma as a tool for lung cancer prediction using extreme gradient boosting (XGBoost) classifier. *Thoracic Cancer* **11(1)**:95–102 DOI 10.1111/1759-7714.13204.

Zhao M, Wang Q, Wang Q, Jia P, Zhao Z. 2013. Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives. *BMC Bioinformatics* **14(S11)**:S1 DOI 10.1186/1471-2105-14-S11-S1.

Zoi K, Cross NCP. 2015. Molecular pathogenesis of atypical CML, CMML and MDS/MPN-unclassifiable. *International Journal of Hematology* **101(3)**:229–242 DOI 10.1007/s12185-014-1670-3.

Zook JM, Catoe D, McDaniel J, Vang L, Spies N, Sidow A, Weng Z, Liu Y, Mason CE, Alexander N, Henaff E, McIntyre ABR, Chandramohan D, Chen F, Jaeger E, Moshrefi A, Pham K, Stedman W, Liang T, Saghbini M, Dzakula Z, Hastie A, Cao H, Deikus G, Schadt E,

Sebra R, Bashir A, Truty RM, Chang CC, Gulbahce N, Zhao K, Ghosh S, Hyland F, Fu Y, Chaisson M, Xiao C, Trow J, Sherry ST, Zaranek AW, Ball M, Bobe J, Estep P, Church GM, Marks P, Kyriazopoulou-Panagiotopoulou S, Zheng GXY, Schnall-Levin M, Ordonez HS, Mudivarti PA, Giorda K, Sheng Y, Rypdal KB, Salit M. 2016. Extensive sequencing of seven human genomes to characterize benchmark reference materials. *Scientific Data* **3(1)**:160025 DOI 10.1038/sdata.2016.25.