Charles University

Faculty of Science

Study program:

Molecular and cell biology, Genetics and virology

Mgr. Kateřina Kvapilová

Význam různých zdrojů a sekvenačních protokolů při zvyšování přesnosti NGS analýz v diagnostických aplikacích

The importance of different sources and sequencing protocols in increasing the accuracy of NGS analysis in diagnostic applications

Doctoral thesis

Supervisor: Zbyněk Kozmik, Ph.D.

Prague, 2024

Prohlášení:

Prohlašuji, že jsem závěrečnou práci vypracovala samostatně, pouze za použití citované literatury a dostupných informačních zdrojů. Tato práce ani žádná její podstatná část nebyla předložena k získání jiného nebo stejného akademického titulu. Souhlasím s uložením elektronické verze mé práce v databázi "Theses.cz."

V Praze dne 21.3.2024

……..……

Kateřina Kvapilová

ACKNOWLEDGEMENTS

I would like to thank my supervisor Dr. Zbyněk Kozmik and co-supervisor Dr. Jan Pačes. Both of them believed in me throughout. Thanks to their support, I was able to work with excellent people who taught, helped, and guided me during my studies.

I would also like to thank my work colleagues who have always been there to help me – especially Pavel Mišenko and Martin Kašný. In addition, the contributions of Ondra Brzoň and Ján Radvanský did not go unnoticed and were greatly appreciated.

Finally, my husband Petr deserves special thanks for being an endless source of support and inspiration.

K

Table of content

ABSTRACT

Continued advances in next-generation sequencing (NGS) technology, such as capacity, speed, and reduced cost per sequenced base, revolutionize personalized medicine and bring genomics into routine clinical practice. Nevertheless, NGS is still under rapid development, and the variability of sequencing protocols and validation procedures is one of its current bottlenecks.

This thesis aimed to study the influence of different sample sources and NGS protocols (NGS library construction-sequencing-data analysis) on the accuracy of NGS analysis in diagnostic applications. In the first study, performed during the COVID-19 pandemic outbreak, we developed NGS protocols suitable for a whole genome analysis of the SARS-CoV-2 virus. Subsequently, in the second study, we examined the suitability of human saliva-derived gDNA for genomic/genetic analysis of selected variant types compared to traditional blood-derived gDNA using validated NGS protocol and statistical comparison of the obtained data.

Whole genome analysis of the SARS-CoV-2 genome was performed using two captured-based approaches and one amplicon-based approach to study the quality and effectivity of the respective NGS protocols. Synthetic controls were employed to verify the accuracy and specificity of the developed NGS protocols. We proved that real-time quantitative PCR-based quantitation of viral load was the right tool since subsequent sample plexing utilizing cycle threshold values resulted in sequencing data with required coverage uniformity between different samples. We found the capture-based NGS protocol the most suitable for whole genome analysis of the SARS-CoV-2 genome.

In the main study, we analyzed whether human saliva may serve as an equal source of gDNA to blood for single nucleotide (SNV) and small insertion and deletions (small-indels) variant analysis. We designed and validated entire NGS protocols for whole exome (WES) and whole genome (WGS) analysis employing the Coriel NA12878 standard sample and the latest human reference genome GRCh38. Consequently, we analyzed NGS data from 10 paired blood-saliva samples obtained by engaging the same Coriel NA12787 NGS protocol, using statistical analysis tools on the F1 score and other selected sequencing parameters. For the WES protocol, the median value of F1 score for ten paired blood-saliva samples was 0.9858 for SNVs and 0.9076 for small-indels. For the WGS protocol, the median value of F1 was 0.9761 for SNVs and 0.9511 for small-indels. The study's comprehensive results demonstrated a high level of concordance between blood and saliva samples compared to Coriel standard results for F1 scores in the case of SNV and small-indels and for both the WES and WGS NGS protocols, respectively.

These studies advanced our understanding of genome sequencing of samples of a different origin and proved saliva as a suitable source of genomic/genetic data comparable to blood. These findings affect further genomic/genetic research and NGS clinical applications.

Souhrn

Neustálé pokroky v technologii sekvenování nové generace (NGS), jakými jsou kapacita, rychlost a snížené náklady na sekvenovaní, vedou k revoluci v personalizované medicíně, tím, jak postupně přináší genomiku do rutinní klinické praxe. NGS se stále rychle vyvíjí, avšak variabilita sekvenačních protokolů a validačních postupů je jedním z jeho problematických rysů.

Cílem této práce bylo zhodnotit význam různých zdrojů vzorků a sekvenačních protokolů (od přípravy sekvenační knihovny-sekvenování-analýzi dat) pro zvýšení přesnosti NGS analýzy v diagnostických aplikacích. V první studii, provedené během vypuknutí pandemie COVID-19, jsme vyvinuli NGS protokoly vhodné pro analýzu celého genomu viru SARS-CoV-2. Následně jsme ve druhé studii zkoumali vhodnost lidské genomické DNA (gDNA) pocházející ze slin pro genomickou/genetickou analýzu vybraných variantních k tradiční gDNA pocházející z krve pomocí validovaného NGS protokolu a statistického srovnání získaných dat.

K ověření účinnosti sekvenačních protokolů pro analýzu celého genomu SARS-CoV-2 byly použity dva sekvenační přístupy založené na zachycení části genomu a jeden sekvenanční přístup založený na amplifikaci částí genomů. K ověření přesnosti a specifičnosti vyvinutých NGS protokolů byly použity syntetické kontroly. Prokázali jsme, že kvantifikace vzorků pomocí kvantitativní PCR v reálném čase byla správným nástrojem pro následné plexování vzorků. Využití prahových hodnot cyklu vedlo k sekvenačním datům s požadovanou uniformitou pokrytí mezi různými vzorky. Zjistili jsme, že NGS protokol založený na zachycení je nejvhodnější pro celogenomovou analýzu genomu SARS-CoV-2.

V hlavní studii jsme analyzovali, zda lidské sliny mohou sloužit jako alternativní zdroj gDNA ke krvi pro analýzu jednonukleotidových záměn (SNV) a malých inzercí a delecí (indely). Navrhli jsme a ověřili NGS protokoly pro analýzu celého exomu (WES) a celého genomu (WGS) s použitím standardního vzorku Coriel NA12878 a nejnovějšího lidského referenčního genomu GRCh38. Následně jsme analyzovali NGS data z 10 párových vzorků krve a slin získaných za použití stejného Coriel NA12787 NGS protokolu s použitím nástrojů statistické analýzy pro F1 skóre a dalších vybraných sekvenačních parametrů. Medián F1 skóre WES protokolu pro deset párových vzorků krví a slin dosahoval 0,9858 pro SNV a 0,9076 pro indely; pro WGS protokol pak medián F1 skóre dosahoval 0,9761 pro SNV a 0,9511 pro indely. Výsledky této komplexní studie prokázaly vysokou úroveň shody mezi vzorky krví a slin ve srovnání se standardními výsledky Coriel pro F1 skóre pro SNV a indely a jak pro WES, tak WGS protokol.

Obě studie posunuly naše chápání genomového sekvenování vzorků různého původu a prokázaly, že sliny jsou vhodným zdrojem genomických/genetických dat srovnatelných s krví. Tato zjištění ovlivňují další genomický/genetický výzkum a klinické aplikace NGS.

1 INTRODUCTION

1.1 Human Genome Sequencing

Since breakthrough studies done by Albrecht Kossel and his students, who were the first to identify five purine and pyrimidine nitrogenous bases as the basic building blocks of nucleic acids, carried out between 1885 – 1901 [1], more than eighty years passed until the first complete genome, bacterial virus ΦX174, was completely sequenced [2]. In 1984, the Epstein-Barr virus was sequenced, it was the first genome assembled without prior knowledge of the genetic profile and this demonstrated the feasibility of the assembly of short sequence fragments into complete genomes [3]. Only one year later, in 1985, after the stunning success of finding a genetic link to Huntington's disease [4], discussion of the ambitious idea of sequencing the entire human genome began [5][6].

Finally, in 1990, under the leadership of Dr. Francis Collins the significant scientific project, The Human Genome Project (HGP), was officially announced [7]. The HGP set several sub-goals which were gradually achieved [8]; see Table 1. The first draft of the sequences and analyses was published in February 2001 on behalf of the Human Genome Sequence Consortium [9] and Celera Genomic [10]. On April 14, 2003, the National Human Genome Research Institute announced the completion of the HGP [11] but it took almost three more years to analyze the sequences from the capillary Sanger-based method and complete at least 99 % of the euchromatin portion (containing 2.85 billion nucleotides interrupted by only 341 gaps) known as reference genome Build 35. New sequencing methods were needed to fill those 341 gaps in the genome that had not been sequenced [12]. In total, 19 countries and more than 200 laboratories participated in the project with a final cost of approximately $2.7 billion [13]. A peculiarity of the project was that the DNA of a single individual was not sequenced, but a composite of a small number of individuals was assembled into a haploid DNA sequence. The HGP brought at least two surprising outcomes. First, the representation of protein-coding genes in humans is approximately the same as that in small mammals [14]. Secondly, the human genome has significantly more segmental duplications than was suspected [12]. The first human genome released by the Human Genome Project, was actually a pangenome, not a single individual. In 2007, sequencing of the first genome of a specific individual (Craig J. Venter) was announced. Venter's genome was sequenced using the capillary Sanger-based sequencing method utilizing the shotgun strategy, which took almost 15 years [15]. Just one year later, in 2008, the third human genome (that of James D. Watson) was the first to be sequenced using next-generation sequencing (NGS) technology, which uses massively parallel DNA sequencing [16]. It only took two months to sequence Dr. Watson's DNA (7.4x average coverage, 454 Genome

Sequencer FLX) at a cost of less than $1 million, which was 1/100th of the cost when compared to capillary Sanger-based sequencing [17]. However, the cost of sequencing an individual human genome was still so high that the concept of personalized diagnostics based on DNA analysis was unreachable.

Area	HGP Goal	Goal Achieved	Date Achieved
Genetic Map	2- to 5-cM resolution map (600 – 1,500 markers)	1-cM resolution map (3,000 markers)	1994
Physical Map	30,000 STSs	52,000 STSs	1998
DNA Sequence	95% of gene-containing part of human sequence finished to 99.99% accuracy	99% of gene-containing part of human sequence finished to 99.99% accuracy	2003
Capacity and Cost of Finished Sequence	Sequence 500 Mb/year at < $0.25 per finished base	Sequence >1,400	2002
		Mb/year at <$0.09 per finished base	
Human Sequence Variation	100,000 mapped human SNPs	3.7 million mapped human SNPs	2003
Gene Identification	Full-length human cDNAs	15,000 full-length human cDNAs	2003
Model Organisms	Complete genome sequences of *E.coli, S. cerevisiae, C. elegans, D. melanogaster*	Finished genome sequences of *E. coli* , *S. cerevisiae* , *C. elegans* , *D. melanogaster* , plus whole-genome drafts of several others, including *C. briggsae* , *D. pseudoobscura* ,mouse and rat	2003
Functional Analysis	Develop genomic-scale technologies	High-throughput oligonucleotide synthesis	1994
		DNA microarrays	1996
		Eukaryotic, whole-genome knockouts (yeast)	1999
		Scale-up of two-hybrid system for protein-protein interaction	2002
Source: Science. 2003 Apr 11;300(5617):286-90. doi: 10.1126/science.1084564			
Table 1: Milestones of HGP			

To reduce the cost and increase the speed of genome sequencing and analysis, targeted exome sequencing was introduced [18]. This novel approach was able to identify candidate genes for Mendelian disorders. [19]. After 2007, thanks to improvements in NGS technology, which resulted in a cost reduction per sequenced base [20], studies using NGS technology proved the clinical value of whole exome and whole genome sequencing [21][22][23]. Some studies also discussed the importance of the sequencing method chosen, the analysis pipeline, and the need for validation methodologies for both the WES and WGS approaches [24][25]. However, it took a long time for it to be demonstrated that whole-exome and whole-genome sequencing provided critical data for diagnosis and that it was a useful and cost-effective approach to clinical guidance and the selection of an appropriate treatment protocol [26].

1.2 From First to Fourth Generation Sequencing Technology

In the 1970s, first-generation sequencing was represented by two methods: Sanger or Maxam-Gilbert sequencing. Neither of these methods could have been developed without prior knowledge of the structure of DNA [27], cloning [28] and polymerase activity [29]. In 1975, Frederic Sanger and his team developed a sequencing method, was used to sequence bacteriophage ΦX174 [30], known as the chain termination method, and only two years later he updated his method so that it was faster and more accurate [31]. A detailed description of the Sanger sequencing technology can be found in Chapter 1.3, Sanger vs. Illumina NGS technology. A rival sequencing method, developed by Allan Maxam and Walter Gilbert, known as chemical sequencing, was introduced in 1976 [32]. Maxam-Gilbert sequencing is based on the partial chemical modification of terminal nucleotides on single-stranded DNA (ssDNA) and the subsequent cleavage of the ssDNA at sites adjacent to the modified nucleotides. The updated Sanger sequencing method was, in comparison to Maxam-Gilbert sequencing, more "user-friendly" plus, it did not require the use of hazardous chemicals and thus became the dominant technique for determining the DNA sequence for the following three decades. As early as 1980, capillary electrophoresis technology was introduced[33]. In 1987, Applied Biosystems (later acquired by Perkin-Elmer company, now Thermo Fisher) launched the first PAGE gel-based automated sequencer (ABI PRISM AB370A, 16 samples up to 450 base pairs (bp) long fragments) on the market and in 1995 the same company subsequently launched the first automated capillary sequencer (ABI PRISM 310, 1 sample up to 600 bp long fragments). In 1998, Applied Biosystems launched the ABI PRISM 3700 with 96 capillaries which allowed faster analysis with more accurate results and allowed the sequencing of even longer DNA fragments (up to 800 bp) [34].These systems played a crucial role in the successful completion of the HGP, since by 1998, only 6 % of the human genome had been sequenced.

Significant advancements in sequencing came in the mid-2000s with the development of next-generation sequencing (NGS) technologies, also called massively parallel or second-generation sequencing. NGS technologies allow massive parallel sequencing, drastically increase throughput and make sequencing more affordable. The differences between the first- and second-generation sequencing are in the technology used to read the sequences. First, NGS sequencing of fragmented genomic DNA or copy DNA is accomplished without prior cloning of DNA fragments into a host cell. Clonal amplification *in vivo* switched to *in vitro* polymerase chain reaction (PCR) amplification. Another difference is the construction of NGS libraries where ligated adapter sequences are added to DNA fragments and the NGS libraries are subsequently amplified on a solid surface (cluster formation) or beads. Second, nucleotide incorporation into sequenced reads is directly monitored by luminescence detection or by changes in electrical charge during the sequencing process [35]. Whereas some next generation sequencing technologies have not been widely implemented on the market [36], at least five sequencing technologies were in use at the time of their launch or are still in use today, such as pyrosequencing technology, represented by 454 Life Sciences, launched in 2005 (later acquired by Roche) [37] and the technology launched by Solexa in 2008 (acquired by Illumina, 2006) [38]. Pyrosequencing and the Illumina technology are both based on the sequencing-by-synthesis principle. Sequencing using ligation technology, utilized by Applied Biosystems (now Thermo Fisher) [39], is no longer in use. On the contrary, today semiconductor sequencing technology, represented by Ion Torrent [40] (now Thermo Fisher), is, next to Illumina technology, the second largest player in the NGS market. The fifth player in the NGS field is technology from Complete Genomics (acquired by BGI, now MGI) and its DNA nanoball sequencing technology that utilizes the sequencing by ligation principle [41]. Complete Genomics is fully oriented towards human genome sequencing, with their awareness of the complexity of sequencing the human genome, in 2019 they came up with the single-tube long fragment read technology for assembling long reads from short read sequences [42] as an alternative to third-generation sequencing.

In parallel with the second-generation sequencing technology, third-generation sequencing technology, also known as single molecule sequencing or long-reads sequencing [36], is coming onto the market. The first single molecule sequencing technology that avoids the need for DNA amplification on a solid surface or beads was adopted by Helicos BioSciences [43]. Nowadays, third-generation sequencing is represented by the Pacific Bioscience Single Molecule Real-Time (SMRT) sequencing technology [44] and Oxford Nanopore sequencing technology (ONT) [45] which both allow real-time single-molecule sequencing. Contrary to SMRT, ONT is also able to sequence short fragments. Both technologies can produce much longer reads than the second-generation sequencing

technologies which is beneficial in *de-novo* genome assembly [46] [47], haplotype phasing [48] [49], epigenetics research [50] [51] or direct RNA sequencing with ONT [52]. The data generated using third-generation sequencing technology has higher error rates [46] [53] than previous sequencing technologies, but this issue can be solved successfully through long read self-correction [54] or hybrid correction by short-reads [55] [56].

The field of nucleic acid sequencing is constantly evolving with new short-read sequencing technologies and many new companies that claim to offer the best sequencing technology at the best price. In 2022 Ultima Genomics introduced mostly natural sequencing-by-synthesis technology through their Ultima system [57]. Element Bioscience introduced a sequencing system based on avidity sequencing by synthesis technology [58] in 2023, promising both better data accuracy and a decrease in the cost per human genome. Complete Genomics (acquired by BGI, now MGI) launched a new sequencer, DNBSEQ-T20×2RS, with a claimed cost per human genome of below $100 [59]. The Singular Genomics G4 sequencer platform is intended for use in clinical laboratories as an accurate and cost-effective alternative for basic NGS applications [60].

1.3 Sanger vs. Illumina NGS technology

Since the thesis is based on the Illumina sequencing technology, this chapter compares the Sanger sequencing technology to the Illumina NGS technology. Early Sanger sequencing technology used radioactively labeled primers, polymerase, and a mix of deoxynucleotides triphosphates (dNTP) plus di-deoxynucleotide triphosphates (ddNTP) "to read" the DNA sequence from an agarose gel. The breakthrough in Sanger sequencing came with the development of PCR [61] and dideoxynucleotide fluorescent dye chemistry [62]. Radioactively labeled primers were replaced by fluorescently labeled primers and the analysis was transferred from agarose gel to an automated sequencer [63]. Under an electric field within a capillary tube, fluorescently labeled DNA fragments are separated based on size. The DNA fragments pass through a capillary and, at the end of it, meet a laser. The laser-excited dye emits light at a characteristic wavelength that is subsequently detected by a light sensor. The software interprets this light signal and translates it into a base call [64]. Nowadays, up to 96 DNA fragments (reads) of 800 bp in length can be sequenced in a single capillary electrophoretic instrument, with a total data output of approximately 76,8 Kbp [65]. In contrast the largest sequencing platform, the NovaSeq X Plus from Illumina, can generate data up to 16 Tbp [66]. Sequencing the human genome with a coverage of one (approx. 3.0 Gbp) by Sanger sequencing requires about 40 000 runs on 96 capillary sequencers, but NovaSeq X Plus can sequence 128 human genomes with 30x coverage in a single run.

NGS Illumina technology utilizes sequencing by synthesis, which is a further development of Sanger sequencing based on the polymerase synthesis of the second strand of DNA fragment. Still, instead of dideoxynucleotide terminators, additional dye-labeled-nucleotides, so-called "reversible terminators "are incorporated into the growing second strand of DNA. The difference in Sanger sequencing is that prior to sequencing, the DNA fragments, locked in NGS libraries, are amplified by PCR (so-called "clustering") on the solid surface of the sequencing flow cell which allows it to be amplified into clusters of identical molecules. Laser-induced excitation of the fluorophore is used to determine the last base incorporated into the growing chain of the unique cluster. Data analyzer software is used to compute the images to identify the final base [38]. The key differences between Sanger sequencing and Illumina NGS sequencing are 1) throughput capacity, 2) the length of the sequenced DNA fragments, 3) sequence accuracy, and 4) cost-effectiveness. Using Sanger sequencing based on capillary electrophoresis, it is possible to accurately sequence up to 96 DNA fragments with a length of 600 to a 1000 base pairs (bp) with a quality score of Q40 (99,99 % base call accuracy) [64] in a single run. This makes Sanger sequencing the preferred choice for the examination of a small number of DNA fragments. In contrast, Illumina NGS technology can sequence fragments from 35 bp up to 600 bp with a quality score of Q30 (99,9% base call accuracy). Illumina declares Q30>85% per 2x150 bp across its sequencing platforms (excluding iSeq 100, Mini|Seq, NextSeq 500/550 series [66], but in reality, the quality score regularly reaches Q35 to Q37 [67], at a large scale – up to millions of fragments in a single run. The ability to analyze a large set of genes/genomes simultaneously clearly makes Illumina NGS sequencing more cost-effective than Sanger Sequencing [9] [36]. Also, its high throughput makes the Illumina NGS an ideal choice for large sequencing projects [68].

Due to the high-quality data output, for many laboratories, Sanger sequencing remains the gold standard for the independent validation of variants detected by NGS. A study carried out by Biesecker et al., [69] raised questions about the reliability of a single round of Sanger sequencing for NGS validation. Instead of NGS validation by Sanger sequencing, the emphasis is now being placed on the importance of appropriate internal quality controls in NGS workflows to detect human and technical errors [70].

1.4 General Sequencing Approaches

1.4.1 Whole Genome Sequencing

Whole genome sequencing allows the comprehensive analysis of entire genomes, both RNA based [71] or DNA based [72] [73]. The entire genome sequence in eukaryotic cells includes

the mitochondrial DNA or chloroplast DNA in plants. WGS allows genomic analysis without prior knowledge of specific targets, making it the first choice for *de-novo* genome assembly from scratch without the support of reference genomic data [74], creating reference standards for specific population cohorts [75], or genomic-based taxonomy [76]. In human diagnosis, WGS is the recommended approach for rare disease diagnosis in preference to WES or small panel sequencing [77]. A recent study demonstrated the clinical value to patients of the simultaneous detection of somatic and germline variants by WGS in oncology patients, providing a comprehensive understanding of the genetic landscape and treatment implications [78] including information about the pharmacogenomic profile [79]. Yet, there are a few challenges when using WGS in human diagnostics. First, the data analysis is challenging due to the enormous amount of data and false-positive results that might be reported. Moreover, WGS data can be re-interpreted in the light of future clinical findings. The periodic re-analysis of negative results can bring clinical value, but it is time and management-consuming [80].

Whole genome sequencing is a suitable approach for wild population studies. The 1000 Genomes Project**,** the first international project that focused on the variability of the human genomes across populations demonstrated the need for a large data set to correctly interpret the data obtained [81]. This project examined 2504 human genomes from 26 populations utilizing a genotyping array-based approach, whole exome sequencing, and low-coverage whole genome sequencing. Population studies only based on a WGS approach have been conducted and are in progress on all continents to meet the goal to create a local reference genome to identify the genetic variability of that population [82] [83] [84]. Also, a large multi-population project is running at the pan-European level. The Million European Genomes Alliance (MEGA) initiative aims to obtain one million genomic sequences from 40 European populations by 2027 [85]. The above projects aim to create a local reference genome and, therefore, prioritize the analysis of the genomes of healthy individuals. In contrast, the UK 100 000 Genome Project, also known as the 100K UK project, analyzed the genomes of approx. 84 000 patients affected by rare diseases or cancers to study the role of genes in health and disease [86]. Under the umbrella of the 100K UK project, more than one hundred thousand genomes were analyzed, with a focus on different conditions and aspects of genome analysis [87].

1.4.2 Targeted Sequencing

Targeted sequencing generates sequencing reads for the regions of interest, which decreases the data output requirements per sample (compared to WGS) and, via multiplexing, increases the ability to

scale the sequencing experiment. There are two types of targeted sequencing: capture-based or amplicon-based. Capture-based sequencing involves hybridization and the subsequent capture of specific genomic regions using specifically designed probes. Amplicon-based sequencing involves amplification of specific genomic regions using pre-defined primers. Whether an amplicon-based or capture-based approach is used for a particular region of the genome is influenced by the primary structure of the nucleic acids. If some target regions are difficult for primers designed for amplicon-based assays, longer probes are designed for capture-based assays. On the other hand, the amplicon approach has a simpler workflow and requires smaller amounts of input DNA. Both methods allow for deep sequencing [88] of specific regions, making them useful for detecting rare somatic variants [89] or genetic mosaicism [90].

Both of these targeted sequencing approaches take priority in certain applications. Amplicon sequencing is most suitable for metagenomics analysis [91], molecular breeding or genotyping [92] or CRISPR screening [93]. Amplicon sequencing can be also used for whole genome analysis for small genomes [94] [95]. Capture-based sequencing is more suitable for whole exome sequencing [19], gene panels for cancer research and diagnosis [96] or gene panels for infectious disease surveillance [97]. In contrast to whole exome sequencing (both coding exonic regions and non-coding exonic regions in varying proportions depending on the manufacturer of the equipment, but in general it is approx. 2 % of the genome [98]) there is clinical exome sequencing (CES), which is only focused on disease-associated relevant genes (genes associated to known clinical phenotypes; approx. 1 % of the genome [99]) and the targeted size of CES can vary depending on the hospital/laboratory requirements. Usually, the list of genes for a clinical exome come from the Online Mendelian Inheritance in Man database (OMIM) (https://www.omim.org/) or the Human Gene Mutation Database (HGMD) (https://www.hgmd.cf.ac.uk/ac/index.php). CES for patients with Mendelian phenotypes shows high clinical utility compared to WES [100].

Targeted sequencing has at least two limitations. First, it requires knowledge of the regions of interest and the primers/probes must be designed using this knowledge [101]. Further, targeted sequencing requires good quality data output in terms of the portion of covered bases and the uniformity of sequencing reads. Uniform read coverage enables an analysis of the maximum number of targeted regions utilizing a low sequencing capacity. Finally, both targeted approaches used for deep sequencing often utilize unique molecular indexes/identifiers to avoid PCR bias [102].

1.4.3 RNA Sequencing

Before the advent of NGS technology, the most affordable gene expression technology was DNA microarray technology [103] [104]. Unfortunately, the limitation of DNA microarray is its reliance on prior sequence information for probe design and issues with cross-hybridization and background signals. NGS technology significantly changed RNA research, making RNA-sequencing (RNA-seq) the preferred method for gene expression analysis and identification of new types of RNA, both utilizing direct sequence identification. RNA-seq is used to analyze continuous changes in cellular coding and non-coding transcriptome and can detect both known and novel RNA transcripts, providing a more comprehensive view of the transcriptome [105]. Although direct sequencing of both poly(A) RNA and non-poly(A) tailed-RNA is available using Nanopore technology [106] most RNA-seq application are based on cDNA NGS library preparation. For transcriptome analysis it is also possible to use the Iso-Seq method with PacBio SMRT sequencing technology which allows the sequencing of full-length cDNA to discover novel transcripts [107].

The commonest approach to RNA-seq is gene expression analysis to study the changes in gene expression levels under different conditions [108] [109]. Gene expression analysis utilizes a single-end sequencing strategy (sequencing only the first read from the NGS library), which only provides the necessary YES/NO information. This approach is not limited to known genes, but masks sample heterogeneity, for this reason RNA-seq is very often called "bulk RNA-seq". On the contrary, single-cell RNA-seq (scRNA-seq) allows the assessment of sample heterogeneity to single-cell resolution, but sample preparation is quite challenging [110]. Bulk RNA-seq or scRNA-seq is performed to quantify the overall mRNA abundance. After the re-evaluation of "junk DNA" [111] non-coding RNA is widely studied using NGS. It involves transfer RNA[112] [113], ribosomal RNA [114], microRNA [115], small-interfering RNA [116] [117] or piwi-interacting RNA [118] and long non-coding RNA sequencing analysis [119] [120]. All types of RNA-seq, regardless of the read length, help to refine the genome annotation [121].

1.4.4 Methylation Sequencing

Methylation sequencing is used to analyze DNA methylation, a key mechanism in epigenetic regulation for both prokaryotes and eukaryotes. N6-methylated adenine (6mA), N4-methylcytosine (4mC) and 5-methylcytosine (5mC) in prokaryotes protects DNA from restriction digestion in a cellular defense pathway and play a role in the regulation of gene expression. Mammals mainly control epigenetics with 5-methylcytosine and 5- hydroxymethylcytosine (5hmC). 5mc is the most common and most studied methylation modification in animals and plants. DNA methylation in

humans is essential for genetic imprinting, X chromosome inactivation [122] or suppression of transposable elements [123]. Aberrant DNA methylation is associated with many diseases, including cancer, autoimmune diseases, inflammatory diseases, and metabolic disorders [124] [125]. Genome-wide hypomethylation is linked to chromosomal instability [126] and hypermethylation of promoters inhibits gene transcription [127].

Bisulfite sequencing is a two-step protocol used to detect 5-methylcytosine by sodium disulfate treatment in single stranded DNA. In the first step, unmethylated cytosines are converted to uracil through deamination using sodium bisulfate. Because sodium bisulfate does not affect methylated cytosines, they remain unchanged. PCR amplification with uracil-tolerant polymerase is then used to amplify the DNA fragments. The amplified DNA has thymine in place of the original unmethylated cytosines and cytosine in place of the original methylated cytosines (5mC). Those DNA fragments are sequenced and compared to the bisulfite-converted reference genome to determine the methylation pattern [128] [129].

The basic approach to bisulfite sequencing is methylation pattern analysis [130] and genomic imprinting analysis [131], and within cancer research it is used to identify new biomarkers [132]. The limitation of bisulfite sequencing lies in its inability to differentiate between 5-methylcytosine and 5-hydroxymethylcytosine, as well as its inability to detect other forms of methylated base modifications in DNA, such as N6-methyladenine, 5-bromodeoxyuridine or N6-methyladenosine in RNA. Methylation sequencing by Nanopore technology directly allows the identification of DNA and RNA base modifications, including 5mC, 5hmC, 6mA, and 5-bromodeoxyuridine in DNA and N6-methyladenosine in RNA plus it also adds the value of long reads. The PacBio long-read technology, similar to Nanopore, allows direct methylation pattern analysis [46] [47] [129].

1.5 Illumina NGS Library Construction

NGS library preparation is a process that involves the conversion of a gDNA/copy DNA into a fragment which can then be sequenced. NGS library preparation is constantly evolving with the introduction of simple and robust workflows, so the experience of NGS library preparation may vary from year to year. Although each sequencing approach has different requirements for NGS library preparation, there are general requirements for all approaches that are described in Chapter 1.4, General Sequencing Approaches. NGS library preparation consists of four steps: DNA fragmentation, adapter ligation, library amplification, and final quality check, including the data analysis as described in Chapter 1.7, Bioinformatics – Sequencing Data Analysis.

In the case of WGS, targeted or methylation sequencing, the recommended input material for NGS library construction is high-quality DNA, 50 ng – 1 μg [133], which must be cleaved into fragments that are approximately 350-550 bases long [134]. A specific sequencing application determines the optimal fragment size. Fragmentation is usually done mechanically (sonication) or enzymatically (non-specific endonuclease mix) [135]. An alternative enzymatic method for enzymatic fragmentation is tagmentation, which utilizes enzyme transposase that simultaneously fragments and inserts adapters into dsDNA [105] [135]. After fragmentation, fragment size selection results in DNA fragments of a defined length. For fragment size selection magnetic beads, spin-up columns, or gel digestion can be used [134]. In targeted sequencing, the regions of interest may be captured and enriched by specifically designed probes or amplified by PCR with custom-designed primers [135].

The Illumina technology does not allow the direct sequencing of RNA. Therefore, it is necessary to transcribe the RNA into copy DNA (cDNA), which is the starter material for the creation of the NGS library for RNA-seq. To transcribe RNA into cDNA, reverse transcription by poly(T) primers, or random hexamer primers are used [136]. Full-length cDNA is then fragmented, as previously described above for gDNA. For long RNA, the RNA molecules can be directly fragmented by hydrolysis. RNA fragmentation is not required for samples with low RNA integrity numbers, typically RNA extracted from formalin-fixed, paraffin-embedded (FFPE) tissue samples. RNA library preparation kits designed for FFPE samples usually utilize probes targeted at known regions to enrich RNA fragments [137].

Adapters (approx. 80 bases in length) attached to the DNA fragments are constructed from three parts: 1) flow-cell binding sequences, 2) primer sequences for amplification, and 3) barcode sequences. Adaptors can be added to both ends of DNA fragments through a ligation step (which involves the repair of the ends of DNA fragments, adding adenosine at the 3′ ends of blunted DNA fragments, and then ligation of the adapters by thymine overhang at the 5′ ends of the adaptor sequence) or during the first step of library preparation if the fragmentation was done through tagmentation. The barcode sequence in the adapter is used for sample indexing. There are many versions of indexing strategy; indexes are used to identify samples to allow many samples to be mixed in a single sequencing experiment (multiplexing). A special kind of index - unique molecular identifiers (UMI) can be used to identify the sample at the level of the individual fragment; see Figure 1. UMI is often used in RNA-seq facilitating gene expression quantification and scRNA-seq to distinguish the molecules from one cell. UMI serves to eliminate PCR bias, as well as PCR duplicates

to accurately identify rare variants. In RNA-seq, the correct choice of adaptors and PCR amplification strategy also helps to distinguish strand specificity. The correct strategy used , single or dual index, also depends on the single-end or pair-end sequencing. Single-end sequencing is mostly used in the cost-effective analysis of gene expression analysis or CHIP-seq analysis. Pair-end sequencing generates high quality data suitable for *de-novo* assembly and variant analysis. Pair-end sequencing requires the same amount of DNA input as single-end sequencing.

Figure 1: NGS library adaptors source: https://www.lubio.ch/blog/ngs-adapters, modified

All NGS libraries must be amplified by PCR before sequencing, with the exception of those prepared using PCR-free library preparation kits. The number of amplification cycles varies with the sequencing approach but typically between 8 and 12 cycles are required to reach the final concentration. Unfortunately, there are several biases incorporated by PCR amplification [138]. PCR bias and CG bias are two phenomena that influence data output. PCR bias refers to the uneven amplification of DNA fragments during NGS library construction. Factors that influence PCR bias include primer design, template sequence, and polymerase efficiency. PCR bias can lead to over- or under-representation of specific genomic regions in the read coverage. GC bias refers to the uneven representation of DNA fragments based on their GC content. GC bias also affects the uniformity of read coverage across the genome, but it is only related to CG content [139].

Quality checks serve to estimate the NGS library concentration and DNA fragment length. Sample quantification is crucial when several samples are pooled within a single sequencing experiment. It has a direct, positive, impact on the quality of the NGS data in terms of even read distribution. Differences in sample molar concentration between the pooled NGS libraries can result in a lack of coverage of the libraries with lower concentrations. NGS libraries are checked twice, both before and after dilution to sequencing loading concentration. The most common approach to quantify NGS libraries is quantitative PCR (qPCR), which determines the concentration of only the amplifiable DNA fragments. In addition to concentration, the size distribution of the NGS library must be checked using an electrophoretic system. In the presence of adapter dimers or with an inappropriate size of NGS library fragments, both can be treated by spin column systems to select fragments of the correct size.

The preparation of NGS libraries requires a good knowledge of the NS library preparation protocol and, thus, an active approach towards the possible prevention and elimination of the potential occurrence of low-quality NGS libraries, but it also requires skill when working with minimal volumes and concentrations. Nowadays, most library preparation kits can be adjusted to allow automation. Automation of NGS library preparation reduces hands-on time, however, automation is only suitable above a certain number of samples, and its implementation in the laboratory requires enough time to be allocated for optimization [134].

1.6 Human Reference Genome

A reference genome (reference assembly) is a digital database of the DNA sequences of a particular organism. Since the reference genome is made by combining DNA sequences from several different donors, the reference assembly does not precisely show the genes of a single individual. The biggest goal and the flagship outcome of the Human Genome Project was a reference human genome [140]. The path to the creation of a complete reference human genome was not straightforward; see Table 2. It began in 1982 when the GenBank database was established to collect all the publicly available nucleotide sequences [141]. The National Center for Biotechnology Information (NCBI) took oversight of GenBank in 1992 [142] and in 1999 the NCBI established a new database, the dbSNP which stood alongside the GenBank database [143]. The goal of the dbSNP is to hold all the identified genetic variations (SNVs, indels, short tandem repeats or microsatellites, ...).

Before the first drafts of the human genomes were assembled in 2001, the "academic" section of the HGP agreed to gradually publish the generated sequence data in the international DNA databases; one of which was GenBank [144]. In 2001 the "academic" section of the HGP published

sequences based on a mapping-based approach using bacterial artificial chromosomes [9], this in comparison to Celera Genomic, the "commercial" section of the HGP, published data based on a whole genome shotgun strategy [10]. The Celera Genomic data analysis struggled with multicopy sequence presence [145] and was of lower quality than the data from the "academic" section. Both of the released assembled genomes were compared to the dbSNP database to allow the gradual validation of already known SNVs/SNPs and the addition of new ones. In 2002, following the HGP, the international project HapMap began. The International HapMap Project aimed to determine the typical patterns of DNA sequence variation in the human genome. DNA from more than 1,184 individuals from 11 populations [146] was analyzed using array-based SNP genotyping. The data from the HapMap project was used to add variants to the first published reference genome.

UCSC version	**Release date**	**Release name**	**Status**
hs1	Jan. 2022	T2T Consortium CHM13v2.0	Available
hg38	Dec. 2013	Genome Reference Consortium GRCh38	Available
hg19	Feb. 2009	Genome Reference Consortium GRCh37	Available
hg18	Mar. 2006	NCBI Build 36.1	Available
hg17	May 2004	NCBI Build 35	Available
hg16	Jul. 2003	NCBI Build 34	Available
hg15	Apr. 2003	NCBI Build 33	Archived
hg13	Nov. 2002	NCBI Build 31	Archived
hg12	Jun. 2002	NCBI Build 30	Archived
hg11	Apr. 2002	NCBI Build 29	Archived (data only)
hg10	Dec. 2001	NCBI Build 28	Archived (data only)
hg8	Aug. 2001	UCSC-assembled	Archived (data only)
hg7	Apr. 2001	UCSC-assembled	Archived (data only)
hg6	Dec. 2000	UCSC-assembled	Archived (data only)
hg5	Oct. 2000	UCSC-assembled	Archived (data only)
hg4	Sep. 2000	UCSC-assembled	Archived (data only)
hg3	Jul. 2000	UCSC-assembled	Archived (data only)
hg2	Jun. 2000	UCSC-assembled	Archived (data only)
hg1	May 2000	UCSC-assembled	Archived (data only)
source: https://genome.ucsc.edu/FAQ/FAQreleases.html#release1			
Table 2: List of human genome releases			

As a consequence of the free access to all existing DNA sequence databases, the first human reference genomes (hg1 - hg8) were released by the University of California, Santa Cruz (UCSC) between May 2000 and August 2001. In December 2001, the NCBI released NCBI Build 28 (hg10)

and subsequently the NCBI verified all reference human genomes released up to NCBI Build 36.1 (hg18). In 2007, the Genome Reference Consortium (GRC) was formed as an international group of institutions that aimed to improve the existing human, mouse, and zebrafish reference genome assemblies. The starting point for the human genome assembly under GCR was NCBI Build 36.1. It took two years to prepare GRCh37 (hg19), released in 2009. Thanks to significant progress in bioinformatic analysis and the massive amount of data from the 1000 Genome Project, in 2013, the most important updates to the reference human genome were made under the name of GRCh38 (hg38). The three-phase 1000 Genome Project (launched in 2008 and completed in 2012 under the leadership of the GRC) collected multiple data sources (a combination of low-coverage whole-genome sequencing, deep-coverage exome sequencing, and array-based SNP genotyping) to describe genetic variability (deletions, insertions, inversions, copy number variations, and single nucleotide polymorphisms). During the 1000 Genomes Project, the genomes of 2,504 individuals from twenty-six populations were reconstructed. This data was analyzed and finally a total of nearly 88 million variants (84.7 million SNVs/SNPs, 3.6 million short insertions/deletions (indels), and 60,000 structural variants) were identified and incorporated into the GRCh38 reference genome [147]. Despite this progress in human genome assembly, about 8% of its content was still missing. In January 2022, a complete human reference genome was released by the Telomere to Telomere (T2T) consortium. The T2T human reference genome, not only used third-generation sequencing technologies (PacBio and Oxford Nanopore) to add previously non-sequenced regions, but for the first time it is not a compilation of many human genomes (pangenome), but a well-defined human cell line (CHM13hTERT cell; 46,XX). Twenty years later, despite the initial efforts to create a human reference, the human reference genome is finally 100 % sequenced [148].

1.7 Bioinformatics – Sequencing Data Analysis

There has always been the simultaneous development of bioinformatics tools alongside NGS technology. Bioinformatics is defined as "the application of tools of computation and analysis to the capture and interpretation of biological data" [149]. Bioinformatics uses computer science, programming, information technology, mathematics and statistics to analyze and interpret enormous amounts of complex biological data, mostly DNA, RNA, and protein sequences. The basic applications of bioinformatics are sequence assembly, variant calling, genome annotation (describing regions of biological interest within a genome, such as, to determine the protein-coding regions, regulatory sequences, non-coding RNA, structural motifs, or repetitive sequences) and gene expression analysis [150]. Although each of these approaches uses different benchmarking tools,

they all require clean input data. In nucleic acid analysis, there are primary, secondary, and tertiary bioinformatic analyses, which have well-defined procedures and quality checkpoints [151].

There are different ideas of what primary bioinformatic analysis entails. In the narrowest sense, primary analysis only includes image analysis and base calling on the sequencer itself, sometimes the FASTQ file generation and demultiplexing, or even FASTQ quality control, is involved, such as MiSeq or NovaSeq X both from Illumina. The FASTQ format associates each nucleotide with an ASCII-encoded quality number that corresponds to a Phred quality score – Q(phred); see Table 3. Illumina instruments use Q(phred) scores for a base of $Q(\text{phred}) = -10 \log(e)$, where e is the estimated probability of a base being wrong. A Phred quality score of 30 indicates the likelihood of finding one incorrect base call among 1000 bases [150]. Low quality base calls are trimmed, and uninterpretable reads are filtered out, but this step can be done manually by the sequencer user. The most widely used FastQC tool [151] is then used to check the quality of the raw Fastq reads. Index-based demultiplexing and adapter removal are performed in parallel with FASTQ generation. The secondary analysis, always performed by the bioinformatician, refers to *de-novo* assembly (assembly and arrangement of sequenced reads and, in longer sequences, based on the overlap-layout strategy) or for much less demanding reads mapping assembly (reads are assembled according to their concordance with the reference sequence using the mutual read overlaps). If the reference genome is known, the preferred method of assembly is reference alignment. Although mapping assembly benefits from the advantage of the reference genome, the algorithm must determine the correct location of every read against the reference genome, including challenging reads such as those from homologous or highly repetitive regions, therefore, the definition of a threshold to distinguish between real genetic variations and misalignments is crucial. The output of secondary analysis is typically in the standardized binary alignment and map (BAM) and variant call formats (VCF), where BAM is basically the alignment of the sequencing reads to the reference sequence and VCF contains information about the detected variants [152]. VCF is a text file that contains information about the chromosomal position of the reference base and the alternative base, or bases identified from the sequenced data. Secondary analysis also can detect more complex structural variants (copy number variations or large genome rearrangements). For this approach long-read sequencing is more suitable than short reads and WGS is more appropriate than WES [153]. Tertiary analysis is the final

Phred quality score Q(phred)	Propabilty of incorect base call	Base call accuracy (%)
10	1/10	90
20	1/100	99
30	1/1000	99.9
40	1/10000	99.99
50	1/100000	99.999

Table 3: Phred Quality Score

stage of next-generation sequencing data analysis. It plays a critical role in the integration of the results from the primary and secondary analyses with other relevant data sources and is the most time-consuming part of the data analysis. The first step is variant annotation, which provides a biological context for all the variants found. The annotated variants are then filtered according to their clinical significance and reported [154]. As this is the most time-consuming area of the NGS data analysis, artificial intelligence (AI) assisted tools are being introduced into NGS-based diagnosis. AI algorithms, known as machine learning and deep learning, can be used to process large and complex genomic datasets. Although AI is the only way to speed up NGS data interpretation, it will always be up to the clinician to report the final diagnosis. [155] [156]. See Figure 2.

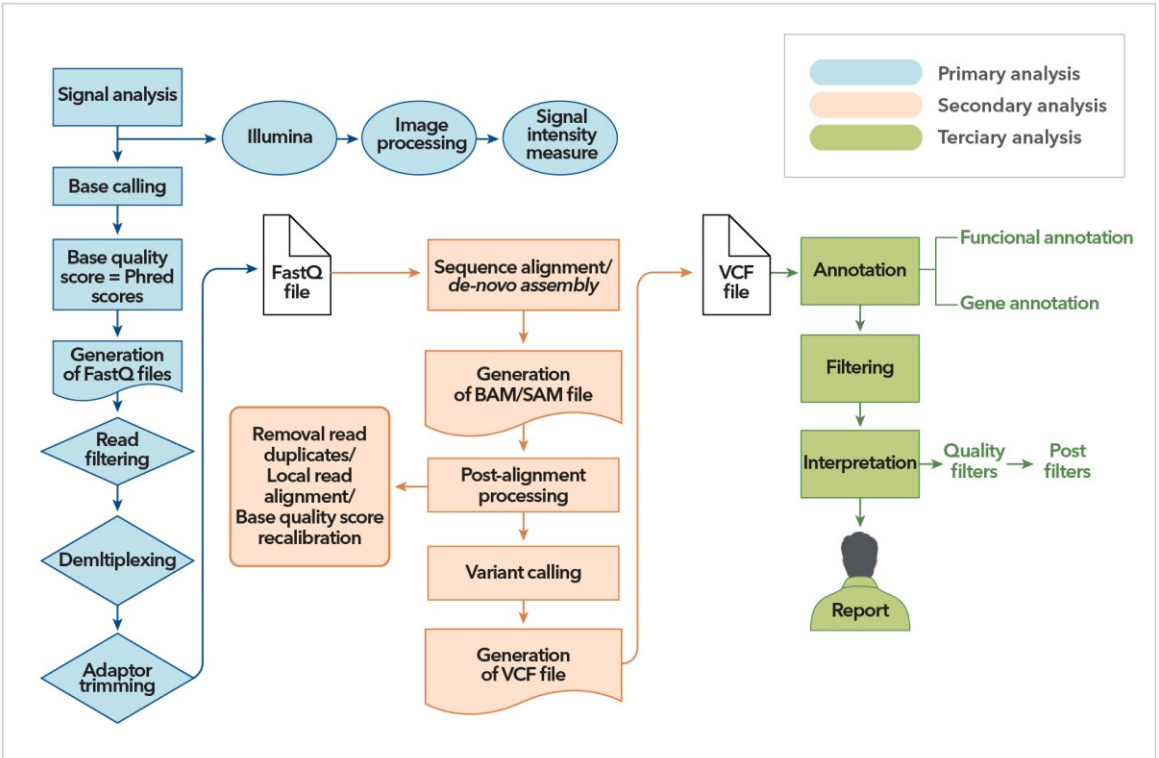

Figure 2: General bioinformatic pipeline source: DOI: 10.3390/jcm9010132, modified

There are widely accepted guidelines for the evaluation of genomic variations obtained through NGS, such as the American College of Medical Genetics and Genomics guidelines [157] and the European Society of Human Genetics [158] guidelines. Standards and guidelines were also published by the Association for Molecular Pathology, the American Society of Clinical Oncology, the College of American Pathologists and the European Society for Medical Oncology [159] for the interpretation and reporting of variants identified with NGS [160].

In fact, there are now so many bioinformatic tools available that it is difficult to choose a specific one. As a result, it can be difficult for different teams to compare their findings and replicate the results of other research groups. Switching from one set of bioinformatic tools to an updated set often requires the re-analysis of data, while switching to brand new bioinformatic tools requires additional training and testing [153]. In the future, more emphasis may be focused on the unification of bioinformatic tools to simplify data flow and interlaboratory reproducibility.

1.8 NGS in Diagnostics

Throughout history, medical diagnosis has evolved from a symptom-based approach to a broader understanding, reflecting growth in medical knowledge and technology that lately includes genetics and personalized medicine. The sequencing of the human genome has revolutionized diagnosis, enabling a deeper understanding of the relationship between genetics, individuality and disease, leading to personalized medical interventions and lifestyle recommendations.

1.8.1 NGS in Clinical Microbiology

Sequencing technology was first applied to public health in 1990 for multi-locus sequence typing of *Neisseria meningitidis*. The authors used amplicon sequencing with primers from multiple housekeeping genes to identify strains with virulent lineages. The authors of the publication correctly assumed that the introduction of sequencing technology to pathogen analysis would simplify pathogen detection and cross-analysis between laboratories by sharing DNA isolates, PCR products, or sequencing data rather than the infectious samples themselves [161]. The biggest advantage of NGS technology is that it can be widely applied the detection and analysis of viruses, bacteria, fungi, parasites, as well as animal and human hosts without prior knowledge of the specific pathogen [162]. However, the entry of NGS technology into pathogen analysis was not straightforward [163], and it had to gain respect alongside traditional methods [164].

Traditional retrospective diagnostic techniques used in microbiological laboratories include culture cultivation, immunoassays (ELISA) [165], pathogen-specific antibody biomarker detection [166], or the molecular identification of microbial DNA or RNA via PCR-based assays [164]. The limitation of most molecular PCR-based assays is that they only target a limited number of pathogens using specific primers or probes [167]. There are approaches where NGS offers advantages over traditional methods. This includes metagenomic studies for more accurate detection and characterization of pathogens, detection of new virus mutations in screening for vaccine escape or detection of antimicrobial resistance. The first NGS metagenomic approach used characterized all the DNA or RNA present in a sample, the entire microbiome, as well as the human host genome in

patient samples [167] and can fundamentally help to improve public health surveillance [168] through the identification or discovery of unexpected or unknown pathogens. Reverse vaccinology uses NGS for vaccine design. After viral genome sequencing special bioinformatic tools are used to predict the most immunogenic epitopes [169]. New mutations that allow partial viral or complete viral vaccination escape can then be detected through regular re-sequencing [170]. The final approach where NGS is used in pathogen detection and identification is antimicrobial resistance (AMR) monitoring. In hospitals in particular, the spread of AMR may lead to a threat to patients. AMR genes, often found on small plasmids, can easily move between bacteria through horizontal gene transfer, leading to the spread of AMR [171]. AMR monitoring can be carried out through whole-genome sequencing as well as targeted plasmid sequencing [172]. Choosing the right sequencing platform is crucial in pathogen analysis. Choosing between short read sequencing and long read sequencing depends on the application. Short read NGS platforms are ideal for high-throughput analysis. Long read technology, represented by ONT, is the perfect choice to provide immediate results [171].

In virology, NGS demonstrated its potential during the SARS-Cov-2 pandemic. In December 2019, the People's Republic of China reported the spread of an unknown virus that caused severe respiratory disease. The virus, later named Severe Acute Respiratory Syndrome Coronavirus-2 (SARS-CoV-2), was identified in the same country on January 7, 2020. Within three months, SARS-CoV-2 had been detected in 144 countries worldwide [173]. The SARS-CoV-2 virus is a single-strand, positive-sense RNA virus with a large RNA genome, approximately 30 kilobases in length, encoding approximately 29 proteins [174]. RNA viruses have a high mutation rate – "offspring" typically differ by 1-2 mutations from their "parent". The ability to rapidly change their genome is essential to allow them to escape the immune response of the host [175]. While other methods, such as RT-qPCR, are widely used for virus screening, the study of the SARS-CoV-2 virus using next-generation sequencing was beneficial for several reasons. Firstly, NGS allows a rapid and detailed analysis of its genetic makeup, which is essential to track the spread of the existing virus types and subtypes and to identify new ones [176] while that information is critical for the development of effective vaccines and therapeutic strategies [170]. Secondly, NGS allows the spread and transmission of the virus to be monitored, which has a significant contribution to public health strategies. Finally, NGS, together with human WGS, provides insight into how the virus interacts with the human genome, improving our understanding of the variability of individual clinical outcomes.

NGS has changed the way pathogens are studied. It enables their rapid identification, helps in the investigation of pathogen outbreaks, and helps to find the new types and subtypes of those pathogens. Making NGS a regular part of research and public health labs is still a challenge, but continued implementation of NGS improvements are leading to advances in the study of infectious diseases.

1.8.2 NGS in Human Genetic Diagnostics

There are three basic areas of human diagnostics where NGS has a big impact: the analysis of inherited genetic variants, oncology diagnosis and non-invasive prenatal testing.

Next to karyotyping and fluorescence in situ hybridization (FISH), array-based comparative genomic hybridization and array-based SNP genotyping are used in human genetics [177]. Generally, the most significant advantage of NGS is the examination of many targets at the same time, as can also be done with array-based SNP genotyping. However, genotyping arrays are recommended for screening rather than for diagnostics [178] as array-based analysis may miss the rare germline coding variants with minor allele frequencies that can be detected by WES [179]. Still WES ignores the majority of the non-coding regions of the genome that are crucial in complete genetic analysis. In the case of rare and ultra-rare genetic diseases, the WGS approach gives the most comprehensive results. WGS in combination with the latest NGS technology and interpretation tools, powered by artificial intelligence, can deliver results in an incredibly short time. This approach is called rapid WGS. The biggest clinical value of rapid WGS is delivered to infants with unknown conditions, where confident diagnosis can be reported in 40–50% of cases [180]. Balancing the clinical and financial value of WGS approaches can be achieved through a combination of whole exome sequencing (coverage 100x) with low pass whole genome sequencing (coverage 2-5x), allowing the discovery of rare coding variants and analysis of variants in the rest of the genome [179]. Part of genetic testing is the analysis of genes responsible for drug efficacy or toxicity – pharmacogenetics. Whereas thousands of pharmacogenetic biomarkers are known, only very few of them are clinically implemented. Partly because the precise determination of the effect of a given variant on drug kinetics is not only influenced by genetic variability in the specific genes, but unlike most Mendelian diseases, drug effect is a composite of genetic as well as clinical (comorbidity) and environmental (drug interactions) factors. Clinically relevant pharmacogenetic SNP markers are mostly analyzed using SNP-based arrays. However, as WES or WGS approaches are increasingly implemented and validated in genetic testing, it can be assumed that pharmacogenetic testing will gradually switch from SNP-based arrays to NGS tests [181].

The clinical value of oncology therapies that target genetic alterations was first demonstrated through imatinib treatment of chronic myeloid leukemia patients with BCR-ABL gene fusion [182] in 2008. In the diagnosis of cancer patients, multiple gene mutations often need to be tested due to the heterogeneity of tumor mutations. Traditional molecular pathology assays are used to target a specific mutation or biomarker, so it might be necessary to perform multiple tests. This kind of testing requires more biopsy tissue, whereas, with NGS technology, these targets can be analyzed in a single assay which minimizes tissue input requirements and the turnover time [183]. A specially designed RNA-seq is used to detect fusion genes, which are responsible for the development of cancer and can be targeted therapeutically. In oncology, NGS technology is an important tool in precision medicine, not only providing information to allow the diagnostic classification of diseases, but also to assist in the selection of the correct therapeutic guidelines and predicted prognosis.

A modern diagnostic approach within prenatal care, introduced for commercial use in 2011, is NGS-based non-invasive prenatal testing (NIPT) that analyzes cell-free fetal DNA in maternal blood to detect fetal chromosomal aneuploidies and large structural aberrations. Although NIPT is used as a screening method and a positive result must be confirmed by invasive tests, NIPT has high sensitivity (true positive rate) and high specificity (false negative rate), ranging from 91% to 100% depending on aneuploidy [184]. Deep ultrasound and cytogenetic testing after a positive NIPT can refine the results to avoid invasive checks, especially in low-risk women who are more likely to have true negative NIPT results [185].

1.8.3 Methodical Challenges in NGS

In 2021, the American College of Medical Genetics and Genomics (ACMG), a highly respected authority, released comprehensive recommended technical standards for genetics laboratories to ensure the consistent implementation of NGS analyses. This chapter summarizes the ACMG recommendations for NGS implementation and validation [186].

After the choice of an appropriate sequencing approach, emphasis is put on end-to-end workflow analytical validation, from DNA isolation, through NGS library preparation to data analysis. It is necessary to critically examine any NGS assay with respect to its technical limitations, whether they relate to limitations due to the nature of the samples or the genomic regions where the variants are detected. For this it is necessary to include well-defined reference materials to determine the technical accuracy of the NGS assay. The Genome in a Bottle Consortium, supported by the (NIST) National Institute for Standards and Technology, released well-defined standardized variant datasets (truth data set) for multiple Coriell samples (gDNA isolates from cell lines), which can be

used for germline end-to-end technical validation. ACMG emphasizes that the truth data set may contain false positive variants because Coriel samples are isolates from passaged cell lines that may contain mutations not captured in the "truth data set" and this must be taken into account in bioinformatic validation. It is not recommended to validate an entire workflow using different platforms and pipelines that may introduce variability into the resulting data. (for example, reference material should not be analyzed using different pipelines).

There are special requirements for the samples. First, it must be possible to isolate nucleic acid of sufficient quality and quantity as defined in the technical standards. Second, some sequencing approaches are not recommended for certain sample types, for example, FFPE for long-read sequencing or saliva samples for WGS. While FFPE samples objectively have low NA quality, saliva samples are excluded from WGS analysis due to the possibility of non-human contamination that may affect data analysis. The source of the extracted genomic DNA must always be declared (blood, saliva, FFPE samples...) so that the possible impact on sequencing data can be determined. Similarly, the minimum required coverage must be determined for each sequencing approach to correctly determine zygosity, mosaicism or heteroplasmy in the case of germline testing. In somatic testing high sequencing coverage (also called deep sequencing) is necessary to allow the detection of rare variants.

Last but not least, it is necessary to build up a reproducible bioinformatics pipeline. A separate chapter on the implementation of NGS in clinical laboratories is devoted to the correct reporting of variants, which, although it ends with the final report, begins with the correct filtering of variants, and the selection of databases for variant classification [186]. The European Society of Human Genetics has issued similar guidelines for genetic testing using NGS [187]. The NGS implementation guidelines are also intended for pathology laboratories in the standards in both Europe [188] and the United States [189].

As genetic and genomic testing using next-generation sequencing becomes commonplace in the life sciences, clinical genomics, and so-called “recreational genomics,” it is necessary to ensure the genomic data generated is of both a high quality and is comparable. The impact of the biological source of the sample on genomic analysis is also widely debated, as studies with conflicting conclusions have been published. Although saliva-derived gDNA has been proven as an alternative to blood-derived gDNA, especially for array-based genotyping approaches [190], including array-based methylation studies [191], and some studies even proved saliva to be suitable for WES or even WGS to analyze SNVs and small-indels [192] [193] there are also studies and recommendations

from authorities that saliva should not be used for WGS analysis due to the negative influence of non-human contamination [157]. We designed a study to systematically compare the sequencing data and qualitative parameters of saliva-derived gDNA with blood-derived gDNA, for both the WES and WGS protocols for SNVs and small-indels.

2 AIMS

The aim of the thesis was to evaluate the accuracy of NGS protocols for the generation of reliable, high-quality genomic data of whole exome and whole genome analysis in a routine laboratory setting. We analyzed the importance of different sample sources and sequencing protocols to determine the effect on the accuracy of NGS analysis for different sequencing approaches. The aims of both studies are presented in this chapter. The individual published studies, which provide a more detailed description, are part of Chapter 9, Appendix.

Aim 1 - Performance evaluation of different approaches to library preparation (capture- and amplicon-based) for a comprehensive analysis of the SARS-CoV-2 genome in a high-throughput laboratory. The individual sub aims were:

1. Evaluation of the suitability of different sequencing approaches utilizing Ct value.
2. Workflow validation by engaging synthetic controls that correspond to different variants of the analyzed genome.
3. Analyze the advantages and limitations of the end-to-end protocol in terms of the time-consumed.

Aim 2 - To test the suitability of saliva-derived genomic DNA for genomic testing in comparison with blood-derived gDNA.

The individual sub aims were:

1. Technical validation of WGS and WES protocols using reference standards.
2. Variant concordance comparison of paired blood/saliva samples.
3. Cross-protocol genotype concordance comparison.
4. Comparison of the sequencing metrics of gDNA derived from blood and saliva.
5. Estimation of the non-human contamination rate in saliva samples and its influence on variant detection.

3 EXPERIMENTS

The experimental work carried out in both studies are presented in this chapter. The publication of the individual studies, which provide more detailed descriptions of the experiments can be found in Chapter 9, Appendix.

3.1 Sample collection

For the first project we collected a total number of 55 isolates from nasopharyngeal swabs and bronchoalveolar lavage fluid from patients with a positive COVID-19 diagnosis from five Czech hospitals – two in Prague, two in Brno, and one in Pilsen. Nasopharyngeal swab samples and bronchoalveolar lavage fluid samples were collected and supervised by hospital staff.

For the second project, we collected paired blood-saliva samples from ten participants with no clinical indications of disease; see Table 4. All participants from the blood-saliva comparison project provided informed consent prior to sample collection.

Sample ID	Age*	Sex	Relationship
Proband 1	16	F	daughter
Proband 2	29	M	son
Proband 3	47	M	mother
Proband 4	56	F	father
Proband 5	28	M	unrelated
Proband 6	34	M	unrelated
Proband 7	34	M	unrelated
Proband 8	36	F	daughter
Proband 9	60	F	mother
Proband 10	62	F	unrelated
* age of proband at time of sample collection			
4 person family			
a couple			
Table 4: Aim 2, Sample identification			

4 mL of venous blood was extracted into Vacuette K2EDTA tubes (Becton Dickinson, USA, cat. number: SKU: 367863 GTIN:00382903678631). After 30 minutes of tempering at room temperature, all tubes were stored at 4°C. Before the saliva sample collection, the participants were given instructions and under supervision cleaned their oral cavities (to minimize non-human sources of contamination) following a strict protocol. Two 1 mL samples of saliva were collected from each participant, placed into an in-house collection buffer and stored at 4°C.

3.2 NA extraction and sample quality control

For the first project, the total nucleic acid from the nasopharyngeal swabs and bronchoalveolar lavage fluid was isolated using the QIASymphony Virus/Pathogen Mini kit (QIAGEN, USA, cat. number: 937036) or the QIAamp Viral RNA Mini kit K (QIAGEN, USA, cat. number: 52906) in the hospital laboratories. Subsequently, we determined the number of SARS-CoV-2 copies in each isolate using RT-qPCR (Direct SARS-CoV-2 RT PCR kit, Institute of Applied Biotechnologies, Czech Republic, cat. number: DSC960DELTA).

For the second project, blood and saliva samples were processed using the QIAamp® DNA Blood Mini Kit (QIAGEN, USA, cat. number: 51104) within 24 hours of sample collection, to extract gDNA. Blood gDNA isolation was carried out using the "DNA Purification from Blood or Body Fluids" protocol. We used a modified version of QIAGEN's "Isolation of genomic DNA from saliva and mouthwash" protocol to isolate the gDNA from saliva samples. We checked the integrity of the blood-derived and saliva-derived gDNA using gel electrophoresis (0.8% agarose gel) and measured the concentrations of gDNA using a 1x dsDNA High Sensitivity (HS) Kit on the Qubit Instrument (Thermo Fisher Scientific, USA, cat. number: Q33231). Finally, we made an evaluation of the purity (A260/280 absorbance ratio) of all the gDNA samples using a NanoPhotometer P300 (Implen, Germany).

3.3 Internal controls

We used control samples to evaluate the protocol accuracy for both projects.

In the first project we used 5ng of human breast tumor RNA (HBT) (Takara Bio, France, cat. number: 636635), as a negative control (NC). Positive controls (PC) 1-4 were prepared by spiking two synthetic RNA controls (Twist Synthetic SARS-CoV-2 RNA Control 1, MT007544.1, cat. number: 102019 and Twist Synthetic SARS-CoV-2 RNA Control 2, MN908947.3, cat. number: 102024, Twist Bioscience, USA) into standard human breast tumor (HBT) total RNA (Takara Bio, France). Both Twist Synthetic controls 1 and 2 were diluted by ten-fold dilution with 5 ng of HBT total RNA so that the final concentration of each of them was 1.000, 100, 10 and 1 copies in positive controls PC1, PC2, PC3 and PC4, respectively. Negative and positive controls were validated by RT-qPCR by measuring their Ct values.

For the second project, as an internal control for technical validation, we used the NIST human genome RS NA12878 sample (Coriell Institute, USA, cat. number: NA12878) in the concentration required by the particular NGS library preparation protocol.

3.4 NGS library preparation

In the first project, in total 127 NGS libraries, including 17 internal controls, were prepared; see Table 5. For detailed information on sample multiplexing see Klempt et al., Supplementary Table S1. Over the course of the second project, a total of 52 NGS libraries were prepared, including 12 controls; see Kvapilova et al., Additional File 1, Table 1 and Table 2.

Sequencing approach	NEB+TWIST		Illumina	Paragon
	NEB+TWIST1	NEB+TWIST2		
num. of samples	24	16	35	35
controls	4x pc/2x nc	1x pc	4x pc/1 nc	4x pc/ 1 nc
num. of prepared NGS libraries	30	17	40	40
num. of plexes	3	2	8	1 pool
num. of sequenced NGS libraries	30	17	39	21
num. of analysed NGS libraries	18	16	32	21
sample Ct values	11.29 - 31.96	13.1 - 25.83	11.29 - 29.98	11.29 - 25.83
Table 5: Aim 1, Sequencing approaches				

Virus Whole Genome NGS library preparation by Hybridization Capture (Twist Bioscience)

The NGS libraries from the 40 isolates, 4 PC, and 2 NC were prepared using the NEBNext Ultra II Directional RNA Library Prep Kit for Illumina (New England Biolabs, USA, cat. number: E7760). Then we prepared 5 multiplexes: three multiplexes were only based on DNA concentration (150 ng or maximum amount) regardless of the Ct value and two multiplexes based on DNA concentration (150 ng or maximum amount) and Ct values. All the multiplexes were subsequently enriched by the Twist SARS-CoV-2 Research Panel kit (Twist Bioscience, USA, cat. number: 102018). Prior to sequencing, the prepared multiplexes were checked using a Qubit 2.0 and 2100 Bioanalyzer and equally pooled according to their concentration.

Virus Whole Genome NGS Library preparation by Hybridization Capture (Illumina)

Prior to library preparation, RNA from all 35 isolates was transcribed into double stranded copy DNA using the NEBNext® RNA First Strand Synthesis Module (New England Biolabs, USA, cat. number: E7525) and the NEBNext® Ultra™ II Directional RNA Second Strand Synthesis Module (New England Biolabs, USA, cat. number: E7550). Next, the NGS libraries were prepared using the Nextera Flex for Enrichment kit (Illumina, USA, cat. number: 20025524, now Illumina DNA Library Prep with Enrichment) kit and divided into eight multiplexes according to Ct value. Then, each multiplex was enriched using the Respiratory Virus Oligo Panel (Illumina, USA, cat. number:

20042472). Prior to sequencing, the multiplexes were checked using the Qubit 2.0 and 2100 Bioanalyzer to ensure equal pooling.

Virus Whole Genome NGS Library preparation by Amplicon (Paragon Genomics)

We used the same 35 isolates in the Illumina approach, but only 21 NGS libraries were successfully prepared using the CleanPlex® SARS-CoV-2 Research and Surveillance Panel (Paragon Genomics, USA, cat. number: SKU: 918010). Those 21 NGS libraries were sequenced in a single pool.

Human Whole Genome NGS library preparation (Illumina)

In the second project, we mechanically fragmented the paired blood-saliva gDNA using the Covaris M220 (Covaris, USA) and then prepared whole genome NGS libraries using the TruSeq DNA PCR-Free kit (Illumina, USA, cat. number: 20015963) according to the manufacturer's guidelines. We measured the concentration of the NGS libraries using the Qubit 1x dsDNA HS Kit (Thermo Fisher Scientific, USA), and then confirmed the fragment length of the NGS libraries using the 2100 Bioanalyzer System with a DNA HS chip (Agilent, USA). The NGS library quality control (QC) parameters can be found in Kvapilova et al., Supplementary Additional File 1, Table 1.

Human Whole Exome NGS library preparation (Twist Bioscience)

We prepared whole exome NGS libraries using the Illumina DNA Prep with the Enrichment library preparation kit (Illumina, USA, cat. number: 20025524), followed by the Alliance VCGS Exome panel (Twist Bioscience, USA, cat. number: 104912) and Mitochondrial DNA panel (Twist Bioscience, USA, cat. number: 102039) hybridization capture enrichment. Finally, we measured the concentrations and fragment length of the whole exome NGS libraries in the same way as the whole genome NGS libraries. The NGS library QC parameters can be found in Kvapilova et al., Supplementary Additional File 1, Table 1.

3.5 Sequencing

The NGS libraries from both projects were pooled after their quality control assessment and these NGS libraries were denatured and diluted as per the guidelines corresponding to each of the Illumina sequencing platforms used:

MiSeq

The NGS libraries, prepared through the NEB+TWIST1 and NEB+TWIST2 approaches were sequenced on the MiSeq platform (Illumina, USA) using the MiSeq Reagent Kit v3 (600 cycle) (Illumina, USA, cat. number: MS-102-3003) and the Reagent Kit v2 (500 cycle) (Illumina, USA,

cat. number: MS-102-2003). The NGS libraries prepared using the NEB+Illumina approach were sequenced using the MiSeq Reagent Kit v3 (600 cycles) (Illumina, USA, cat. number: MS-102-3003), and the Paragon NGS library. The loading concentration and sequencing configuration can be found in Klemt et al., Supplementary Table S2. For all libraries we targeted 0,5 mil PE reads per sample. The number of total PE reads can be found in Klempt et al., Supplementary Table S2.

iSeq

Prior to NovaSeq 6000 sequencing, the human whole genome NGS libraries from the second project, were pre-sequenced on an iSeq instrument (i1 Reagent v2, 300 cycles sequencing kit; Illumina, USA, cat. number: 20031374) in a single pool, to check the non-human contamination ratio; see Kvapilova et al., Supplementary Additional File 1, Table 1. We targeted 0.5 mil PE reads per sample. The number of total PE reads are listed in Kvapilova et al., Supplementary Additional File 1, Table 2.

NovaSeq 6000

The whole human exome and genome NGS libraries from the second project were sequenced using the S4 Reagent Kit v1.5 (300 cycles) (Illumina, USA, cat. number: 20028312) using the XP4 workflow (Illumina, USA, cat. number: 20043131). The results from the pre-sequencing iSeq run were used to adjust for proper pooling of the WGS libraries to ensure delivery of an average output of 800 million pair-end reads per sample (30x average coverage). With the WGS libraries we targeted 60,000 pair-end human reads (100x average coverage). The number of total PE reads are in Kvapilova et al., Supplementary Additional File 1, Table 2.

3.6 Bioinformatic analysis

Primary analysis

Both projects utilized the same primary data analysis workflow, the bcl files were demultiplexed and converted into fastq files using bcl2fastq v2.20.0.422 [194] with its default settings, followed by quality control using FastQC. v0.11.8 [151]. Adapter trimming and removal of sequencing artefacts (short reads, low quality reads) from the datasets was done using fastp tool v0.20.0. [195][192].

Secondary analysis

In the first project, all the sample sequencing reads were aligned to the SARS-Cov-2 Wuhan-Hu-1 reference genome NM9088947.3 using bowtie2 [196] with the default sensitivity settings. To mark duplicate sequences and measure the sequence depth we used the Picard tool [197]. Variants were identified using the Freebayes method with default settings [198]. Finally, we calculated the median

coverage for all the coding parts of the SARS-CoV-2 genome regions (ORF1a, ORF1ab, S, ORF3a, E, M, ORF6, ORF7a, ORF7b, ORF8, N, ORF1) using the Picard tool [197].

In the second project, a secondary analysis was performed as follows: we mapped the sequencing reads onto the Illumina DRAGEN Graph reference genome GRCh38. Then we used the DRAGEN Bio-IT Platform DNA ENRICHMENT pipeline v3.10 (Illumina, USA) for WES data analysis and the DRAGEN Bio-IT Platform DNA GERMLINE pipeline v3.10 (Illumina, USA) for the WGS data analysis [199]. All mapping reads with a quality level >1 were used for variant calling with a hard threshold of AF >5%. We used the vcfstats tool output and DRAGEN metrics for mapping and variant calling quality control [200]. Targeted regions for WES analysis were defined by the Twist panels (Alliance Exome + Mitochondial, BED files available online). For the analysis of the non-human DNA ratio, we used the FastqScreen tool [151] to map unmapped reads against an oral human microbiome database (HOMD) [201] and the bcftools isec toll [202] for variant comparison against the ACMG SF v3.2 list [203] for reporting of secondary findings in whole exome and whole genome sequencing.

For the tertiary analysis of paired blood-saliva called variants the EMEDGENE pipeline v32.0.25 was used.

Concordance comparison formula

For concordance comparison, we used the F1 score, which is the harmonic mean of precision (truth positive predictions relative to total predicted positives, in other words, correctly identified positives) and recall (truth positive predictions relative to total actual positives, in other words incorrectly identified as positive). The F1 score has a maximum value of 1 (perfect precision and recall) and a minimum of 0. In our case, we measured the concordance between two numerically identical variants/genotypes datasets [204]. The generated VCF files were compared by hyp.py tool v.0.3.14 [205].

Statistical methods

Statistical computation was only done as part of the second project. We analyzed the pairwise concordance between various experimental conditions using a mixed analysis of variance (ANOVA). We used post-hoc tests to calculate the statistical significance of individual conditions and we corrected the p-values obtained through Bonferroni adjustments [206]. Statistical significance was set to $p \geq 0.05$. We used the non-parametric Mann–Whitney test [207] to compare the sequencing metrics between the groups of blood and saliva samples. All the statistical analyses

and visualizations were performed using Python software packages.

4 RESULTS

The results of both projects are presented in this chapter. The individual studies, which provide more detailed descriptions of the experiments may be found in Chapter 9, Appendix.

Aim 1

Sub aim 1: Evaluation of the suitability and bottlenecks of different sequencing approaches utilizing Ct values

Initially, thirty NGS libraries were prepared using the NEBNext® Ultra™ II Directional RNA Library Prep Kit for the Illumina and Twist SARS-CoV-2 Research Panel. NGS libraries were grouped into three multiplexes, each containing ten samples. The analysis of the sequenced data revealed significant variability among the samples, within both the entire experiment, and the individual multiplexes. The analysis performance (mean coverage and the number of mapped reads) of the individual samples within each multiplex varied significantly, as shown by the standard deviation (SD) of 4.65, 3.93, and 4.63 million PE reads for multiplexes 1, 2, and 3, respectively; see Klempt et al., Supplementary Table S1. Despite this sample variability, the distribution of total reads across all three multiplexes was relatively even, with 19, 23.6, and 20.1 million PE reads which suggests correct pooling efficiency; see Klempt et al., Figure 1, green lines 1-3 and Supplementary Table S1. These results led us to rethink the process of sample quantitation and quality scoring prior library preparation. We used an in-house reverse transcription quantitative PCR (RT-qPCR) assay rather than relying on cycle threshold (Ct) measurement provided by external laboratories. Further, we pooled the NGS libraries based on these Ct values, so in next experiment (NEB+TWIST2 approach), we grouped 17 NGS libraries into two multiplexes according to our own Ct values. This adjustment resulted in a more uniform coverage distribution among the 17 NGS libraries. The SD within these two multiplexes was notably lower, at 1.05 and 1.78 million PE reads for multiplexes 4 and 5; see Klempt et al., Supplementary Table S1.

The next step was the preparation of 40 NGS libraries (including four positive and one negative control) using Nextera Flex for Enrichment and the Respiratory Virus Oligo Panel (both Illumina) combined with NEB solutions for the first and second-strand cDNA generation (NEB+Illumina). In these libraries, 5 samples were pooled into each multiplex, to provide a total of 8 multiplexes, with the Ct range within each multiplex no greater than 3 (multiplex 8 included PCs with higher Ct values than the rest of the samples). We observed a similar effect on the coverage uniformity as we found in the NEB+TWIST2 approach. We see better coverage uniformity in multiplexes, where the samples were grouped based on Ct values; see Klempt et al., Figure 1, green line 4 and 5 plus orange line 1 - 8. We also observed that the coverage decreases in samples with

lower viral load (high Ct values); see Klempt et al., Figure 1, orange line 7 and 8. The NEB+Illumina approach generally resulted in a lower percentage of mapped reads compared to other approaches, as indicated in the Klempt et al., Supplementary Table S1. A lower percentage of mapped reads could be a consequence of the lower Respiratory Oligo Viral Panel specificity.

The Paragon protocol does not require multiplexing, but it allows the individual quality control of each NGS library. Only 21 (out of the 40 prepared) NGS libraries passed quality control and they were mixed into a single sequencing pool. The Ct values of the sequenced NGS libraries were between 11.29 and 25.83. Although the amplicon-based samples mostly had higher Ct values (Ct≥23), after sequencing six of the NGS libraries exhibited a lower number of reads, but compared to the capture-based approaches, the rest of the NGS libraries had higher coverage; see Klempt et al., Figure 3, therefore uniform coverage over the entire pool was not achieved; see Klempt et al., Figure 1 and Supplementary Table S1.

The Ct values negatively correlated with the percentage of reads mapped to the reference genome and most significantly correlated with the percentage of aligned reads at 20, 50 and 100x coverage. In fact given that the Ct values were an indication of the actual viral load, as illustrated in Klempt et al., Figure 2 and Figure 3, nearly all bioinformatic processing parameters showed this trend. A less pronounced negative correlation was observed between Ct values and median coverage, when the 13 NGS libraries common for all three approaches were investigated; see Klempt et al., Figure 4. This is likely because all the samples fell within the optimal Ct range of 11–23, this is the optimal Ct value to achieve the best ratio between mapped and total reads for a coverage >20x.

Sub aim 2: Workflow validation (wet lab and bioinformatics pipeline) by engaging synthetic controls corresponding to different variants of the analyzed genome

As shown in Klempt at al., Table 1, we validated the wet lab workflow and reference-mapping-based bioinformatic pipeline using the synthetic Twist controls MT007544.1 and MN9088947.3. utilizing knowledge of three single nucleotide variants present in the control samples (SNPs 19065T > C, 26144G > T, 22303T > G; and deletion 29749 ACGATCGAGTG > A). Each of them compared the detected variants found in all sequencing approaches for both of the positive controls. The sequencing data obtained from the amplicon-based approach only detected two of these variants due to low coverage at certain positions, despite the higher overall number of called variants across all the samples sequenced using this approach. On the contrary, the positive controls prepared using capture-based approaches detected all four variants, moreover all the NGS libraries prepared using these approaches showed high sequencing similarity across all samples; see Klempt et al., Figure 5.

Sub aim 3: Analysis of the advantages and limitations of the end-to-end protocol in terms of the time-consumed

Both approaches (capture- and amplicon-based) showed advantages and disadvantages in terms of the sample input requirements, time consumed and output data quality. The amplicon-based approach had a simpler workflow, which requires both very low hands-on (approx. 6 hours) and instrument time (approx. 4 hours). In our hands however, the amplicon-based approach resulted in a high PCR bias; see Klempt et al., Figure 5. The capture-based approach required two more hours of hands-on work in comparison to the amplicon-based approach but then required 24 hours of instrument time, mostly due to the hybridization procedure, but it then produced better coverage uniformity. In terms of the workflow, both capture-based approaches are more laborious than amplicon-based approach; see Klempt et al., Supplementary Figure S1. Although capture- and amplicon-based approaches also differ in their sample input requirements, it was difficult to objectively evaluate this. Since the concentration of nucleic acids could not be measured in most cases, we used the required input volume regardless of the potential sample concentration.

Aim 2

Sub aim 1&2: Technical validation of WGS and WES protocols using reference standard and variant concordance comparison of paired blood-saliva samples

Using an RS NA12878 sample, we aimed to estimate the error rate in our WGS and WES protocols, individually for SNVs and small-indels. We performed a) a pairwise-triplicate-based comparison and b) a truth dataset (TDS)-based comparison utilizing RS NA12878. Secondly, we did the same analyses for the paired blood–saliva samples. We calculated the variant concordance comparison for both the WGS and WES protocols, using the F1 score:

- restricted to the high confidence region (HCR) of the RS NA12878 (WGS_HCR or WES_HCR);
- restricted to the non-difficult region (NDR) (WGS_NDR or WES_NDR);
- restricted to the intersection of HCR and NDR regions (WGS_HCR_NDR or WES_HCR_NDR);
- restricted to the problematic genomic regions (HARD) (WGS_HARD or WES_HARD);

We used the F1 score based evaluation as the major parameter for paired blood-saliva sample comparison, calculating the absolute concordance between two data sets of the same size (number

of bases). The resulting medians of F1 scores are summarized in Table 6. For the WES protocol, the median value of F1 score for ten paired blood-saliva samples was 0.9858 for SNVs and 0.9076 for small-indels. For the WGS protocol, the median value of F1 was 0.9761 for SNVs and 0.9511 for small-indels. Generally, SNVs have higher accuracy than small-indels in all analyzed genomic regions (with the exception of small-indels detected in WES_HCR_NDR RS NA12878). The highest F1 scores were obtained for those parts of the genome restricted to HCR plus NDR. In contrast, HARD regions showed the lowest concordance. The distribution of F1 scores was more heterogeneous in the WES protocol compared to the WGS protocol with respect to genomic regions, this is more noticeable in small-indels; see Kvapilova et al., Fig. 2 and Supplementary Additional File 1, Table 3. Importantly, the F1 score distribution of paired blood-saliva samples copied the distribution of F1 score RS NA12878 iterations over the whole genome, whole exome, and overall restricted regions.

WES						
Comparison Type	**Variant Type**	**WES**	**WES_HCR**	**WES_HCR_NDR**	**WES_NDR**	**WES_HARD**
WB-S	SNV	0,9858	0,9980	0,9998	0,9993	0,9131
	small-indels	0,9076	0,9607	0,9988	0,9956	0,8140
TDS (RS NA12878)	SNV		0,9931	0,9984		
	small-indels		0,9492	0,9934		
RS NA12878	SNV	0,9832	0,9971	0,9994	0,9987	0,9132
	small-indels	0,8985	0,9549	1,0000	0,9967	0,8166
WGS						
Comparison Type	**Variant Type**	**WGS**	**WGS_HCR**	**WGS_HCR_NDR**	**WGS_NDR**	**WGS_HARD**
WB-S	SNV	0,9761	0,9985	0,9997	0,9992	0,8230
	small-indels	0,9511	0,9956	0,9988	0,9963	0,8984
TDS (RS NA12878)	SNV		0,9982	0,9996		
	small-indels		0,9969	0,9996		
RS NA12878	SNV	0,9752	0,9989	0,9996	0,9990	0,8189
	small-indels	0,9537	0,9972	0,9997	0,9966	0,9072

Table 6: F1-scores medians for WGS and WES analyses individually. Results are grouped, according to the used protocol (WES, WGS), comarison type (WB_S, TDS, PAIRS) and variant type (SNV, SMALL-INDEL): Each group contains F1-Score values for all comparisons in four different genomic regions whole reference genome or exome (WGS or WES), high-confidence region (WGS_HCR or WES_HCR), non-difficult regions (WGS_NDR or WES_NDR), HCR and NDR simultaneously (WGS_HCR-NDR or WES_HCR_NDR) and regions outside NDR intersected with regions outside HCR (WES_HARD and WGS_HARD). Each group is supplemented with median.

Sub aim 3: Cross-protocol genotype concordance comparison

We performed cross-protocol genotype concordance to determine whether the same variants were found, regardless of the whether the WES or WGS protocol was used. For this analysis, we used the pre-defined BED file "file4truth" (BED file of the Twist Alliance VCGS Exome panel, HCR of RS NA12878 and NDR of GRCh38; a total of 27,031,362 bp). We did two comparisons for three iterations of RS NA12878. In the first comparison we analyzed the concordance with TDS. This resulted in a concordance of 99.58 % for SNVs and 98.43 % for small-indels. In the second comparison we analyzed the concordance of all of the RS NA12878 iterations; the concordance was 99.82 % for SNVs and 98.95% for small-indels. The same cross-protocol analysis was done separately for blood and saliva samples. The cross-protocol genotype concordance for the blood samples was 99.89 % (SNVs) and 98.99 % (small-indels); for the saliva samples it was 99.90 % (SNVs) and 99.24 % (small-indels). The results of these cross-protocol genotype concordance comparisons are summarized in Kvapilova et al., Fig. 3 and Additional File 1, Table 4.

To better visualize the results, we plotted a set of variant-related parameters against chromosomal locations to explore discrepancies between the paired blood-saliva samples and the three iterations of RS NA12878 samples. For all the paired blood-saliva sample parameters, we see the same pattern as for the RS NA12878 triplicates; see Kvapilova et al., Fig. 4.

Sub aim 4: Comparison of the sequencing metrics of blood or saliva derived gDNA

We compared the sequencing parameters from saliva-derived DNA samples with those from blood-derived samples for the WGS and WES protocols. We found the following differences: the variant calling quality, in both SNVs and small-indels, was better in saliva, especially when using WGS, but it was only significantly better for small-indels. The number of mapped reads (reads aligned to human reference genome GRCh38, was lower for saliva - this was significant for both the WGS and WES samples but more so for the WGS protocol. The saliva sample also produced more duplicate reads, which was only significant with the WES protocol. The saliva samples had fewer MAPQ10 reads, which was only significant for the WGS protocol. Lastly, the length of the DNA fragments was shorter in the saliva samples, and this was significant for both the WGS and WES protocols. For all the parameter comparisons; see Kvapilova et al., Fig. 5.

Sub aim 5: Estimation of non-human contamination rate in saliva samples and its influence on variant detection

We also focused on a determination of the non-human contamination ratio in saliva samples and its influence on the number of called variants. We used two methods to determine the contamination ratio. First we used low-pass iSeq sequencing to determine the level of contamination to allows us to calculate the NovaSeq 6000 loading concentration necessary to achieve the required number of mapped human reads. Then we recalculated the contamination rates after NovaSeq 6000 sequencing to check the final non-human contamination level. We found that between 8% and 45% of all the WGS NovaSeq 6000 reads did not match the human GRCh38 reference genome. In contrast to the RS NA12878 and blood-derived samples, where only 4 to 5% of all reads did not match; see Kvapilova et al., Fig. 6 and Supplementary Additional File 1, Table 5. The WES library protocol significantly reduced the amount of non-human DNA in both the blood and saliva samples. However, the contamination rates of WGS saliva samples were more heterogenous than those of the WGS blood samples. Nonetheless, when we plotted the F1 scores against the contamination level for each saliva sample, we did not find any pattern that might suggest the contamination levels affected the sequencing accuracy; see Kvapilova et al, Fig. 7 and Supplementary Additional File 1, Table 6.

Beyond the scope of the published study, we performed variant comparison against the ACMG SF v3.2 list to report secondary findings from the paired blood-saliva samples for both sequencing approaches. The average value of the SNV variant calls concordance with the ACMG gene set were 98,73 % for WGS and 99,81 % for WGS, restricted to the exonic regions only, and 99,54 % for WES. Concordance of small-indels calls were on average 92,88 % for WGS and 97,77 % for exonic regions in WGS, and 70,51 % for WES; see Table 7. For the tertiary analysis of paired blood-saliva samples we used the Emedgene software package to find all relevant variants located in the ACMG genes. The variants found using Emedgene were the same for both the blood and saliva samples as well for both the WES and WGS approaches; see Table 8.

WGS SNVs													
samples	1	2	3	4	5	6	7	8	9	10	RS NA12878		
blood unique variants	50	39	54	29	56	61	45	47	41	63	30	27	30
saliva unique variants	62	34	60	73	81	53	49	44	90	32	61	37	51
variant intersect	8408	8831	8125	8582	8069	8474	8073	7850	8140	8293	8125	8128	8135

WES SNVs													
samples	1	2	3	4	5	6	7	8	9	10	RS NA12878		
blood unique variants	0	1	1	0	1	0	0	0	1	0	0	0	0
saliva unique variants	2	2	0	2	0	0	0	2	2	0	0	0	0
variant intersect	341	368	326	305	285	320	321	255	268	338	287	287	287

WGS SNVs restricted to VCGS exome													
samples	1	2	3	4	5	6	7	8	9	10	RS NA12878		
blood unique variants	0	0	0	0	2	0	0	0	0	1	1	0	0
saliva unique variants	1	0	0	0	0	1	1	0	0	0	0	1	0
variant intersect	343	370	327	308	286	321	321	258	272	338	289	289	289

WGS small-indels													
samples	1	2	3	4	5	6	7	8	9	10	RS NA128787		
blood unique variants	71	68	72	64	108	68	71	80	81	72	68	67	63
saliva unique variants	102	85	91	88	106	89	82	83	85	84	76	50	88
variant intersect	2172	2276	2157	2299	2105	2228	2067	2038	2080	2150	2166	2167	2154

WES small-indels													
samples	1	2	3	4	5	6	7	8	9	10	RS NA12878		
blood unique variants	4	4	4	9	6	6	7	7	4	7	7	5	4
saliva unique variants	3	4	4	7	6	4	5	6	5	4	5	9	6
variant intersect	25	27	25	25	22	23	26	25	27	25	19	19	22

WGS small-indels restricted to VCGS exome													
samples	1	2	3	4	5	6	7	8	9	10	RS NA12878		
blood unique variants	0	0	0	0	2	0	1	0	0	0	1	0	0
saliva unique variants	0	0	1	0	1	0	2	0	0	0	0	1	0
variant intersect	28	27	26	28	27	28	32	33	29	29	23	23	23

Table 7: Variant concordance of paired blood-saliva samples in ACMG SF v3.2 list of genes. Comparison of the variant concordance in the set of genes defined by the American College of Medical Genetics and Genomics in paired blood-saliva samples. Unique variants represent variants found in only one type of the sample type, intersect represents variants found in both types of samples.

sample	blood		saliva	
	WES	**WGS**	**WES**	**WGS**
1	NM_000410.4 c.187C>G	NM_000410.4 c.187C>G	NM_000410.4 c.187C>G	NM_000410.4 c.187C>G
2	NM_000410.4 c.187C>G NM_000545.8 c.79A>C	NM_000410.4 c.187C>G NM_000545.8 c.79A>C	NM_000410.4 c.187C>G NM_000545.8 c.79A>C	NM_000410.4 c.187C>G NM_000545.8 c.79A>C
3	NM_000545.8 c.79A>C	NM_000545.8 c.79A>C	NM_000545.8 c.79A>C	NM_000545.8 c.79A>C
4	NM_000410.4 c.187C>G NM_001304717.5 c.10G>A NM_000545.8 c.79A>C	NM_000410.4 c.187C>G NM_001304717.5 c.10G>A NM_000545.8 c.79A>C	NM_000410.4 c.187C>G NM_001304717.5 c.10G>A NM_000545.8 c.79A>C	NM_000410.4 c.187C>G NM_001304717.5 c.10G>A NM_000545.8 c.79A>C
5	NM_000551.4 c.74C>T NM_024675.4 c.2993G>A NM_000152.5 c.2065G>A	NM_000551.4 c.74C>T NM_024675.4 c.2993G>A NM_000152.5 c.2065G>A	NM_000551.4 c.74C>T NM_024675.4 c.2993G>A NM_000152.5 c.2065G>A	NM_000551.4 c.74C>T NM_024675.4 c.2993G>A NM_000152.5 c.2065G>A
6	/	/	/	/
7	NM_000410.4 c.187C>G NM_000410.4 c.845G>A	NM_000410.4 c.187C>G NM_000410.4 c.845G>A	NM_000410.4 c.187C>G NM_000410.4 c.845G>A	NM_000410.4 c.187C>G NM_000410.4 c.845G>A
8	NM_000410.4 c.187C>G NM_000545.8 c.79A>C	NM_000410.4 c.187C>G NM_000545.8 c.79A>C	NM_000410.4 c.187C>G NM_000545.8 c.79A>C	NM_000410.4 c.187C>G NM_000545.8 c.79A>C
9	NM_000410.4 c.187C>G NM_001304717.5 c.10G>A NM_000545.8 c.79A>C	NM_000410.4 c.187C>G NM_001304717.5 c.10G>A NM_000545.8 c.79A>C	NM_000410.4 c.187C>G NM_001304717.5 c.10G>A NM_000545.8 c.79A>C	NM_000410.4 c.187C>G NM_001304717.5 c.10G>A NM_000545.8 c.79A>C
10	NM_001304717.5 c.10G>A NM_000545.8 c.79A>C	NM_001304717.5 c.10G>A NM_000545.8 c.79A>C	NM_001304717.5 c.10G>A NM_000545.8 c.79A>C	NM_001304717.5 c.10G>A NM_000545.8 c.79A>C

Table 8: Emedgene Analysis. Relevant variants in ACMG SF v3.2 list of genes detected by EMEDGENE software in paired blood and saliva samples.

5 DISCUSSION

Aim 1

As a consequence of the COVID-19 pandemic that broke out in 2020 and lasted for next two years, we used the SARS-CoV-2 coronavirus as a model organism (human gDNA samples were difficult to collect during the pandemic) with the goal to develop wet lab and data analysis protocol using the defined standard for protocol validation that would allow for a comprehensive analysis of the whole genome.

The first three SARS-CoV-2 positive cases in the Czech Republic were reported on March 1st, 2020, and the first-ever sequence of the "Czech" SARS CoV-2 virus was published on GISAID (GISAID, ID: EPI_ISL_414477) on March 12, 2020. This first "Czech" sample was partially sequenced using ONT technology. No sequencing guidelines for whole genome analysis of virus genome were established in the Czech Republic or elsewhere, and each laboratory attempted to sequence following their best practice and technology available. Although during the first months of the pandemic, thousands of SARS-CoV-2 genomes were published at GISAID (gisaid.org), only a part of them was sufficient for appropriate whole genome variant analysis. Since the outbreak of the pandemic, we had no doubt that the NGS approach is the only one capable of monitoring and tracking the SARS-CoV-2 virus genome development.

Samples for the first study were collected in April and May 2020 when, on average, 200 positively tested patients across the Czech Republic per day were identified by PCR tests [208]. Our goal was not to use NGS to detect the virus but to develop and validate sequencing protocol that would allow a high-quality SARS-CoV-2 whole-genome analysis - potentially including *de-novo* assembly of the entire SARS-CoV-2 genome. Samples were delivered in the format of RNA isolates, most of them of poor quality without any QC performed, with only SARS-CoV-2 positivity declaration. We used two capture-based approaches based on available test kits on the market; SARS-CoV-2 Research Panel (Twist Bioscience) that targets the complete coronavirus genome with approximately 1,000 probes, designed against the SARS-CoV-2 genome (GenBank: MN908947.3) and Virus panel (Illumina) that target over 40 respiratory viruses including SARS-CoV-2. Virus panel kit employs more than 7.800 probes designed against a set of viruses. Then, we used an amplicon-based approach utilizing CleanPlex® SARS-CoV-2 Research and Surveillance Panel (Paragon Genomics) with designed amplicon primers for the SARS-CoV-2 virus. To validate our library preparation, sequencing, and data analysis protocols from the point of view of variant analysis accuracy and specificity, we used the only standard available on the market at this time - synthetic

controls of Twist Bioscience. The use of synthetic controls allowed us to objectively evaluate the capture-based approach as superior to the amplicon-based approach, as the capture-based approach allowed us to detect and characterize all four variants defined in the synthetic control genome.

The whole analysis was complicated by the fact that we were not able to quantify the total amount of RNA via standard methods (ultraviolet absorbance, fluorescence or capillary electrophoresis) in the isolates due to their very low RNA concentration. We were able to evaluate RNA concentration only for approx. 30% of the isolates plus this concentration parameter represented total RNA in the sample, not specifically viral RNA [209]. After the first round of sequencing, where we sequenced 24 NGS libraries prepared by capture-based approach NEB+TWIST1 (see Experiments) using standard NGS library preparation protocol according to manufacturer's recommendations with no modifications, we observed high variability of genomic data quality among samples as well as among multiplexes. To adjust coverage uniformity within the multiplexes, we employed the RT-qPCR-based semi-quantitation method. We used the Direct SARS-CoV-2 RT-qPCR kit (Institute of Applied Biotechnologies) to measure Ct values of the respective samples [210]. Samples with similar Ct were subsequently multiplexed together in the second NEB+TWIST2 sequencing run. This approach resulted in higher uniformity of the median coverage of the respective viral genomes within the sequenced multiplexes. Subsequently, we multiplexed samples for the NEB+Illumina approach utilizing the same RT-qPCR procedure.

In addition to the capture-based approach, we analyzed an amplicon-based approach, represented by CleanPlex® SARS-CoV-2 Research and Surveillance Panel (Paragon Genomics), which should be able to sequence even one copy of the virus. Unfortunately, this NGS library preparation kit didn't work so effectively in our hands. Out of 40 isolates, we successfully prepared only 20 NGS libraries that met the required quality parameters for whole genome SARS-Cov-2 analysis, so we assumed this protocol was insufficient for clinical samples with very low RNA concentrations.

Ct values showed to be a reliable predictor for sample plexing during NGS library preparation from the perspective of whole genome variant analysis. The sample with the lowest Ct value of 11.29 (compared to Ct value of 25.8 for positive control PC4 with 1000 copies of synthetic control/reaction, this sample may contain around 10 million virus copies/reaction) provided good results for all three sequencing approaches. Good quality results (obtaining >20x coverage at least over 50 % bases) are evident until Ct value of 23 (compared to PC4, those sample contains approx. 10 000 of copies/reaction). Samples with a higher Ct value (≥ 23) exhibit, in general, a lower number of total

reads as well as a lower ratio of mapped reads and showed not to be appropriate for whole genome analysis.

We adopted the findings of this study to achieve the best possible results in the subsequent SARS-CoV-2 study, in which, in cooperation with leading clinical centers in the Czech Republic, we sequenced and analyzed the genomes of 229 samples collected during the first year of the SARS-CoV-2 pandemic in the Czech Republic [176].

Aim 2

After completing the first study, we employed all the methodology findings once we designed the major study, which focused on comparing genomic data quality obtained from paired human blood and saliva samples. This study systematically analyzed the comparison of genomic data obtained from saliva-derived gDNA and blood-derived gDNA for single nucleotide variants and small insertions and deletions from germline sequences for WGS and WES validated protocols.

The given findings would prove the usage of saliva as a suitable alternative material for genomic analysis compared to blood: firstly, saliva samples can be easily non-invasively collected without the need for medical expert assistance. Secondly, saliva samples provide sufficient gDNA as a starting material for whole genome analysis using a PCR-free NGS library preparation protocol. Also, the cell content of saliva, which is at least partially comparable to blood, was a decisive factor in the choice. Whole blood contains various blood cells, i.e., erythrocytes, leukocytes, and thrombocytes. Saliva samples also contain various cells, such as epithelial cells and leukocytes, as well as some microorganisms that vary from sample to sample [211]. Since gDNA from blood samples is mainly isolated from leukocytes, the leukocyte content makes saliva samples the best alternative material to blood compared to other oral specimens such as mouthwash or buccal swabs. It should be noted that a major disadvantage of saliva samples is their inherent microbial contamination, which can generate a large number of false-positive variants if NGS library preparation, sequencing, and/or data analysis is not performed properly [212].

What distinguishes our study from other studies [193] [213] [214] which somewhat addressed the suitability of saliva as an alternative material to blood, is that our study is the only one prospectively designed study to determine the variant concordance of paired blood-saliva samples by the end-to-end validated protocol. Also, compared to previous studies, we used the GRCh38 reference genome for sequencing alignment. First, we validated the WES and WGS protocols using the NIST reference standard Coriel sample NA 12878 (RS NA12878). We used the F1 score to calculate the variant concordance ratio in pairwise RS NA12878 samples. Our WGS

and WES protocol validation includes a pairwise triplicate-based comparison of RS NA12878 iterations against each other to simulate the comparison of two 100% identical samples. Additionally, the comparison restricted to the available TDS for RS NA12878 provided an exact comparison against a known set of variants. Then, we compared called variants from paired blood-saliva samples using the same end-to-end protocol used for the RS NA12878 comparison analysis. The mutual comparison of F1 scores of RS NA12878 iterations and paired blood-saliva samples was performed on pre-defined genome regions for WES and WGS protocols (whole genome, whole exome, non-difficult-to-sequence regions (NDR), high-confidence regions (HCR), region restricted to the intersection of both (HCR_NDR) and hard to sequence genomic regions (HARD). Since we used two sequencing protocols (whole-exome and whole-genome sequencing), we analyzed whether one of these approaches is more suitable for variant detection. For this purpose, we performed a cross-protocol variant comparison. We also examined the ratio of human to non-human sequencing reads in saliva-derived gDNA samples compared to blood-derived and RS NA 12878 ones, considering the potential impact of contamination on sequencing and genotyping accuracy. Finally, we analyzed the sequencing metrics of all sequencing runs. Detailed study design is described in Kvapilova et al., Fig. 1.

To determine the technical accuracy of the WGS and WES protocols, we used the fact that RS NA12878 is characterized by a well-defined truth data set of variants (TDS) of in high-confidence regions (HCR) [215]. Biologically, RS NA12878 is gDNA isolated from cultured cell lines [216], which can exhibit newly occurring "error" variants due to a natural mutation process that slowly accumulates variants from batch to batch. In the end, the cell line may carry variants not reflected in the TDS, so the TDS contains false-positive/negative variants in unknown proportion [186]. This was confirmed by a separate analysis of three iterations of RS NA 12878 followed by TDS comparison only. It is important to say that we also compared different variant callers using these RS NA 12878 iterations and TDS region at the beginning of our study to find the most appropriate one and selected Dragen version 3.10 as the one giving us the best concordance using the F1 score and other measures (unpublished data). When we examined cross-protocol genotype concordance in the individual RS NA12878 iterations, we saw high number of variants found exclusively (specifically) in the TDS or only in all three iterations for both SNVs and small-indels, in the WGS as well as WES data, respectively; see Kvapilova et al., Additional File 2, Fig. 5. However, this finding can (at least partially) also be explained as a technical-systematic error that affects both the WGS as well as WES protocols in the same way, creating real false positives/negatives variants compared to the TDS. Cross-protocol genotype concordance in selective blood or saliva samples may serve as evidence of

sequencing accuracy, as we see the same variants found in both sequencing approaches. The concordance in small-indels is slightly lower due to the WES protocol rather than due to sequencing inaccuracy [217].

We used the F1 score formula calculating the absolute concordance between two data sets of the same size. The F1 score is a harmonic mean of precision (truth positive predictions relative to total predicted positives, in other words, correctly identified positives) and recall (truth positive predictions relative to total actual positives, in other words incorrectly identified as positive). The F1 score reaches its best value at 1 (absolute concordance) and worst at 0 (no concordance). The of F1 score of 10 paired blood-saliva samples plus RS NA12878 triplicates ranged between 0.9712-0.9883 for SNVs and between 0.8880-0.9169 for small-indels in the case of the WES protocol, in the case of WGS F1 score ranged between 0.9742-0.9781 for SNVs and between 0.9489-0.9545 for small-indels.

The F1 scores of the paired blood-saliva samples have similar comparable medians and distributions to the F1 scores of the RS NA12878 pairwise triplicate comparison. The most significant difference from the perspective of the degree of variant concordance between the two paired blood-saliva samples is the genome region in which we compare the called variants. Generally, SNVs showed better concordance than small-indels in most of the genomic areas we studied. When we focused on specific parts of the genome region such as HCR/, HCR_NDR, or NDR, we saw an improvement in the median F1 scores for those regions in both WGS and WES approaches, for WGS and WES compared to analysis performed for whole genome or whole exome regions individually. As expected, the lowest F1 scores were found in the genomic regions that are hard to sequence in both WGS and WES protocols for SNVs and small-indels. All these findings support recent advice to evaluate and check the quality of sequencing data approaches for different types of genetic variations in different above-defined parts of the genome separately [218].

Non-human sequences found in saliva samples mainly belong to the human oral microbiome (was not studied thoroughly). Although analysis of the human oral microbiome is very important in assessing an individual's overall health [219], this was not part of our study. Non-human sequences found in saliva samples mainly belong to the human oral microbiome (was not studied thoroughly). Although analysis of the human oral microbiome is very important in assessing an individual's overall health [218], this was not part of our study. However, further investigation of these non-human sequences may enhance the overall clinical value of the sequencing data obtained from saliva samples using WGS approach. The blood samples contained the same proportion of human-mappable reads and non-human-mappable reads as the RS NA12878 samples. The blood samples

contained the same proportion of human-mappable reads, and HOMD-mappable reads as the RS NA12878 samples (around 4-5%). In contrast, the proportion of non-human-mappable reads in saliva samples was much higher than in blood or RS NA12878 samples. Moreover, despite the same saliva sample collection protocol followed by all study participants, the observed contamination was highly variable (8%-45% of WGS reads did not map to the human GRCh38 genome reference). Nevertheless, the variation in F1 scores between paired blood-saliva samples was slight, with no statistically significant influence of contamination on variant concordance when using our optimized protocol. Similar to Sosonkina et al. [220], we optimized the sample output of the NovaSeq 6000 sequencing experiment to achieve a specific minimum coverage of human-mappable reads by adding an extra low-pass sequencing step. Although low-pass pre-sequencing requires some time and money, it ensures uniform coverage of all samples without over-increasing the sequencing capacity.

Finally, we compared the quality of obtained sequencing analysis parameters of saliva and blood samples for both WES and WGS protocols. Surprisingly, variant calling quality measures were rather better in the saliva samples, at least for the WGS protocol, which was most likely the result of slightly higher coverage of WGS saliva samples compared to WGS blood samples. We also see shorter fragment lengths in saliva samples compared to blood samples. It seems that due to the aggressive microenvironment in the mouth, a higher amount of gDNA fragments are already naturally present in saliva samples.

Beyond the published study's scope, we compared variants' occurrence in the set of genes defined by the American College of Medicine genomics (unpublished data). The ACMG authority recommends analyzing this set of genes if germline data from WES or WGS are available. The ACMG v3.2 genes represent about 6.9 Mbp of the genome. When we apply our VCGS WES panel to cover the ACMG genes, we reduce the region of interest to only 0.5 Mb. Despite such a significant reduction of the analyzed area, we still analyze about 82% of the known pathogenic and likely pathogenic single nucleotide variants and 100 % of the pathogenic and likely pathogenic small-indels of the ACMG v3.2 genes according to the ClinVar database. We analyzed variant concordance in paired blood and saliva samples as well as in pairwise concordance in RS NA12878 iterations. Ideally, we should find all called variants in the respective intersections. This ideal match was seen in RS NA12878 iterations and only for WES SNVs data.

We inspected all unique SNVs and small-indels in WES and WGS_WES restricted data using the Integrative Genomics Viewer (IGV) [221]. SNVs unique only to saliva or blood samples, respectively, have in common that they are all found only in 6 genome loci. These loci are

characterized by high repetitiveness or homology, leading to the mapping of low-quality sequencing reads, so those unique variants are primarily false positives since sequencing reads are mapped with zero quality by our protocol mapping settings. The unique small-indels are found in the same loci as unique SNVs, challenging the same repetitiveness/homology issue, which is more evident in WES data. The IGV inspection of WES and WGS_WES restricted data of small-indels, exhibit differences in the length of insertion/deletion of repetitive bases, suggesting, after more complex reanalysis, small-indels concordance. The finding of extra variants in the WGS restricted to the VCGS exome dataset also points to false negative variants in the WES data, as IVG inspection proved. Finally, we use the EMEDGEN (Illumina) tool to search for and further interpret relevant variants in ACMG genes in paired blood-saliva samples. Except for sample 6, where no variants of interest were found in paired blood-saliva samples, the same variants of interest were found in both blood and saliva samples for WES and WGS data. In samples 1 and 2, we were able to detect hereditary influence from sample 4; in sample 8, we saw an inherited variant from sample 9.

In recent years, NGS has made significant progress in all areas of genomics and genetics. From single gene sequencing through to a small, restricted number of genes, to whole exome sequencing, and up to whole genome sequencing. Due to the continuously decreasing price per sequenced base, quality tools for analysis of data output, more straightforward interpretation, and thanks to the implementation of artificial intelligence in clinical diagnostics, WES is currently under increasing pressure from WGS [222]. There are still challenges in storing large amounts of data, but the recent clinical NGS genetic approaches target WGS. Indeed, the correct protocol comprising isolation of high-quality DNA, library preparation, sequencing strategy, and data analysis and interpretation is essential, and our study provides a comprehensive view and possible solutions leading to the reliable interpretation of NGS data from various sources for whole-genome sequencing and analysis.

6 CONCLUSIONS

Aim 1

In the case of finding the most appropriate sequencing approach for small viral genomes of SARS-CoV-2 in a mixed sample where both human material and viral content are presented, we evaluated the capture-based approach as more appropriate than the amplicon-based approach for whole genome analysis. To correctly determine the viral nucleic acid concentration using classical methods, we employed quantitative PCR to determine the viral content and established a benchmark level for accurate genomic analysis. The required quality of sequencing data for whole genome analysis in the case of the capture-based approach was obtained when the Ct value ranged between 16 and 23, and individual samples were multiplexed according to their Ct values. Although the capture-based approach is more time-consuming than the amplicon-based approach, the quality of the capture-based approach for whole genome analysis and variant detection was proved by implementing synthetic internal controls. We designed the protocol, including internal standards, by which we subsequently analyzed hundreds of SARS-CoV-2 genomes.

Aim 2

Human samples of different origins, blood and saliva, collected simultaneously from 10 individuals and processed with the same protocol from isolation, library prep, sequencing, up to the bioinformatic workflow, exhibit a concordance rate in paired blood-saliva samples for SNV and small-indels in predefined genomic regions significantly high enough to declare saliva as a suitable alternative material for population studies as well as clinical use compared to blood samples for WGS and WES analysis. Paired blood-saliva samples show the same distribution pattern of different sequencing parameters with no statistically significant differences. In addition, the NIST Coriel 12878 standard sample used for technical protocol validation showed a similar profile of sequencing parameters and variant distribution as the paired blood-saliva samples. Saliva samples exhibited a higher rate of non-human DNA compared to blood samples; however, there was no significant correlation between contamination levels and variant detection in saliva samples compared to blood samples. Cross-protocol genotype concordance revealed high concordance rates for both SNVs and small-indels in both sequencing protocols, indicating consistent variant detection regardless of the sequencing approach. Beyond the scope of the published study, we analyzed SNVs and small-indels in paired blood-saliva samples for concordance in the number of relevant variants in the set of genes

defined by the American College of Medical Genetics and Genomics. Significant concordance was achieved across the paired blood-saliva samples as well as for the sequencing approaches for SNVs and small-indels.

LIST OF ABBREVIATIONS

ACMG SF	American College of Medical Genetics and Genomics, secondary findings
5mC	5-methylcytosine
6mA	N6-methyladenine
5hmC	5- hydroxymethylcytosine
AMR	antimicrobial resistance
BAM	binary alignment and map
BED	browser extensible data
bp	base pairs
BrdU	5-bromodeoxyuridine
cDNA	copy DNA
CES	clinical exome sequencing
CNV	copy number variant
Ct	cycle threshold
DNA	deoxyribonucleic acid
dNTP	deoxynucleotide triphosphate
ddNTP	di-deoxynucleotide triphosphate
ELISA	the enzyme-linked immunosorbent assay
FISH	fluorescence in situ hybridization
FFPE	formalin-fixed, paraffin-embedded
gDNA	genomic DNA
GIAB	Genome in the Bottle
GRC	Genome Reference Consortium
HGP	Human Genome Project
HCR	high-confidence region

HOMD	Human Oral Microbiome Database
HGP	Human Genome Project
IGV	Integrative Genomics Viewer
m6A	N6-methyladenosine (m6A)
MAPQ10	mapping quality 10
MEGA	Million European Genomes Alliance
NA	nucleic acid
NCBI	National Center for Biotechnology Information
NGS	next generation sequencing
NIPT	non-invasive prenatal testing
NIST	National Institute for Standards and Technology
ONT	Oxford Nanopore sequency technology
PCR	polymerase chain reaction
PE	paired-end
QC	quality control
qPCR	quantitative PCR
RNA	ribonucleic acid
RNA-seq	RNA sequencing
RS	reference standard
S	saliva
ssDNA	single-strended DNA
scRNA-seq	single cell RNA sequencing
SNV	single nucleotide variant
SNP	single nucleotide polymorphism

ssDNA	single-stranded DNA
SMRT	Single Molecule Real-Time sequencing technology
Indel	insertion/deletion
ISS	In situ sequencing
N6mA	N6-methyladenine
NCBI	National Center for Biotechnology Information
NDR	non-difficult regions
T2T	Telomere to Telomere Consortium
TDS	truth dataset (ground truth variants in NA12878 reference sample)
TR	tandem repeat
UMI	unique molecular identifiers
VCF	variant call format
WB	whole blood
WES	whole exome sequencing
WGS	whole genome sequencing

7 CONTRIBUTION TO THE PUBLICATIONS

Aim 1

Klempt, P.; Brož, P.; Kašný, M.; Novotný, A.; Kvapilová, K.; Kvapil, P. Performance of Targeted Library Preparation Solutions for SARS-CoV-2 Whole Genome Analysis. *Diagnostics* **2020**, *10*, 769. https://doi.org/10.3390/diagnostics10100769

I made proposals for the methods to be used in the study, to ensure that our research approach was well designed and scientifically rigorous. I created a graphical overview for the process scheme time consumption for the library preparation approaches and was involved in writing the first draft of the manuscript.

Aim 2

Kvapilova, K., Misenko, P., Radvanszky, J. *et al.* Validated WGS and WES protocols proved saliva-derived gDNA as an equivalent to blood-derived gDNA for clinical and population genomic analyses. *BMC Genomics* **25**, 187 (2024). https://doi.org/10.1186/s12864-024-10080-0

I proposed the topic of the study and was responsible for the design of the entire study. I was actively involved in sample collection, participated in the NGS library preparation, quality control of NGS libraries and sequencing. I played a major role in writing the first draft of the manuscript, as well the coordination of the revision of the manuscript after comments from reviewers.

REFERENCES

This thesis was based on the following publications:

[1] JONES, Mary Ellen. Albrecht Kossel, a biographical sketch. *The Yale journal of biology and medicine*. September 1953, č. 26, s. 80-97.

[2] SANGER, F.; AIR, G. M.; BARRELL, B. G.; BROWN, N. L.; COULSON, A. R. et al. Nucleotide sequence of bacteriophage φX174 DNA. Online. *Nature*. 1977, roč. 265, č. 5596, s. 687-695. ISSN 0028-0836. Dostupné z: https://doi.org/10.1038/265687a0. [cit. 2024-01-19].

[3] BAER, R.; BANKIER, A. T.; BIGGIN, M. D.; DEININGER, P. L.; FARRELL, P. J. et al. DNA sequence and expression of the B95-8 Epstein—Barr virus genome. Online. *Nature*. 1984, roč. 310, č. 5974, s. 207-211. ISSN 0028-0836. Dostupné z: https://doi.org/10.1038/310207a0. [cit. 2024-01-23].

[4] GUSELLA, James F.; WEXLER, Nancy S.; CONNEALLY, P. Michael; NAYLOR, Susan L.; ANDERSON, Mary Anne et al. A polymorphic DNA marker genetically linked to Huntington's disease. Online. *Nature*. 1983, roč. 306, č. 5940, s. 234-238. ISSN 0028-0836. Dostupné z: https://doi.org/10.1038/306234a0. [cit. 2024-01-23].

[5] SINSHEIMER, Robert L. The Santa Cruz Workshop—May 1985. Online. *Genomics*. 1989, roč. 5, č. 4, s. 954-956. ISSN 08887543. Dostupné z: https://doi.org/10.1016/0888-7543(89)90142-0. [cit. 2024-02-04].

[6] DULBECCO, Renato. A Turning Point in Cancer Research: Sequencing the Human Genome. Online. *Science*. 1986, roč. 231, č. 4742, s. 1055-1056. ISSN 0036-8075. Dostupné z: https://doi.org/10.1126/science.3945817. [cit. 2024-01-24].

[7] NIH. *Understanding Our Genetic Inheritance: The US Human Genome Project, the First Five Years FY 1991--1995*. Online. 1990. Dostupné z: https://digital.library.unt.edu/ark:/67531/metadc1191272/. [cit. 2024-01-23].

[8] COLLINS, Francis S.; MORGAN, Michael a PATRINOS, Aristides. The Human Genome Project: Lessons from Large-Scale Biology. Online. *Science*. 2003, roč. 300, č. 5617, s. 286-290. ISSN 0036-8075. Dostupné z: https://doi.org/10.1126/science.1084564. [cit. 2024-01-26].

[9] LANDER, Eric S.; LINTON, Lauren M.; BIRREN, Bruce; NUSBAUM, Chad; ZODY, Michael C. et al. Initial sequencing and analysis of the human genome. Online. *Nature*. 2001, roč. 409, č. 6822, s. 860-921. ISSN 0028-0836. Dostupné z: https://doi.org/10.1038/35057062. [cit. 2024-01-19].

[10] VENTER, J. Craig; ADAMS, Mark D.; MYERS, Eugene W.; LI, Peter W.; MURAL, Richard J. et al. The Sequence of the Human Genome. Online. *Science*. 2001, roč. 291, č. 5507, s. 1304-1351. ISSN 0036-8075. Dostupné z: https://doi.org/10.1126/science.1058040. [cit. 2024-01-19].

[11] *Https://www.genome.gov/27555238/april-2013-the-10year-anniversary-of-the-human-genome-project-commemorating-and-reflecting*. Online. [cit. 2024-01-24].

[12] Finishing the euchromatic sequence of the human genome. Online. *Nature*. 2004, roč. 431, č. 7011, s. 931-945. ISSN 0028-0836. Dostupné z: https://doi.org/10.1038/nature03001. [cit. 2024-01-19].

[13] *NIH, Human Genome Project*. Online. August 24, 2022, August 24, 2022. Dostupné z: https://www.genome.gov/about-genomics/educational-resources/fact-sheets/human-genome-project. [cit. 2024-01-19].

[14] VEERAMACHANENI, Vamsi; MAKALOWSKI, Wojciech; GALDZICKI, Michal; SOOD, Raman a MAKALOWSKA, Izabela. Mammalian Overlapping Genes: The Comparative Perspective. Online. *Genome Research*. 2004, roč. 14, č. 2, s. 280-286. ISSN 1088-9051. Dostupné z: https://doi.org/10.1101/gr.1590904. [cit. 2024-01-24].

[15] LEVY, Samuel; SUTTON, Granger; NG, Pauline C; FEUK, Lars; HALPERN, Aaron L et al. The Diploid Genome Sequence of an Individual Human. Online. *PLoS Biology*. 2007, roč. 5, č. 10, s. 2113 - 2144. ISSN 1545-7885. Dostupné z: https://doi.org/10.1371/journal.pbio.0050254. [cit. 2024-01-19].

[16] MARGULIES, Marcel; EGHOLM, Michael; ALTMAN, William E.; ATTIYA, Said a BADER, Joel S. Genome sequencing in microfabricated high-density picolitre reactors. Online. *Nature*. 2005, roč. 437, č. 7057, s. 376-380. ISSN 0028-0836. Dostupné z: https://doi.org/10.1038/nature03959. [cit. 2024-01-23].

[17] WHEELER, David A.; SRINIVASAN, Maithreyan; EGHOLM, Michael; SHEN, Yufeng; CHEN, Lei et al. The complete genome of an individual by massively parallel DNA sequencing. Online. *Nature*. 2008, roč. 452, č. 7189, s. 872-876. ISSN 0028-0836. Dostupné z: https://doi.org/10.1038/nature06884. [cit. 2024-01-19].

[18] MAMANOVA, Lira; COFFEY, Alison J; SCOTT, Carol E; KOZAREWA, Iwanka; TURNER, Emily H et al. Target-enrichment strategies for next-generation sequencing. Online. *Nature Methods*.

2010, roč. 7, č. 2, s. 111-118. ISSN 1548-7091. Dostupné z: https://doi.org/10.1038/nmeth.1419. [cit. 2024-01-20].

[19] NG, Sarah B.; TURNER, Emily H.; ROBERTSON, Peggy D.; FLYGARE, Steven D.; BIGHAM, Abigail W. et al. Targeted capture and massively parallel sequencing of 12 human exomes. Online. *Nature*. 2009, roč. 461, č. 7261, s. 272-276. ISSN 0028-0836. Dostupné z: https://doi.org/10.1038/nature08250. [cit. 2024-01-19].

[20] NIH. *Human Genome Cost*. Online. 2021, November 1, 2021. Dostupné z: https://www.genome.gov/about-genomics/fact-sheets/Sequencing-Human-Genome-cost. [cit. 2024-01-23].

[21] BUCOSSI, Serena; POLIMANTI, Renato; VENTRIGLIA, Mariacarla; MARIANI, Stefania; SIOTTO, Mariacristina et al. Intronic rs2147363 Variant in ATP7B Transcription Factor-Binding Site Associated with Alzheimer's Disease. Online. *Journal of Alzheimer's Disease*. 2013, roč. 37, č. 2, s. 453-459. ISSN 18758908. Dostupné z: https://doi.org/10.3233/JAD-130431. [cit. 2024-01-20].

[22] BACH, Elisa; WOLF, Beat; OLDENBURG, Johannes; MÜLLER, Clemens a ROST, Simone. Identification of deep intronic variants in 15 haemophilia A patients by next generation sequencing of the whole factor VIII gene. Online. *Thrombosis and Haemostasis*. 2017, roč. 114, č. 10, s. 757-767. ISSN 0340-6245. Dostupné z: https://doi.org/10.1160/TH14-12-1011. [cit. 2024-01-20].

[23] GELFMAN, Sahar; WANG, Quanli; MCSWEENEY, K. Melodi; REN, Zhong a GOLDSTEIN, David B. Annotating pathogenic non-coding variants in genic regions. Online. *Nature Communications*. 2017, roč. 8, č. 1, s. 1-11. ISSN 2041-1723. Dostupné z: https://doi.org/10.1038/s41467-017-00141-2. [cit. 2024-01-20].

[24] ALIOTO, Tyler S.; BUCHHALTER, Ivo; DERDAK, Sophia; HUTTER, Barbara; ELDRIDGE, Matthew D. et al. A comprehensive assessment of somatic mutation detection in cancer using whole-genome sequencing. Online. *Nature Communications*. 2015, roč. 6, č. 1, s. Results. ISSN 2041-1723. Dostupné z: https://doi.org/10.1038/ncomms10001. [cit. 2024-01-20].

[25] MARSHALL, Christian R.; CHOWDHURY, Shimul; TAFT, Ryan J.; LEBO, Mathew S. a BUCHAN, Jillian G. Best practices for the analytical validation of clinical whole-genome sequencing intended for the diagnosis of germline disease. Online. *Npj Genomic Medicine*. 2020, roč. 5, č. 1, s. Test Validation. ISSN 2056-7944. Dostupné z: https://doi.org/10.1038/s41525-020-00154-9. [cit. 2024-01-20].

[26] SMITH, Hadley Stevens; SWINT, J. Michael; LALANI, Seema R.; YAMAL, Jose-Miguel a DE OLIVEIRA OTTO, Marcia C. Clinical Application of Genome and Exome Sequencing as a

Diagnostic Tool for Pediatric Patients: a Scoping Review of the Literature. Online. *Genetics in Medicine*. 2019, roč. 21, č. 1, s. 3-16. ISSN 10983600. Dostupné z: https://doi.org/10.1038/s41436-018-0024-6. [cit. 2024-01-20].

[27] WATSON, J. D. a CRICK, F. H. C. Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. Online. *Nature*. 1953, roč. 171, č. 4356, s. 737-738. ISSN 0028-0836. Dostupné z: https://doi.org/10.1038/171737a0. [cit. 2024-01-24].

[28] JACKSON, David A.; SYMONS, Robert H. a BERG, Paul. Biochemical Method for Inserting New Genetic Information into DNA of Simian Virus 40: Circular SV40 DNA Molecules Containing Lambda Phage Genes and the Galactose Operon of Escherichia coli. Online. *Proceedings of the National Academy of Sciences*. 1972, roč. 69, č. 10, s. 2904-2909. ISSN 0027-8424. Dostupné z: https://doi.org/10.1073/pnas.69.10.2904. [cit. 2024-02-18].

[29] LEHMAN, I.R.; BESSMAN, Maurice J.; SIMMS, Ernest S. a KORNBERG, Arthur. Enzymatic Synthesis of Deoxyribonucleic Acid. Online. *Journal of Biological Chemistry*. 1958, roč. 233, č. 1, s. 163-170. ISSN 00219258. Dostupné z: https://doi.org/10.1016/S0021-9258(19)68048-8. [cit. 2024-01-24].

[30] SANGER, F. a COULSON, A.R. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. Online. *Journal of Molecular Biology*. 1975, roč. 94, č. 3, s. 441-448. ISSN 00222836. Dostupné z: https://doi.org/10.1016/0022-2836(75)90213-2. [cit. 2024-01-20].

[31] SANGER, F.; NICKLEN, S. a COULSON, A. R. DNA sequencing with chain-terminating inhibitors. Online. *Proceedings of the National Academy of Sciences*. 1977, roč. 74, č. 12, s. 5463-5467. ISSN 0027-8424. Dostupné z: https://doi.org/10.1073/pnas.74.12.5463. [cit. 2024-01-20].

[32] MAXAM, A M a GILBERT, W. A new method for sequencing DNA. Online. *Proceedings of the National Academy of Sciences*. 1977, roč. 74, č. 2, s. 560-564. ISSN 0027-8424. Dostupné z: https://doi.org/10.1073/pnas.74.2.560. [cit. 2024-01-20].

[33] JORGENSON, James W. a LUKACS, Krynn DeArman. Zone electrophoresis in open-tubular glass capillaries. Online. *Analytical Chemistry*. 1981, roč. 53, č. 8, s. 1298-1302. ISSN 0003-2700. Dostupné z: https://doi.org/10.1021/ac00231a037. [cit. 2024-01-24].

[34] KOUMI, Pieris; GREEN, Helen E.; HARTLEY, Susan; JORDAN, Darren; LAHEC, Sharon et al. Evaluation and validation of the ABI 3700, ABI 3100, and the MegaBACE 1000 capillary array electrophoresis instruments for use with short tandem repeat microsatellite typing in a forensic

environment. Online. *ELECTROPHORESIS*. 2004, roč. 25, č. 14, s. 2227-2241. ISSN 0173-0835. Dostupné z: https://doi.org/10.1002/elps.200305976. [cit. 2024-01-24].

[35] KULSKI, Jerzy K. Next-Generation Sequencing — An Overview of the History, Tools, and "Omic" Applications. Online. In: KULSKI, Jerzy K. (ed.). *Next Generation Sequencing - Advances, Applications and Challenges*. InTech, 2016, s. 1-60. ISBN 978-953-51-2240-1. Dostupné z: https://doi.org/10.5772/61964. [cit. 2024-02-12].

[36] HEATHER, James M. a CHAIN, Benjamin. The sequence of sequencers: The history of sequencing DNA. Online. *Genomics*. 2016, roč. 107, č. 1, s. 1-8. ISSN 08887543. Dostupné z: https://doi.org/10.1016/j.ygeno.2015.11.003. [cit. 2024-01-24].

[37] ROTHBERG, Jonathan M a LEAMON, John H. The development and impact of 454 sequencing. Online. *Nature Biotechnology*. 2008, roč. 26, č. 10, s. 1117-1124. ISSN 1087-0156. Dostupné z: https://doi.org/10.1038/nbt1485. [cit. 2024-01-24].

[38] BENTLEY, David R.; BALASUBRAMANIAN, Shankar; SWERDLOW, Harold P.; SMITH, Geoffrey P.; MILTON, John et al. Accurate whole human genome sequencing using reversible terminator chemistry. Online. *Nature*. 2008, roč. 456, č. 7218, s. 53-59. ISSN 0028-0836. Dostupné z: https://doi.org/10.1038/nature07517. [cit. 2024-01-20].

[39] VOELKERDING, Karl V; DAMES, Shale A a DURTSCHI, Jacob D. Next-Generation Sequencing: From Basic Research to Diagnostics. Online. *Clinical Chemistry*. 2009, roč. 55, č. 4, s. 641-658. ISSN 0009-9147. Dostupné z: https://doi.org/10.1373/clinchem.2008.112789. [cit. 2024-01-24].

[40] MERRIMAN, Barry; R&D TEAM, Ion Torrent a ROTHBERG, Jonathan M. Progress in Ion Torrent semiconductor chip based sequencing. Online. *ELECTROPHORESIS*. 2012, roč. 33, č. 23, s. 3397-3417. ISSN 0173-0835. Dostupné z: https://doi.org/10.1002/elps.201200424. [cit. 2024-01-24].

[41] DRMANAC, Radoje; SPARKS, Andrew B.; CALLOW, Matthew J.; HALPERN, Aaron L.; BURNS, Norman L. et al. Human Genome Sequencing Using Unchained Base Reads on Self-Assembling DNA Nanoarrays. Online. *Science*. 2010, roč. 327, č. 5961, s. 78-81. ISSN 0036-8075. Dostupné z: https://doi.org/10.1126/science.1181498. [cit. 2024-02-11].

[42] WANG, Ou; CHIN, Robert; CHENG, Xiaofang; WU, Michelle Ka Yan; MAO, Qing et al. Efficient and unique cobarcoding of second-generation sequencing reads from long DNA molecules enabling cost-effective and accurate sequencing, haplotyping, and de novo assembly. Online. *Genome Research*. 2019, roč. 29, č. 5, s. 798-808. ISSN 1088-9051. Dostupné z: https://doi.org/10.1101/gr.245126.118. [cit. 2024-02-11].

[43] BOWERS, Jayson; MITCHELL, Judith; BEER, Eric; BUZBY, Philip R; CAUSEY, Marie et al. Virtual terminator nucleotides for next-generation DNA sequencing. Online. *Nature Methods*. 2009, roč. 6, č. 8, s. 593-595. ISSN 1548-7091. Dostupné z: https://doi.org/10.1038/nmeth.1354. [cit. 2024-01-24].

[44] MASHAYEKHI, Foad a RONAGHI, Mostafa. Analysis of read length limiting factors in Pyrosequencing chemistry. Online. *Analytical Biochemistry*. 2007, roč. 363, č. 2, s. 275-287. ISSN 00032697. Dostupné z: https://doi.org/10.1016/j.ab.2007.02.002. [cit. 2024-01-20].

[45] JAIN, Miten; OLSEN, Hugh E.; PATEN, Benedict a AKESON, Mark. The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. Online. *Genome Biology*. 2016, roč. 17, č. 1, s. 1-11. ISSN 1474-760X. Dostupné z: https://doi.org/10.1186/s13059-016-1103-0. [cit. 2024-01-20].

[46] RHOADS, Anthony a AU, Kin Fai. PacBio Sequencing and Its Applications. Online. *Genomics, Proteomics & Bioinformatics*. 2015, roč. 13, č. 5, s. 278-289. ISSN 16720229. Dostupné z: https://doi.org/10.1016/j.gpb.2015.08.002. [cit. 2024-01-24].

[47] LU, Hengyun; GIORDANO, Francesca a NING, Zemin. Oxford Nanopore MinION Sequencing and Genome Assembly. Online. *Genomics, Proteomics & Bioinformatics*. 2016, roč. 14, č. 5, s. 265-279. ISSN 16720229. Dostupné z: https://doi.org/10.1016/j.gpb.2016.05.004. [cit. 2024-01-24].

[48] TOURDOT, Richard W.; BRUNETTE, Gregory J.; PINTO, Ricardo A. a ZHANG, Cheng-Zhong. Determination of complete chromosomal haplotypes by bulk DNA sequencing. Online. *Genome Biology*. 2021, roč. 22, č. 1, s. 1-31. ISSN 1474-760X. Dostupné z: https://doi.org/10.1186/s13059-021-02330-1. [cit. 2024-01-24].

[49] AMMAR, Ron; PATON, Tara A.; TORTI, Dax; SHLIEN, Adam a BADER, Gary D. Long read nanopore sequencing for detection of HLA and CYP2D6 variants and haplotypes. Online. *F1000Research*. 2015, roč. 4, s. 1-19. ISSN 2046-1402. Dostupné z: https://doi.org/10.12688/f1000research.6037.2. [cit. 2024-01-24].

[50] CHEN, Jinfeng; CHENG, Jingfei; CHEN, Xiufei; INOUE, Masato; LIU, Yibin et al. Whole-genome long-read TAPS deciphers DNA methylation patterns at base resolution using PacBio SMRT sequencing technology. Online. *Nucleic Acids Research*. 2022, roč. 50, č. 18, s. e104-e104. ISSN 0305-1048. Dostupné z: https://doi.org/10.1093/nar/gkac612. [cit. 2024-01-24].

[51] LIU, Yang; ROSIKIEWICZ, Wojciech; PAN, Ziwei; JILLETTE, Nathaniel; WANG, Ping et al. DNA methylation-calling tools for Oxford Nanopore sequencing: a survey and human epigenome-wide

evaluation. Online. *Genome Biology*. 2021, roč. 22, č. 1, s. 1-33. ISSN 1474-760X. Dostupné z: https://doi.org/10.1186/s13059-021-02510-z. [cit. 2024-01-24].

[52] COZZUTO, Luca; LIU, Huanle; PRYSZCZ, Leszek P.; PULIDO, Toni Hermoso; DELGADO-TEJEDOR, Anna et al. MasterOfPores: A Workflow for the Analysis of Oxford Nanopore Direct RNA Sequencing Datasets. Online. *Frontiers in Genetics*. 2020, roč. 11, s. 1-11. ISSN 1664-8021. Dostupné z: https://doi.org/10.3389/fgene.2020.00211. [cit. 2024-01-24].

[53] SAHLIN, Kristoffer a MEDVEDEV, Paul. Error correction enables use of Oxford Nanopore technology for reference-free transcriptome analysis. Online. *Nature Communications*. 2021, roč. 12, č. 1, s. 1-13. ISSN 2041-1723. Dostupné z: https://doi.org/10.1038/s41467-020-20340-8. [cit. 2024-01-24].

[54] MORISSE, Pierre; MARCHET, Camille; LIMASSET, Antoine; LECROQ, Thierry a LEFEBVRE, Arnaud. Scalable long read self-correction and assembly polishing with multiple sequence alignment. Online. *Scientific Reports*. 2021, roč. 11, č. 1, s. Results. ISSN 2045-2322. Dostupné z: https://doi.org/10.1038/s41598-020-80757-5. [cit. 2024-01-24].

[55] SALMELA, Leena a RIVALS, Eric. LoRDEC: accurate and efficient long read error correction. Online. *Bioinformatics*. 2014, roč. 30, č. 24, s. 3506-3514. ISSN 1367-4811. Dostupné z: https://doi.org/10.1093/bioinformatics/btu538. [cit. 2024-01-24].

[56] MADOUI, Mohammed-Amin; ENGELEN, Stefan; CRUAUD, Corinne; BELSER, Caroline; BERTRAND, Laurie et al. Genome assembly using Nanopore-guided long and error-free DNA reads. Online. *BMC Genomics*. 2015, roč. 16, č. 1, s. 3506–3514. ISSN 1471-2164. Dostupné z: https://doi.org/10.1186/s12864-015-1519-z. [cit. 2024-01-24].

[57] Cost-efficient whole genome-sequencing using novel mostly natural sequencing-by-synthesis chemistry and open fluidics platform. *BioRxiv*. May 30, 2022, roč. 11, s. 1-8.

[58] ARSLAN, Sinan; GARCIA, Francisco J.; GUO, Minghao a KELLINGER, Matthew W. Sequencing by avidity enables high accuracy with low reagent consumption. Online. *Nature Biotechnology*. 2024, roč. 42, č. 1, s. 132-138. ISSN 1087-0156. Dostupné z: https://doi.org/10.1038/s41587-023-01750-7. [cit. 2024-01-24].

[59] PRNEWSWIRE. *MGI Secured First Corporate Order of Ultra-high Throughput Sequencer DNBSEQ-T20×2**. Online. PRNEWSWIRE. MGI Secured First Corporate Order of Ultra-high Throughput Sequencer DNBSEQ-T20×2*. 26 Apr, 2023n. l., 2024. Dostupné z:

https://www.prnewswire.com/news-releases/mgi-secured-first-corporate-order-of-ultra-high-throughput-sequencer-dnbseq-t20O2-301807850.html. [cit. 2024-02-11].

[60] GOUIN, Kenneth; LAMARCA, Liz; XIANG, Yu; DECKER, Anne; SHORE, Sabrina et al. Abstract 220: Performance assessment of the novel G4 sequencing platform for cancer research applications. Online. *Cancer Research*. 2023, roč. 83, č. 7_Supplement, s. 220-220. ISSN 1538-7445. Dostupné z: https://doi.org/10.1158/1538-7445.AM2023-220. [cit. 2024-02-11].

[61] SMITH, Lloyd M.; FUNG, Steven; HUNKAPILLER, Michael W.; HUNKAPILLER, Tim J. a HOOD, Leroy E. The synthesis of oligonucleotides containing an aliphatic amino group at the 5′ terminus: synthesis of fluorescent DNA primers for use in DNA sequence analysis. Online. *Nucleic Acids Research*. 1985, roč. 13, č. 7, s. 2399-2412. ISSN 0305-1048. Dostupné z: https://doi.org/10.1093/nar/13.7.2399. [cit. 2024-01-28].

[62] PROBER, James M.; TRAINOR, George L.; DAM, Rudy J.; HOBBS, Frank W.; ROBERTSON, Charles W. et al. A System for Rapid DNA Sequencing with Fluorescent Chain-Terminating Dideoxynucleotides. Online. *Science*. 1987, roč. 238, č. 4825, s. 336-341. ISSN 0036-8075. Dostupné z: https://doi.org/10.1126/science.2443975. [cit. 2024-01-29].

[63] SMITH, Lloyd M.; SANDERS, Jane Z.; KAISER, Robert J.; HUGHES, Peter; DODD, Chris et al. Fluorescence detection in automated DNA sequence analysis. Online. *Nature*. 1986, roč. 321, č. 6071, s. 674-679. ISSN 0028-0836. Dostupné z: https://doi.org/10.1038/321674a0. [cit. 2024-02-11].

[64] SLATKO, Barton E.; GARDNER, Andrew F. a AUSUBEL, Frederick M. Overview of Next-Generation Sequencing Technologies. Online. *Current Protocols in Molecular Biology*. 2018, roč. 122, č. 1, s. 2-15. ISSN 1934-3639. Dostupné z: https://doi.org/10.1002/cpmb.59. [cit. 2024-01-29].

[65] *3730xl DNA Analyzer*. Online. ThermoFisher. 2024. Dostupné z: https://www.thermofisher.com/order/catalog/product/A41046?SID=srch-srp-A41046).. [cit. 2024-01-28].

[66] ILLUMINA. *Illumina Systems*. Online. ILLUMINA. Illumina Systems. 2024, 2024. Dostupné z: https://emea.illumina.com/systems.html. [cit. 2024-02-15].

[67] ILLUMINA. *Sequencing Quality Scores*. Online. ILLUMINA. Sequencing Quality Scores. 2024, 2024. Dostupné z: https://emea.illumina.com/science/technology/next-generation-sequencing/plan-experiments/quality-scores.html. [cit. 2024-02-15].

[68] SOSINSKY, Alona; AMBROSE, John; CROSS, William; TURNBULL, Clare; HENDERSON, Shirley et al. Insights for precision oncology from the integration of genomic and clinical data of

13,880 tumors from the 100,000 Genomes Cancer Programme. Online. *Nature Medicine*. 2024, roč. 30, č. 1, s. 279-289. ISSN 1078-8956. Dostupné z: https://doi.org/10.1038/s41591-023-02682-0. [cit. 2024-01-29].

[69] BECK, Tyler F; MULLIKIN, James C a BIESECKER, Leslie G. Systematic Evaluation of Sanger Validation of Next-Generation Sequencing Variants. Online. *Clinical Chemistry*. 2016, roč. 62, č. 4, s. 647-654. ISSN 0009-9147. Dostupné z: https://doi.org/10.1373/clinchem.2015.249623. [cit. 2024-01-20].

[70] ARTECHE-LÓPEZ, A.; ÁVILA-FERNÁNDEZ, A.; ROMERO, R.; RIVEIRO-ÁLVAREZ, R.; LÓPEZ-MARTÍNEZ, M. A. et al. Sanger sequencing is no longer always necessary based on a single-center validation of 1109 NGS variants in 825 clinical exomes. Online. *Scientific Reports*. 2021, roč. 11, č. 1, s. 1-7. ISSN 2045-2322. Dostupné z: https://doi.org/10.1038/s41598-021-85182-w. [cit. 2024-01-28].

[71] FIERS, W.; CONTRERAS, R.; DUERINCK, F.; HAEGEMAN, G.; ISERENTANT, D. et al. Complete nucleotide sequence of bacteriophage MS2 RNA: primary and secondary structure of the replicase gene. Online. *Nature*. 1976, roč. 260, č. 5551, s. 500-507. ISSN 0028-0836. Dostupné z: https://doi.org/10.1038/260500a0. [cit. 2024-01-29].

[72] FLEISCHMANN, Robert D.; ADAMS, Mark D.; WHITE, Owen; CLAYTON, Rebecca A.; KIRKNESS, Ewen F. et al. Whole-Genome Random Sequencing and Assembly of Haemophilus influenzae Rd. Online. *Science*. 1995, roč. 269, č. 5223, s. 496-512. ISSN 0036-8075. Dostupné z: https://doi.org/10.1126/science.7542800. [cit. 2024-01-29].

[73] FRASER, Claire M.; GOCAYNE, Jeannine D.; WHITE, Owen; ADAMS, Mark D.; CLAYTON, Rebecca A. et al. The Minimal Gene Complement of Mycoplasma genitalium. Online. *Science*. 1995, roč. 270, č. 5235, s. 397-404. ISSN 0036-8075. Dostupné z: https://doi.org/10.1126/science.270.5235.397. [cit. 2024-01-29].

[74] SOHN, Jang-il a NAM, Jin-Wu. The present and future of de novo whole-genome assembly. Online. *Briefings in Bioinformatics*. 2018, roč. 19, č. 1, s. 23-40. ISSN 1467-5463. Dostupné z: https://doi.org/10.1093/bib/bbw096. [cit. 2024-02-24].

[75] WONG, Karen H. Y.; LEVY-SAKIN, Michal a KWOK, Pui-Yan. De novo human genome assemblies reveal spectrum of alternative haplotypes in diverse populations. Online. *Nature Communications*. 2018, roč. 9, č. 1, s. 1-9. ISSN 2041-1723. Dostupné z: https://doi.org/10.1038/s41467-018-05513-w. [cit. 2024-01-30].

[76] PAUL, Bobby; DIXIT, Gunjan; MURALI, Thokur Sreepathy; SATYAMOORTHY, Kapaettu a HAO, W. Genome-based taxonomic classification. Online. *Genome*. 2019, roč. 62, č. 2, s. 45-52. ISSN 0831-2796. Dostupné z: https://doi.org/10.1139/gen-2018-0072. [cit. 2024-01-29].

[77] TURRO, Ernest; ASTLE, William J.; MEGY, Karyn; GRÄF, Stefan; GREENE, Daniel et al. Whole-genome sequencing of patients with rare diseases in a national health system. Online. *Nature*. 2020, roč. 583, č. 7814, s. 96-102. ISSN 0028-0836. Dostupné z: https://doi.org/10.1038/s41586-020-2434-2. [cit. 2024-01-31].

[78] SOSINSKY, Alona; AMBROSE, John; CROSS, William; TURNBULL, Clare; HENDERSON, Shirley et al. Insights for precision oncology from the integration of genomic and clinical data of 13,880 tumors from the 100,000 Genomes Cancer Programme. Online. *Nature Medicine*. 2024, roč. 30, č. 1, s. 279-289. ISSN 1078-8956. Dostupné z: https://doi.org/10.1038/s41591-023-02682-0. [cit. 2024-01-31].

[79] CASPAR, Sylvan Manuel; SCHNEIDER, Timo; STOLL, Patricia; MEIENBERG, Janine a MATYAS, Gabor. Potential of whole-genome sequencing-based pharmacogenetic profiling. Online. *Pharmacogenomics*. 2021, roč. 22, č. 3, s. 177-190. ISSN 1462-2416. Dostupné z: https://doi.org/10.2217/pgs-2020-0155. [cit. 2024-01-29].

[80] MORRIS, Huw R; HOULDEN, Henry a POLKE, James. Whole-genome sequencing. Online. *Practical Neurology*. 2021, roč. 21, č. 4, s. 322-327. ISSN 1474-7758. Dostupné z: https://doi.org/10.1136/practneurol-2020-002561. [cit. 2024-01-31].

[81] An integrated map of genetic variation from 1,092 human genomes. Online. *Nature*. 2012, roč. 491, č. 7422, s. 56-65. ISSN 0028-0836. Dostupné z: https://doi.org/10.1038/nature11632. [cit. 2024-02-23].

[82] CHOUDHURY, Ananyo; RAMSAY, Michèle; HAZELHURST, Scott; ARON, Shaun; BARDIEN, Soraya et al. Whole-genome sequencing for an enhanced understanding of genetic variation among South Africans. Online. *Nature Communications*. 2017, roč. 8, č. 1, s. 1-11. ISSN 2041-1723. Dostupné z: https://doi.org/10.1038/s41467-017-00663-9. [cit. 2024-02-23].

[83] NAGASAKI, Masao; YASUDA, Jun; KATSUOKA, Fumiki; NARIAI, Naoki; KOJIMA, Kaname et al. Rare variant discovery by deep whole-genome sequencing of 1,070 Japanese individuals. Online. *Nature Communications*. 2015, roč. 6, č. 1, s. 1-13. ISSN 2041-1723. Dostupné z: https://doi.org/10.1038/ncomms9018. [cit. 2024-02-23].

[84] RAZALI, Rozaimi Mohamad; RODRIGUEZ-FLORES, Juan; GHORBANI, Mohammadmersad; NAEEM, Haroon; AAMER, Waleed et al. Thousands of Qatari genomes inform human migration

history and improve imputation of Arab haplotypes. Online. *Nature Communications*. 2021, roč. 12, č. 1, s. 1-16. ISSN 2041-1723. Dostupné z: https://doi.org/10.1038/s41467-021-25287-y. [cit. 2024-02-23].

[85] GENOMEWEB. *European Countries Step up Efforts to Share Genomic Data as Part of 1+ Million Genomes Initiative*. Online. GENOMEWEB. European Countries Step up Efforts to Share Genomic Data as Part of 1+ Million Genomes Initiative. 2024, Feb 15, 2024. Dostupné z: https://www.genomeweb.com/informatics/european-countries-step-efforts-share-genomic-data-part-1-million-genomes-initiative. [cit. 2024-02-23].

[86] GENOMICS ENGLAND. *100,000 Genomes Project*. Online. 2024. Dostupné z: https://www.genomicsengland.co.uk/initiatives/100000-genomes-project. [cit. 2024-02-28].

[87] SOSINSKY, Alona; AMBROSE, John; CROSS, William; TURNBULL, Clare; HENDERSON, Shirley et al. Insights for precision oncology from the integration of genomic and clinical data of 13,880 tumors from the 100,000 Genomes Cancer Programme. Online. *Nature Medicine*. 2024, roč. 30, č. 1, s. 279-289. ISSN 1078-8956. Dostupné z: https://doi.org/10.1038/s41591-023-02682-0. [cit. 2024-02-23].

[88] SIMS, David; SUDBERY, Ian; ILOTT, Nicholas E.; HEGER, Andreas a PONTING, Chris P. Sequencing depth and coverage: key considerations in genomic analyses. Online. *Nature Reviews Genetics*. 2014, roč. 15, č. 2, s. 121-132. ISSN 1471-0056. Dostupné z: https://doi.org/10.1038/nrg3642. [cit. 2024-01-31].

[89] BHAI, Pratibha; TUROWEC, Jacob; SANTOS, Stephanie; KERKHOF, Jennifer; PICKARD, LeeAnne et al. Molecular profiling of solid tumors by next-generation sequencing: an experience from a clinical laboratory. Online. *Frontiers in Oncology*. 2023, roč. 13, s. 1-13. ISSN 2234-943X. Dostupné z: https://doi.org/10.3389/fonc.2023.1208244. [cit. 2024-01-31].

[90] CAO, Ye; TOKITA, Mari J.; CHEN, Edward S.; GHOSH, Rajarshi; CHEN, Tiansheng et al. A clinical survey of mosaic single nucleotide variants in disease-causing genes detected by exome sequencing. Online. *Genome Medicine*. 2019, roč. 11, č. 1, s. 1-11. ISSN 1756-994X. Dostupné z: https://doi.org/10.1186/s13073-019-0658-2. [cit. 2024-02-25].

[91] LIU, Yong-Xin; QIN, Yuan; CHEN, Tong; LU, Meiping; QIAN, Xubo et al. A practical guide to amplicon and metagenomic analysis of microbiome data. Online. *Protein & Cell*. 2021, roč. 12, č. 5, s. 315-330. ISSN 1674-800X. Dostupné z: https://doi.org/10.1007/s13238-020-00724-8. [cit. 2024-01-31].

[92] BERNARDO, Amy; ST. AMAND, Paul; LE, Ha Quang; SU, Zhenqi a BAI, Guihua. Multiplex restriction amplicon sequencing: a novel next-generation sequencing-based marker platform for high-throughput genotyping. Online. *Plant Biotechnology Journal*. 2020, roč. 18, č. 1, s. 254-265. ISSN 1467-7644. Dostupné z: https://doi.org/10.1111/pbi.13192. [cit. 2024-01-31].

[93] IIDA, Midori; SUZUKI, Miyuki; SAKANE, Yuto; NISHIDE, Hiroyo; UCHIYAMA, Ikuo et al. A simple and practical workflow for genotyping of CRISPR–Cas9-based knockout phenotypes using multiplexed amplicon sequencing. Online. *Genes to Cells*. 2020, roč. 25, č. 7, s. 498-509. ISSN 1356-9597. Dostupné z: https://doi.org/10.1111/gtc.12775. [cit. 2024-01-31].

[94] TAYLOR, Mariah K.; WILLIAMS, Evan P.; WONGSURAWAT, Thidathip; JENJAROENPUN, Piroon; NOOKAEW, Intawat et al. Amplicon-Based, Next-Generation Sequencing Approaches to Characterize Single Nucleotide Polymorphisms of Orthohantavirus Species. Online. *Frontiers in Cellular and Infection Microbiology*. 2020, roč. 10, č. 565591, s. 1-18. ISSN 2235-2988. Dostupné z: https://doi.org/10.3389/fcimb.2020.565591. [cit. 2024-01-31].

[95] KLEMPT, Petr; BROŽ, Petr; KAŠNÝ, Martin; NOVOTNÝ, Adam; KVAPILOVÁ, Kateřina et al. Performance of Targeted Library Preparation Solutions for SARS-CoV-2 Whole Genome Analysis. Online. *Diagnostics*. 2020, roč. 10, č. 10, s. 1-12. ISSN 2075-4418. Dostupné z: https://doi.org/10.3390/diagnostics10100769. [cit. 2024-01-31].

[96] OTTESTAD, Anine Larsen; HUANG, Mo; EMDAL, Elisabeth Fritzke; MJELLE, Robin; SKARPETEIG, Veronica et al. Assessment of Two Commercial Comprehensive Gene Panels for Personalized Cancer Treatment. Online. *Journal of Personalized Medicine*. 2023, roč. 13, č. 1, s. 1-12. ISSN 2075-4426. Dostupné z: https://doi.org/10.3390/jpm13010042. [cit. 2024-01-31].

[97] HERNÁNDEZ-NEUTA, Iván; MAGOULOPOULOU, Anastasia; PINEIRO, Flor; LISBY, Jan Gorm; GULBERG, Mats et al. Highly multiplexed targeted sequencing strategy for infectious disease surveillance. Online. *BMC Biotechnology*. 2023, roč. 23, č. 1, s. 1-10. ISSN 1472-6750. Dostupné z: https://doi.org/10.1186/s12896-023-00804-7. [cit. 2024-01-31].

[98] ASPDEN, Julie L.; WALLACE, Edward W.J. a WHIFFIN, Nicola. Not all exons are protein coding: Addressing a common misconception. Online. *Cell Genomics*. 2023, roč. 3, č. 4, s. 1-4. ISSN 2666979X. Dostupné z: https://doi.org/10.1016/j.xgen.2023.100296. [cit. 2024-02-11].

[99] GREEN, Robert C.; BERG, Jonathan S.; GRODY, Wayne W.; KALIA, Sarah S.; KORF, Bruce R. et al. ACMG recommendations for reporting of incidental findings in clinical exome and genome sequencing. Online. *Genetics in Medicine*. 2013, roč. 15, č. 7, s. 565-574. ISSN 10983600. Dostupné z: https://doi.org/10.1038/gim.2013.73. [cit. 2024-01-31].

[100] ALIX, Tom; CHÉRY, Céline; JOSSE, Thomas; BRONOWICKI, Jean-Pierre; FEILLET, François et al. Predictors of the utility of clinical exome sequencing as a first-tier genetic test in patients with Mendelian phenotypes: results from a referral center study on 603 consecutive cases. Online. *Human Genomics*. 2023, roč. 17, č. 1, s. 1-15. ISSN 1479-7364. Dostupné z: https://doi.org/10.1186/s40246-023-00455-x. [cit. 2024-01-31].

[101] GULILAT, Markus; LAMB, Tyler; TEFT, Wendy A.; WANG, Jian; DRON, Jacqueline S. et al. Targeted next generation sequencing as a tool for precision medicine. Online. *BMC Medical Genomics*. 2019, roč. 12, č. 1, s. 1-17. ISSN 1755-8794. Dostupné z: https://doi.org/10.1186/s12920-019-0527-2. [cit. 2024-01-31].

[102] CLEMENT, Kendell; FAROUNI, Rick; BAUER, Daniel E a PINELLO, Luca. AmpUMI: design and analysis of unique molecular identifiers for deep amplicon sequencing. Online. *Bioinformatics*. 2018, roč. 34, č. 13, s. i202-i210. ISSN 1367-4803. Dostupné z: https://doi.org/10.1093/bioinformatics/bty264. [cit. 2024-01-31].

[103] SCHENA, Mark; SHALON, Dari; DAVIS, Ronald W. a BROWN, Patrick O. Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray. Online. *Science*. 1995, roč. 270, č. 5235, s. 467-470. ISSN 0036-8075. Dostupné z: https://doi.org/10.1126/science.270.5235.467. [cit. 2024-02-01].

[104] CLARK, Tyson A; SCHWEITZER, Anthony C; CHEN, Tina X; STAPLES, Michelle K; LU, Gang et al. Discovery of tissue-specific exons using comprehensive human exon microarrays. Online. *Genome Biology*. 2007, roč. 8, č. 4, s. 1-16. ISSN 14656906. Dostupné z: https://doi.org/10.1186/gb-2007-8-4-r64. [cit. 2024-02-01].

[105] HRDLICKOVA, Radmila; TOLOUE, Masoud a TIAN, Bin. RNA -Seq methods for transcriptome analysis. Online. *WIREs RNA*. 2017, roč. 8, č. 1. ISSN 1757-7004. Dostupné z: https://doi.org/10.1002/wrna.1364. [cit. 2024-01-20].

[106] WONGSURAWAT, Thidathip; JENJAROENPUN, Piroon a NOOKAEW, Intawat. Direct Sequencing of RNA and RNA Modification Identification Using Nanopore. Online. In: DEVAUX, Frédéric (ed.). *Yeast Functional Genomics*. 1. Methods in Molecular Biology. New York, NY: Springer US, 2022, s. 71-77. ISBN 978-1-0716-2256-8. Dostupné z: https://doi.org/10.1007/978-1-0716-2257-5_5. [cit. 2024-02-01].

[107] FENG, Shouli; XU, Min; LIU, Fujie; CUI, Changjiang a ZHOU, Baoliang. Reconstruction of the full-length transcriptome atlas using PacBio Iso-Seq provides insight into the alternative splicing in

Gossypium australe. Online. *BMC Plant Biology*. 2019, roč. 19, č. 1, s. 1-16. ISSN 1471-2229. Dostupné z: https://doi.org/10.1186/s12870-019-1968-7. [cit. 2024-02-01].

[108] EPHRAIM, Ramya; FRASER, Sarah; DEVEREAUX, Jeannie; STAVELY, Rhian; FEEHAN, Jack et al. Differential Gene Expression of Checkpoint Markers and Cancer Markers in Mouse Models of Spontaneous Chronic Colitis. Online. *Cancers*. 2023, roč. 15, č. 19, s. 2-21. ISSN 2072-6694. Dostupné z: https://doi.org/10.3390/cancers15194793. [cit. 2024-02-01].

[109] ROBERTS, Aedan G K; CATCHPOOLE, Daniel R a KENNEDY, Paul J. Identification of differentially distributed gene expression and distinct sets of cancer-related genes identified by changes in mean and variability. Online. *NAR Genomics and Bioinformatics*. 2022, roč. 4, č. 1, s. 1-14. ISSN 2631-9268. Dostupné z: https://doi.org/10.1093/nargab/lqab124. [cit. 2024-02-01].

[110] HAQUE, Ashraful; ENGEL, Jessica; TEICHMANN, Sarah A. a LÖNNBERG, Tapio. A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. Online. *Genome Medicine*. 2017, roč. 9, č. 1, s. 1-12. ISSN 1756-994X. Dostupné z: https://doi.org/10.1186/s13073-017-0467-4. [cit. 2024-02-01].

[111] PALAZZO, Alexander F.; GREGORY, T. Ryan a AKEY, Joshua M. The Case for Junk DNA. Online. *PLoS Genetics*. 2014, roč. 10, č. 5, s. 1-8. ISSN 1553-7404. Dostupné z: https://doi.org/10.1371/journal.pgen.1004351. [cit. 2024-02-01].

[112] MARCHAND, Virginie; PICHOT, Florian; THÜRING, Kathrin; AYADI, Lilia; FREUND, Isabel et al. Next-Generation Sequencing-Based RiboMethSeq Protocol for Analysis of tRNA 2′-O-Methylation. Online. *Biomolecules*. 2017, roč. 7, č. 4, s. 2-21. ISSN 2218-273X. Dostupné z: https://doi.org/10.3390/biom7010013. [cit. 2024-02-01].

[113] LUCAS, Morghan C.; PRYSZCZ, Leszek P.; MEDINA, Rebeca; MILENKOVIC, Ivan; CAMACHO, Noelia et al. Quantitative analysis of tRNA abundance and modifications by nanopore RNA sequencing. Online. *Nature Biotechnology*. 2024, roč. 42, č. 1, s. 72-86. ISSN 1087-0156. Dostupné z: https://doi.org/10.1038/s41587-023-01743-6. [cit. 2024-02-01].

[114] FLURIN, Laure; HEMENWAY, Joseph J.; FISHER, Cody R.; VAILLANT, James J.; AZAD, Marisa et al. Clinical Use of a 16S Ribosomal RNA Gene-Based Sanger and/or Next Generation Sequencing Assay to Test Preoperative Synovial Fluid for Periprosthetic Joint Infection Diagnosis. Online. *MBio*. 2022, roč. 13, č. 6, s. e01322-22. ISSN 2150-7511. Dostupné z: https://doi.org/10.1128/mbio.01322-22. [cit. 2024-02-01].

[115] WANG, Jin; CHEN, Jinyun a SEN, Subrata. MicroRNA as Biomarkers and Diagnostics. Online. *Journal of Cellular Physiology*. 2016, roč. 231, č. 1, s. 25-30. ISSN 0021-9541. Dostupné z: https://doi.org/10.1002/jcp.25056. [cit. 2024-02-01].

[116] CHAKRABORTY, Chiranjib; SHARMA, Ashish Ranjan; SHARMA, Garima; DOSS, C. George Priya a LEE, Sang-Soo. Therapeutic miRNA and siRNA: Moving from Bench to Clinic as Next Generation Medicine. Online. *Molecular Therapy - Nucleic Acids*. 2017, roč. 8, č. 1, s. 132-143. ISSN 21622531. Dostupné z: https://doi.org/10.1016/j.omtn.2017.06.005. [cit. 2024-02-01].

[117] DARD-DASCOT, Cloelia; NAQUIN, Delphine; D'AUBENTON-CARAFA, Yves; ALIX, Karine; THERMES, Claude et al. Systematic comparison of small RNA library preparation protocols for next-generation sequencing. Online. *BMC Genomics*. 2018, roč. 19, č. 1, s. 1-16. ISSN 1471-2164. Dostupné z: https://doi.org/10.1186/s12864-018-4491-6. [cit. 2024-02-01].

[118] JIA, Shanshan; ZHANG, Qiang; WANG, Yu; WANG, Yanfu; LIU, Dan et al. PIWI-interacting RNA sequencing profiles in maternal plasma-derived exosomes reveal novel non-invasive prenatal biomarkers for the early diagnosis of nonsyndromic cleft lip and palate. Online. *EBioMedicine*. 2021, roč. 65, č. 1, s. 1-15. ISSN 23523964. Dostupné z: https://doi.org/10.1016/j.ebiom.2021.103253. [cit. 2024-02-01].

[119] YAMADA, Atsushi; YU, Pingjian; LIN, Wei; OKUGAWA, Yoshinaga; BOLAND, C. Richard et al. A RNA-Sequencing approach for the identification of novel long non-coding RNA biomarkers in colorectal cancer. Online. *Scientific Reports*. 2018, roč. 8, č. 1, s. 1-10. ISSN 2045-2322. Dostupné z: https://doi.org/10.1038/s41598-017-18407-6. [cit. 2024-02-01].

[120] BEYLERLI, Ozal; GAREEV, Ilgiz; SUFIANOV, Albert; ILYASOVA, Tatiana a GUANG, Yang. Long noncoding RNAs as promising biomarkers in cancer. Online. *Non-coding RNA Research*. 2022, roč. 7, č. 2, s. 66-70. ISSN 24680540. Dostupné z: https://doi.org/10.1016/j.ncrna.2022.02.004. [cit. 2024-02-01].

[121] SALZBERG, Steven L. Next-generation genome annotation: we still struggle to get it right. Online. *Genome Biology*. 2019, roč. 20, č. 1, s. 1-3. ISSN 1474-760X. Dostupné z: https://doi.org/10.1186/s13059-019-1715-2. [cit. 2024-02-18].

[122] BARTOLOMEI, Marisa S.; OAKEY, Rebecca J. a WUTZ, Anton. Genomic imprinting: An epigenetic regulatory system. Online. *PLOS Genetics*. 2020, roč. 16, č. 8, s. 1-3. ISSN 1553-7404. Dostupné z: https://doi.org/10.1371/journal.pgen.1008970. [cit. 2024-02-12].

[123] LAWLOR, Matthew A a ELLISON, Christopher E. Evolutionary dynamics between transposable elements and their host genomes: mechanisms of suppression and escape. Online. *Current Opinion in*

Genetics & Development. 2023, roč. 82, č. 1, s. 1-9. ISSN 0959437X. Dostupné z: https://doi.org/10.1016/j.gde.2023.102092. [cit. 2024-02-12].

[124] LI, Jiaqi; LI, Lifang; WANG, Yimeng; HUANG, Gan; LI, Xia et al. Insights Into the Role of DNA Methylation in Immune Cell Development and Autoimmune Disease. Online. *Frontiers in Cell and Developmental Biology*. 2021, roč. 9, č. 1, s. 1-13. ISSN 2296-634X. Dostupné z: https://doi.org/10.3389/fcell.2021.757318. [cit. 2024-02-12].

[125] BARRES, Romain a ZIERATH, Juleen R. DNA methylation in metabolic disorders. Online. *The American Journal of Clinical Nutrition*. 2011, roč. 93, č. 4, s. 897S-900S. ISSN 00029165. Dostupné z: https://doi.org/10.3945/ajcn.110.001933. [cit. 2024-02-12].

[126] BESSELINK, Nicolle; KEIJER, Janneke; VERMEULEN, Carlo; BOYMANS, Sander; DE RIDDER, Jeroen et al. The genome-wide mutational consequences of DNA hypomethylation. Online. *Scientific Reports*. 2023, roč. 13, č. 1, s. 1-12. ISSN 2045-2322. Dostupné z: https://doi.org/10.1038/s41598-023-33932-3. [cit. 2024-02-15].

[127] EHRLICH, Melanie. DNA hypermethylation in disease: mechanisms and clinical relevance. Online. *Epigenetics*. 2019, roč. 14, č. 12, s. 1141-1163. ISSN 1559-2294. Dostupné z: https://doi.org/10.1080/15592294.2019.1638701. [cit. 2024-02-15].

[128] LI, Yuanyuan a TOLLEFSBOL, Trygve O. DNA Methylation Detection: Bisulfite Genomic Sequencing Analysis. Online. In: TOLLEFSBOL, Trygve O. (ed.). *Epigenetics Protocols*. 2011. Methods in Molecular Biology. Totowa, NJ: Humana Press, 2011, s. 11-21. ISBN 978-1-61779-315-8. Dostupné z: https://doi.org/10.1007/978-1-61779-316-5_2. [cit. 2024-01-31].

[129] BLEWITT, Marnie; GOUIL, Quentin a KENIRY, Andrew. Latest techniques to study DNA methylation. Online. *Essays in Biochemistry*. 2019, roč. 63, č. 6, s. 639-648. ISSN 0071-1365. Dostupné z: https://doi.org/10.1042/EBC20190027. [cit. 2024-01-31].

[130] LI, Ning; YE, Mingzhi; LI, Yingrui; YAN, Zhixiang; BUTCHER, Lee M. et al. Whole genome DNA methylation analysis based on high throughput sequencing technology. Online. *Methods*. 2010, roč. 52, č. 3, s. 203-212. ISSN 10462023. Dostupné z: https://doi.org/10.1016/j.ymeth.2010.04.009. [cit. 2024-01-31].

[131] HE, Wanhong; SUN, Yuhua; ZHANG, Sufen; FENG, Xing; XU, Minjie et al. Profiling the DNA methylation patterns of imprinted genes in abnormal semen samples by next-generation bisulfite sequencing. Online. *Journal of Assisted Reproduction and Genetics*. 2020, roč. 37, č. 9, s. 2211-2221. ISSN 1058-0468. Dostupné z: https://doi.org/10.1007/s10815-020-01839-x. [cit. 2024-01-31].

[132] FATEMI, Nayeralsadat; TIERLING, Sascha; ES, Hamidreza Aboulkheyr; VARKIANI, Maryam; MOJARAD, Ehsan Nazemalhosseini et al. DNA methylation biomarkers in colorectal cancer: Clinical applications for precision medicine. Online. *International Journal of Cancer*. 2022, roč. 151, č. 12, s. 2068-2081. ISSN 0020-7136. Dostupné z: https://doi.org/10.1002/ijc.34186. [cit. 2024-01-31].

[133] PARKINSON, Nicholas J.; MASLAU, Siarhei; FERNEYHOUGH, Ben; ZHANG, Gang; GREGORY, Lorna et al. Preparation of high-quality next-generation sequencing libraries from picogram quantities of target DNA. Online. *Genome Research*. 2012, roč. 22, č. 1, s. 125-133. ISSN 1088-9051. Dostupné z: https://doi.org/10.1101/gr.124016.111. [cit. 2024-02-02].

[134] HESS, J.F.; KOHL, T.A.; KOTROVÁ, M.; RÖNSCH, K.; PAPROTKA, T. et al. Library preparation for next generation sequencing: A review of automation strategies. Online. *Biotechnology Advances*. 2020, roč. 41, č. 1, s. 1-14. ISSN 07349750. Dostupné z: https://doi.org/10.1016/j.biotechadv.2020.107537. [cit. 2024-02-02].

[135] HEAD, Steven R.; KOMORI, H. Kiyomi; LAMERE, Sarah A.; WHISENANT, Thomas; VAN NIEUWERBURGH, Filip et al. Library construction for next-generation sequencing: Overviews and challenges. Online. *BioTechniques*. 2014, roč. 56, č. 2, s. 61-77. ISSN 0736-6205. Dostupné z: https://doi.org/10.2144/000114133. [cit. 2024-02-02].

[136] VERWILT, Jasper; MESTDAGH, Pieter a VANDESOMPELE, Jo. Artifacts and biases of the reverse transcription reaction in RNA sequencing. Online. *RNA*. 2023, roč. 29, č. 7, s. 889-897. ISSN 1355-8382. Dostupné z: https://doi.org/10.1261/rna.079623.123. [cit. 2024-02-02].

[137] PUCHTA, Marta; BOCZKOWSKA, Maja a GROSZYK, Jolanta. Low RIN Value for RNA-Seq Library Construction from Long-Term Stored Seeds: A Case Study of Barley Seeds. Online. *Genes*. 2020, roč. 11, č. 10, s. 1-15. ISSN 2073-4425. Dostupné z: https://doi.org/10.3390/genes11101190. [cit. 2024-02-12].

[138] AIRD, Daniel; ROSS, Michael G; CHEN, Wei-Sheng; DANIELSSON, Maxwell; FENNELL, Timothy et al. Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. Online. *Genome Biology*. 2011, roč. 12, č. 2, s. 1-14. ISSN 1465-6906. Dostupné z: https://doi.org/10.1186/gb-2011-12-2-r18. [cit. 2024-02-12].

[139] BROWNE, Patrick Denis; NIELSEN, Tue Kjærgaard; KOT, Witold; AGGERHOLM, Anni; GILBERT, M Thomas P et al. GC bias affects genomic and metagenomic reconstructions, underrepresenting GC-poor organisms. Online. *GigaScience*. 2020, roč. 9, č. 2, s. 1-14. ISSN 2047-217X. Dostupné z: https://doi.org/10.1093/gigascience/giaa008. [cit. 2024-02-12].

[140] BIRNEY, Ewan. The International Human Genome Project. Online. *Human Molecular Genetics*. 2021, roč. 30, č. R2, s. R161-R163. ISSN 0964-6906. Dostupné z: https://doi.org/10.1093/hmg/ddab198. [cit. 2024-02-04].

[141] BENSON, D. A.; KARSCH-MIZRACHI, I.; LIPMAN, D. J.; OSTELL, J. a WHEELER, D. L. GenBank. Online. *Nucleic Acids Research*. 2007, roč. 36, č. Database, s. D25-D30. ISSN 0305-1048. Dostupné z: https://doi.org/10.1093/nar/gkm929. [cit. 2024-02-05].

[142] BENTON, David. Recent changes in the GenBank ® On-line Service. Online. *Nucleic Acids Research*. 1990, roč. 18, č. 6, s. 1517-1520. ISSN 0305-1048. Dostupné z: https://doi.org/10.1093/nar/18.6.1517. [cit. 2024-02-05].

[143] SHERRY, S. T. DbSNP: the NCBI database of genetic variation. Online. *Nucleic Acids Research*. 2001, roč. 29, č. 1, s. 308-311. ISSN 13624962. Dostupné z: https://doi.org/10.1093/nar/29.1.308. [cit. 2024-02-05].

[144] MEETING, Notes from the a GUYER, statement compiled by Mark. Statement on the Rapid Release of Genomic DNA Sequence. Online. *Genome Research*. 1998, roč. 8, č. 5, s. 413-413. ISSN 1088-9051. Dostupné z: https://doi.org/10.1101/gr.8.5.413. [cit. 2024-02-04].

[145] LI, Shuyu; CUTLER, Gene; LIU, Jane Jijun; HOEY, Timothy; CHEN, Liangbiao et al. A comparative analysis of HGSC and Celera human genome assemblies and gene sets. Online. *Bioinformatics*. 2003, roč. 19, č. 13, s. 1597-1605. ISSN 1367-4811. Dostupné z: https://doi.org/10.1093/bioinformatics/btg219. [cit. 2024-02-04].

[146] The International HapMap Project. Online. *Nature*. 2003, roč. 426, č. 6968, s. 789-796. ISSN 0028-0836. Dostupné z: https://doi.org/10.1038/nature02168. [cit. 2024-02-05].

[147] SCHNEIDER, Valerie A.; GRAVES-LINDSAY, Tina; HOWE, Kerstin; BOUK, Nathan; CHEN, Hsiu-Chuan et al. Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. Online. *Genome Research*. 2017, roč. 27, č. 5, s. 849-864. ISSN 1088-9051. Dostupné z: https://doi.org/10.1101/gr.213611.116. [cit. 2024-02-05].

[148] NURK, Sergey; KOREN, Sergey; RHIE, Arang; RAUTIAINEN, Mikko a BZIKADZE, Andrey V. The complete sequence of a human genome. Online. *Science*. 2022, roč. 376, č. 6588, s. 44-53. ISSN 0036-8075. Dostupné z: https://doi.org/10.1126/science.abj6987. [cit. 2024-02-06].

[149] BAYAT, A. Science, medicine, and the future: Bioinformatics. Online. *BMJ*. 2002, roč. 324, č. 7344, s. 1018-1022. ISSN 09598138. Dostupné z: https://doi.org/10.1136/bmj.324.7344.1018. [cit. 2024-02-26].

[150] DEL FABBRO, Cristian; SCALABRIN, Simone; MORGANTE, Michele; GIORGI, Federico M. a SEO, Jeong-Sun. An Extensive Evaluation of Read Trimming Effects on Illumina NGS Data Analysis. Online. *PLoS ONE*. 2013, roč. 8, č. 12, s. 1-12. ISSN 1932-6203. Dostupné z: https://doi.org/10.1371/journal.pone.0085024. [cit. 2024-02-18].

[151] WINGETT, Steven W. a ANDREWS, Simon. FastQ Screen: A tool for multi-genome mapping and quality control. Online. *F1000Research*. 2018, roč. 7. ISSN 2046-1402. Dostupné z: https://doi.org/10.12688/f1000research.15931.2. [cit. 2024-01-22].

[152] GONDANE, Aishwarya a ITKONEN, Harri M. Revealing the History and Mystery of RNA-Seq. Online. *Current Issues in Molecular Biology*. 2023, roč. 45, č. 3, s. 1860-1874. ISSN 1467-3045. Dostupné z: https://doi.org/10.3390/cimb45030120. [cit. 2024-02-18].

[153] PEREIRA, Rute; OLIVEIRA, Jorge a SOUSA, Mário. Bioinformatics and Computational Tools for Next-Generation Sequencing Analysis in Clinical Genetics. Online. *Journal of Clinical Medicine*. 2020, roč. 9, č. 1, s. 1-30. ISSN 2077-0383. Dostupné z: https://doi.org/10.3390/jcm9010132. [cit. 2024-02-18].

[154] SONG, Jinming a HUSSAINI, Mohammad. Adopting solutions for annotation and reporting of next generation sequencing in clinical practice. Online. *Practical Laboratory Medicine*. 2020, roč. 19, č. 1, s. 1-9. ISSN 23525517. Dostupné z: https://doi.org/10.1016/j.plabm.2020.e00154. [cit. 2024-02-18].

[155] DIAS, Raquel a TORKAMANI, Ali. Artificial intelligence in clinical and genomic diagnostics. Online. *Genome Medicine*. 2019, roč. 11, č. 1, s. 1-12. ISSN 1756-994X. Dostupné z: https://doi.org/10.1186/s13073-019-0689-8. [cit. 2024-03-07].

[156] ARADHYA, Swaroop; FACIO, Flavia M.; METZ, Hillery; MANDERS, Toby; COLAVIN, Alexandre et al. Applications of artificial intelligence in clinical laboratory genomics. Online. *American Journal of Medical Genetics Part C: Seminars in Medical Genetics*. 2023, roč. 193, č. 3, s. 1-15. ISSN 1552-4868. Dostupné z: https://doi.org/10.1002/ajmg.c.32057. [cit. 2024-03-07].

[157] MILLER, David T.; LEE, Kristy; GORDON, Adam S.; AMENDOLA, Laura M.; ADELMAN, Kathy et al. Recommendations for reporting of secondary findings in clinical exome and genome sequencing, 2021 update: a policy statement of the American College of Medical Genetics and Genomics (ACMG). Online. *Genetics in Medicine*. 2021, roč. 23, č. 8, s. 1391-1398. ISSN 10983600. Dostupné z: https://doi.org/10.1038/s41436-021-01171-4. [cit. 2024-02-18].

[158] HOUGE, Gunnar; LANER, Andreas; CIRAK, Sebahattin; DE LEEUW, Nicole; SCHEFFER, Hans et al. Stepwise ABC system for classification of any type of genetic variant. Online. *European*

Journal of Human Genetics. 2022, roč. 30, č. 2, s. 150-159. ISSN 1018-4813. Dostupné z: https://doi.org/10.1038/s41431-021-00903-z. [cit. 2024-02-18].

[159] EUROPEAN SOCIETY FOR MEDICAL ONCOLOGY. *Guidelines by topic*. Online. EUROPEAN SOCIETY FOR MEDICAL ONCOLOGY. Guidelines by topic. 2024. Dostupné z: https://www.esmo.org/guidelines/guidelines-by-topic. [cit. 2024-02-28].

[160] LI, Marilyn M.; DATTO, Michael; DUNCAVAGE, Eric J.; KULKARNI, Shashikant; LINDEMAN, Neal I. et al. Standards and Guidelines for the Interpretation and Reporting of Sequence Variants in Cancer. Online. *The Journal of Molecular Diagnostics*. 2017, roč. 19, č. 1, s. 4-23. ISSN 15251578. Dostupné z: https://doi.org/10.1016/j.jmoldx.2016.10.002. [cit. 2024-02-28].

[161] MAIDEN, Martin C. J.; BYGRAVES, Jane A.; FEIL, Edward; MORELLI, Giovanna; RUSSELL, Joanne E. et al. Multilocus sequence typing: A portable approach to the identification of clones within populations of pathogenic microorganisms. Online. *Proceedings of the National Academy of Sciences*. 1998, roč. 95, č. 6, s. 3140-3145. ISSN 0027-8424. Dostupné z: https://doi.org/10.1073/pnas.95.6.3140. [cit. 2024-02-07].

[162] GWINN, Marta; MACCANNELL, Duncan a ARMSTRONG, Gregory L. Next-Generation Sequencing of Infectious Pathogens. Online. *JAMA*. 2019, roč. 321, č. 9, s. 1-7. ISSN 0098-7484. Dostupné z: https://doi.org/10.1001/jama.2018.21669. [cit. 2024-02-07].

[163] GRAF, Erin. The Emergence of Pathogen Genomics in Diagnostic Laboratories. Online. *American Society for Clinical Laboratory Science*. Oct 2019, roč. 35, č. 1, s. 1-8. ISSN 0894-959X. Dostupné z: https://doi.org/10.29074/ascls.119.001776. [cit. 2024-02-08].

[164] SIMNER, Patricia J; MILLER, Steven a CARROLL, Karen C. Understanding the Promises and Hurdles of Metagenomic Next-Generation Sequencing as a Diagnostic Tool for Infectious Diseases. Online. *Clinical Infectious Diseases*. 2018, roč. 66, č. 5, s. 778-788. ISSN 1058-4838. Dostupné z: https://doi.org/10.1093/cid/cix881. [cit. 2024-02-07].

[165] LAW, Jodi Woan-Fei; AB MUTALIB, Nurul-Syakima; CHAN, Kok-Gan a LEE, Learn-Han. Rapid methods for the detection of foodborne bacterial pathogens: principles, applications, advantages and limitations. Online. *Frontiers in Microbiology*. 2015, roč. 5, č. 3, s. 1-23. ISSN 1664-302X. Dostupné z: https://doi.org/10.3389/fmicb.2014.00770. [cit. 2024-02-25].

[166] EXUM, Natalie G.; PISANIC, Nora; GRANGER, Douglas A.; SCHWAB, Kellogg J.; DETRICK, Barbara et al. Use of Pathogen-Specific Antibody Biomarkers to Estimate Waterborne Infections in

Population-Based Settings. Online. *Current Environmental Health Reports*. 2016, roč. 3, č. 3, s. 322-334. ISSN 2196-5412. Dostupné z: https://doi.org/10.1007/s40572-016-0096-x. [cit. 2024-02-25].

[167] CHIU, Charles Y. a MILLER, Steven A. Clinical metagenomics. Online. *Nature Reviews Genetics*. 2019, roč. 20, č. 6, s. 341-355. ISSN 1471-0056. Dostupné z: https://doi.org/10.1038/s41576-019-0113-7. [cit. 2024-02-08].

[168] BESSER, J.; CARLETON, H.A.; GERNER-SMIDT, P.; LINDSEY, R.L. a TREES, E. Next-generation sequencing technologies and their application to the study and control of bacterial infections. Online. *Clinical Microbiology and Infection*. 2018, roč. 24, č. 4, s. 335-341. ISSN 1198743X. Dostupné z: https://doi.org/10.1016/j.cmi.2017.10.013. [cit. 2024-02-08].

[169] PASIK, Katarzyna a DOMAŃSKA-BLICHARZ, Katarzyna. High-throughput sequencing in vaccine research. Online. *Journal of Veterinary Research*. 2021, roč. 65, č. 2, s. 131-137. ISSN 2450-8608. Dostupné z: https://doi.org/10.2478/jvetres-2021-0029. [cit. 2024-02-08].

[170] QUER, Josep; COLOMER-CASTELL, Sergi; CAMPOS, Carolina; ANDRÉS, Cristina; PIÑANA, Maria et al. Next-Generation Sequencing for Confronting Virus Pandemics. Online. *Viruses*. 2022, roč. 14, č. 3, s. 1-23. ISSN 1999-4915. Dostupné z: https://doi.org/10.3390/v14030600. [cit. 2024-02-08].

[171] BERBERS, Bas; CEYSSENS, Pieter-Jan; BOGAERTS, Pierre; VANNESTE, Kevin; ROOSENS, Nancy H. C. et al. Development of an NGS-Based Workflow for Improved Monitoring of Circulating Plasmids in Support of Risk Assessment of Antimicrobial Resistance Gene Dissemination. Online. *Antibiotics*. 2020, roč. 9, č. 8, s. 1-29. ISSN 2079-6382. Dostupné z: https://doi.org/10.3390/antibiotics9080503. [cit. 2024-02-08].

[172] WHEELER, Nicole E; PRICE, Vivien; CUNNINGHAM-OAKES, Edward; TSANG, Kara K; NUNN, Jamie G et al. Innovations in genomic antimicrobial resistance surveillance. Online. *The Lancet Microbe*. 2023, roč. 4, č. 12, s. e1063-e1070. ISSN 26665247. Dostupné z: https://doi.org/10.1016/S2666-5247(23)00285-9. [cit. 2024-02-08].

[173] WU, Fan; ZHAO, Su; YU, Bin; CHEN, Yan-Mei; WANG, Wen et al. A new coronavirus associated with human respiratory disease in China. Online. *Nature*. 2020, roč. 579, č. 7798, s. 265-269. ISSN 0028-0836. Dostupné z: https://doi.org/10.1038/s41586-020-2008-3. [cit. 2024-03-04].

[174] WU, Fan; ZHAO, Su; YU, Bin; CHEN, Yan-Mei; WANG, Wen et al. A new coronavirus associated with human respiratory disease in China. Online. *Nature*. 2020, roč. 579, č. 7798, s. 265-269. ISSN 0028-0836. Dostupné z: https://doi.org/10.1038/s41586-020-2008-3. [cit. 2024-03-04].

[175] DUFFY, Siobain. Why are RNA virus mutation rates so damn high? Online. *PLOS Biology*. 2018, roč. 16, č. 8, s. 1-6. ISSN 1545-7885. Dostupné z: https://doi.org/10.1371/journal.pbio.3000003. [cit. 2024-03-04].

[176] KLEMPT, Petr; BRZOŇ, Ondřej; KAŠNÝ, Martin; KVAPILOVÁ, Kateřina; HUBÁČEK, Petr et al. Distribution of SARS-CoV-2 Lineages in the Czech Republic, Analysis of Data from the First Year of the Pandemic. Online. *Microorganisms*. 2021, roč. 9, č. 8, s. 1-12. ISSN 2076-2607. Dostupné z: https://doi.org/10.3390/microorganisms9081671. [cit. 2024-02-26].

[177] OOSTLANDER, AE; MEIJER, GA a YLSTRA, B. Microarray-based comparative genomic hybridization and its applications in human genetics. Online. *Clinical Genetics*. 2004, roč. 66, č. 6, s. 488-495. ISSN 0009-9163. Dostupné z: https://doi.org/10.1111/j.1399-0004.2004.00322.x. [cit. 2024-02-26].

[178] WITSCH-BAUMGARTNER, Martina; SCHWANINGER, Gunda; SCHNAITER, Simon; KOLLMANN, Franziska; BURKHARD, Silja et al. Array genotyping as diagnostic approach in medical genetics. Online. *Molecular Genetics & Genomic Medicine*. 2022, roč. 10, č. 9, s. 1-11. ISSN 2324-9269. Dostupné z: https://doi.org/10.1002/mgg3.2016. [cit. 2024-02-20].

[179] BHÉRER, Claude; EVELEIGH, Robert; TRAJANOSKA, Katerina; ST-CYR, Janick; PACCARD, Antoine et al. A cost-effective sequencing method for genetic studies combining high-depth whole exome and low-depth whole genome. Online. *Npj Genomic Medicine*. 2024, roč. 9, č. 1, s. 1-12. ISSN 2056-7944. Dostupné z: https://doi.org/10.1038/s41525-024-00390-3. [cit. 2024-02-20].

[180] EISENSTEIN, Michael. Super-speedy sequencing puts genomic diagnosis in the fast lane. Online. *Nature*. 2024, roč. 626, č. 8000, s. 915-917. ISSN 0028-0836. Dostupné z: https://doi.org/10.1038/d41586-024-00483-0. [cit. 2024-02-26].

[181] SCHWARZ, Ute I.; GULILAT, Markus a KIM, Richard B. The Role of Next-Generation Sequencing in Pharmacogenetics and Pharmacogenomics. Online. *Cold Spring Harbor Perspectives in Medicine*. 2019, roč. 9, č. 2, s. 1-15. ISSN 2157-1422. Dostupné z: https://doi.org/10.1101/cshperspect.a033027. [cit. 2024-02-28].

[182] DRUKER, Brian J. Translation of the Philadelphia chromosome into therapy for CML. Online. *Blood*. 2008, roč. 112, č. 13, s. 4808-4817. ISSN 0006-4971. Dostupné z: https://doi.org/10.1182/blood-2008-07-077958. [cit. 2024-02-20].

[183] DAHUI, Qin. Next-generation sequencing and its clinical application. Online. *Cancer Biology & Medicine*. 2019, roč. 16, č. 1, s. 4-10. ISSN 20953941. Dostupné z: https://doi.org/10.20892/j.issn.2095-3941.2018.0055. [cit. 2024-02-20].

[184] JAYASHANKAR, Siva Shantini; NASARUDDIN, Muhammad Luqman; HASSAN, Muhammad Faiz; DASRILSYAH, Rima Anggrena; SHAFIEE, Mohamad Nasir et al. Non-Invasive Prenatal Testing (NIPT): Reliability, Challenges, and Future Directions. Online. *Diagnostics*. 2023, roč. 13, č. 15, s. 1-21. ISSN 2075-4418. Dostupné z: https://doi.org/10.3390/diagnostics13152570. [cit. 2024-02-26].

[185] ZHEN, Li; LI, Yu-Juan; YANG, Yan-Dong a LI, Dong-Zhi. The role of ultrasound in women with a positive NIPT result for trisomy 18 and 13. Online. *Taiwanese Journal of Obstetrics and Gynecology*. 2019, roč. 58, č. 6, s. 798-800. ISSN 10284559. Dostupné z: https://doi.org/10.1016/j.tjog.2019.09.012. [cit. 2024-02-26].

[186] REHDER, Catherine; BEAN, Lora J.H.; BICK, David; CHAO, Elizabeth; CHUNG, Wendy et al. Next-generation sequencing for constitutional variants in the clinical laboratory, 2021 revision: a technical standard of the American College of Medical Genetics and Genomics (ACMG). Online. *Genetics in Medicine*. 2021, roč. 23, č. 8, s. 1399-1415. ISSN 10983600. Dostupné z: https://doi.org/10.1038/s41436-021-01139-4. [cit. 2024-02-20].

[187] MATTHIJS, Gert; SOUCHE, Erika; ALDERS, Mariëlle; CORVELEYN, Anniek; ECK, Sebastian et al. Guidelines for diagnostic next-generation sequencing. Online. *European Journal of Human Genetics*. 2016, roč. 24, č. 1, s. 2-5. ISSN 1018-4813. Dostupné z: https://doi.org/10.1038/ejhg.2015.226. [cit. 2024-02-20].

[188] HORGAN, Denis; CURIGLIANO, Giuseppe; RIESS, Olaf; HOFMAN, Paul; BÜTTNER, Reinhard et al. Identifying the Steps Required to Effectively Implement Next-Generation Sequencing in Oncology at a National Level in Europe. Online. *Journal of Personalized Medicine*. 2022, roč. 12, č. 1, s. 1-27. ISSN 2075-4426. Dostupné z: https://doi.org/10.3390/jpm12010072. [cit. 2024-02-20].

[189] JENNINGS, Lawrence J.; ARCILA, Maria E.; CORLESS, Christopher; KAMEL-REID, Suzanne; LUBIN, Ira M. et al. Guidelines for Validation of Next-Generation Sequencing–Based Oncology Panels. Online. *The Journal of Molecular Diagnostics*. 2017, roč. 19, č. 3, s. 341-365. ISSN 15251578. Dostupné z: https://doi.org/10.1016/j.jmoldx.2017.01.011. [cit. 2024-02-20].

[190] GUDISEVA, Harini V.; HANSEN, Mark; GUTIERREZ, Linda; COLLINS, David W.; HE, Jie et al. Saliva DNA quality and genotyping efficiency in a predominantly elderly population. Online. *BMC*

Medical Genomics. 2016, roč. 9, č. 1, s. 1-8. ISSN 1755-8794. Dostupné z: https://doi.org/10.1186/s12920-016-0172-y. [cit. 2024-03-22].

[191] BRUINSMA, Fiona J.; JOO, Jihoon E.; WONG, Ee Ming; GILES, Graham G. a SOUTHEY, Melissa C. The utility of DNA extracted from saliva for genome-wide molecular research platforms. Online. *BMC Research Notes*. 2018, roč. 11, č. 1, s. 1-6. ISSN 1756-0500. Dostupné z: https://doi.org/10.1186/s13104-017-3110-y. [cit. 2024-03-22].

[192] KIDD, Jeffrey M; SHARPTON, Thomas J; BOBO, Dean; NORMAN, Paul J; MARTIN, Alicia R et al. Exome capture from saliva produces high quality genomic and metagenomic data. Online. *BMC Genomics*. 2014, roč. 15, č. 1, s. 1-17. ISSN 1471-2164. Dostupné z: https://doi.org/10.1186/1471-2164-15-262. [cit. 2024-03-22].

[193] TROST, Brett; WALKER, Susan; HAIDER, Syed A; SUNG, Wilson W L; PEREIRA, Sergio et al. Impact of DNA source on genetic variant detection from human whole-genome sequencing data. Online. *Journal of Medical Genetics*. 2019, roč. 56, č. 12, s. 809-817. ISSN 0022-2593. Dostupné z: https://doi.org/10.1136/jmedgenet-2019-106281. [cit. 2024-03-11].

[194] *Bcl2fastq*. Online. 2024. Dostupné z: https://support.illumina.com/sequencing/sequencing_software/bcl2fastq-conversion-software.html. [cit. 2024-01-23].

[195] CHEN, Shifu; ZHOU, Yanqing; CHEN, Yaru a GU, Jia. Fastp: an ultra-fast all-in-one FASTQ preprocessor. Online. *Bioinformatics*. 2018, roč. 34, č. 17, s. i884-i890. ISSN 1367-4803. Dostupné z: https://doi.org/10.1093/bioinformatics/bty560. [cit. 2024-01-22].

[196] LANGMEAD, Ben a SALZBERG, Steven L. Fast gapped-read alignment with Bowtie 2. Online. *Nature Methods*. 2012, roč. 9, č. 4, s. 357-359. ISSN 1548-7091. Dostupné z: https://doi.org/10.1038/nmeth.1923. [cit. 2024-01-23].

[197] BROAD INSTITUTE. *Picard tools*. Online. 2024. Dostupné z: https://broadinstitute.github.io/picard/. [cit. 2024-01-23].

[198] Haplotype-based variant detection from short-read sequencing. *ArXiv: 1207.3907v2*. July 2012, roč. 2012, s. 1-9.

[199] *Illumina DRAGEN Bio-IT Platform v3.10*. Online. 2022. Dostupné z: https://support-docs.illumina.com/SW/DRAGEN_v310/Content/SW/FrontPages/DRAGEN.htm. [cit. 2024-01-22].

[200] *GitHub*. Online. 2024. Dostupné z: https://github.com/pwwang/vcfstats. [cit. 2024-01-22].

[201] DEWHIRST, Floyd E.; CHEN, Tuste; IZARD, Jacques; PASTER, Bruce J.; TANNER, Anne C. R. et al. The Human Oral Microbiome. Online. *Journal of Bacteriology*. 2010, roč. 192, č. 19, s. 5002-5017. ISSN 0021-9193. Dostupné z: https://doi.org/10.1128/JB.00542-10. [cit. 2024-01-23].

[202] DANECEK, Petr; BONFIELD, James K; LIDDLE, Jennifer; MARSHALL, John; OHAN, Valeriu et al. Twelve years of SAMtools and BCFtools. Online. *GigaScience*. 2021, roč. 10, č. 2, s. 1-4. ISSN 2047-217X. Dostupné z: https://doi.org/10.1093/gigascience/giab008. [cit. 2024-03-07].

[203] MILLER, David T.; LEE, Kristy; ABUL-HUSN, Noura S.; AMENDOLA, Laura M.; BROTHERS, Kyle et al. ACMG SF v3.2 list for reporting of secondary findings in clinical exome and genome sequencing: A policy statement of the American College of Medical Genetics and Genomics (ACMG). Online. *Genetics in Medicine*. 2023, roč. 25, č. 8, s. 1-6. ISSN 10983600. Dostupné z: https://doi.org/10.1016/j.gim.2023.100866. [cit. 2024-03-08].

[204] SØRENSEN, T. A Method of Establishing Groups of Equal Amplitude in Plant Sociology Based on Similarity of Species Content and Its Application to Analyses of the Vegetation on Danish Commons. *Kongelige Danske Videnskabernes Selskab, Biologiske Skrifter*. 1948, č. 5, s. 1-34.

[205] GITHUB. *Illumina/hap.py*. Online. GITHUB. Illumina/hap.py. 2024. Dostupné z: https://github.com/Illumina/hap.py. [cit. 2024-02-29].

[206] VALLAT, Raphael. Pingouin: statistics in Python. Online. *Journal of Open Source Software*. 2018, roč. 3, č. 31, s. 1. ISSN 2475-9066. Dostupné z: https://doi.org/10.21105/joss.01026. [cit. 2024-01-25].

[207] OLIPHANT, Travis E. Python for Scientific Computing. Online. *Computing in Science & Engineering*. 2007, roč. 9, č. 3, s. 10-20. ISSN 1521-9615. Dostupné z: https://doi.org/10.1109/MCSE.2007.58. [cit. 2024-01-25].

[208] MZČR. *Onemocnění aktuálně*. Online. MZČR. Onemocnění aktuálně. 2024. Dostupné z: https://onemocneni-aktualne.mzcr.cz/covid-19. [cit. 2024-03-04].

[209] THERMO FISHER SCIENTIFIC. *Qubit™ RNA High Sensitivity (HS), Broad Range (BR), and Extended Range (XR) Assay Kits*. Online. THERMO FISHER SCIENTIFIC. Qubit™ RNA High Sensitivity (HS), Broad Range (BR), and Extended Range (XR) Assay Kits. 2024. Dostupné z: https://www.thermofisher.com/order/catalog/product/Q32852. [cit. 2024-03-04].

[210] KRIEGOVA, Eva; FILLEROVA, Regina a KVAPIL, Petr. Direct-RT-qPCR Detection of SARS-CoV-2 without RNA Extraction as Part of a COVID-19 Testing Strategy: From Sample to Result in

One Hour. Online. *Diagnostics*. 2020, roč. 10, č. 8, s. 1-10. ISSN 2075-4418. Dostupné z: https://doi.org/10.3390/diagnostics10080605. [cit. 2024-03-06].

[211] THEDA, Christiane; HWANG, Seo Hye; CZAJKO, Anna; LOKE, Yuk Jing; LEONG, Pamela et al. Quantitation of the cellular content of saliva and buccal swab samples. Online. *Scientific Reports*. 2018, roč. 8, č. 1, s. 1-8. ISSN 2045-2322. Dostupné z: https://doi.org/10.1038/s41598-018-25311-0. [cit. 2024-03-11].

[212] SAMSON, C. A.; WHITFORD, W.; SNELL, R. G.; JACOBSEN, J. C. a LEHNERT, K. Contaminating DNA in human saliva alters the detection of variants from whole genome sequencing. Online. *Scientific Reports*. 2020, roč. 10, č. 1, s. 1-9. ISSN 2045-2322. Dostupné z: https://doi.org/10.1038/s41598-020-76022-4. [cit. 2024-03-11].

[213] HERZIG, Anthony F.; VELO-SUÁREZ, Lourdes; LE FOLGOC, Gaëlle; BOLAND, Anne; BLANCHÉ, Hélène et al. Evaluation of saliva as a source of accurate whole-genome and microbiome sequencing data. Online. *Genetic Epidemiology*. 2021, roč. 45, č. 5, s. 537-548. ISSN 0741-0395. Dostupné z: https://doi.org/10.1002/gepi.22386. [cit. 2024-03-11].

[214] YAO, Roderick A.; AKINRINADE, Oyediran; CHAIX, Marie a MITAL, Seema. Quality of whole genome sequencing from blood versus saliva derived DNA in cardiac patients. Online. *BMC Medical Genomics*. 2020, roč. 13, č. 1, s. 1-10. ISSN 1755-8794. Dostupné z: https://doi.org/10.1186/s12920-020-0664-7. [cit. 2024-03-11].

[215] ZOOK, Justin M.; MCDANIEL, Jennifer; OLSON, Nathan D.; WAGNER, Justin; PARIKH, Hemang et al. An open resource for accurately benchmarking small variant and reference calls. Online. *Nature Biotechnology*. 2019, roč. 37, č. 5, s. 561-566. ISSN 1087-0156. Dostupné z: https://doi.org/10.1038/s41587-019-0074-6. [cit. 2024-03-11].

[216] CORIELL INSTITUTE FOR MEDICAL RESEARCH. *NA12878*. Online. 2024. Dostupné z: https://catalog.coriell.org/0/Sections/Search/Sample_Detail.aspx?Ref=NA12878&Product=DNA. [cit. 2024-03-11].

[217] KUBIRITOVA, Zuzana; GYURASZOVA, Marianna; NAGYOVA, Emilia; HYBLOVA, Michaela; HARSANYOVA, Maria et al. On the critical evaluation and confirmation of germline sequence variants identified using massively parallel sequencing. Online. *Journal of Biotechnology*. 2019, roč. 298, s. 64-75. ISSN 01681656. Dostupné z: https://doi.org/10.1016/j.jbiotec.2019.04.013. [cit. 2024-04-02].

[218] KRUSCHE, Peter; TRIGG, Len; BOUTROS, Paul C.; MASON, Christopher E.; DE LA VEGA, Francisco M. et al. Best practices for benchmarking germline small-variant calls in human genomes.

Online. *Nature Biotechnology*. 2019, roč. 37, č. 5, s. 555-560. ISSN 1087-0156. Dostupné z: https://doi.org/10.1038/s41587-019-0054-x. [cit. 2024-03-11].

[219] ŞENEL, Sevda. An Overview of Physical, Microbiological and Immune Barriers of Oral Mucosa. Online. *International Journal of Molecular Sciences*. 2021, roč. 22, č. 15, s. 1-15. ISSN 1422-0067. Dostupné z: https://doi.org/10.3390/ijms22157821. [cit. 2024-03-11].

[220] SOSONKINA, Nadiya; KELLY, Melissa; HOLT, James; BICK, David a NAKOUZI, Ghunwa. EP403: Finding merit in impurity. Online. *Genetics in Medicine*. 2022, roč. 24, č. 3, s. S253-S254. ISSN 10983600. Dostupné z: https://doi.org/10.1016/j.gim.2022.01.438. [cit. 2024-03-11].

[221] THORVALDSDOTTIR, H.; ROBINSON, J. T. a MESIROV, J. P. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Online. *Briefings in Bioinformatics*. 2013, roč. 14, č. 2, s. 178-192. ISSN 1467-5463. Dostupné z: https://doi.org/10.1093/bib/bbs017. [cit. 2024-03-22].

[222] ALABDI, Lama; SHAMSELDIN, Hanan E.; KHOUJ, Ebtissal; HELABY, Rana; ALJAMAL, Bayan et al. Beyond the exome: utility of long-read whole genome sequencing in exome-negative autosomal recessive diseases. Online. *Genome Medicine*. 2023, roč. 15, č. 1, s. 1-16. ISSN 1756-994X. Dostupné z: https://doi.org/10.1186/s13073-023-01270-8. [cit. 2024-03-12].

APPENDIX

Klempt, P.; Brož, P.; Kašný, M.; Novotný, A.; Kvapilová, K.; Kvapil, P. Performance of Targeted Library Preparation Solutions for SARS-CoV-2 Whole Genome Analysis. *Diagnostics 10*, 769 (**2020**). https://doi.org/10.3390/diagnostics10100769

Kvapilova, K., Misenko, P., Radvanszky, J. *et al.* Validated WGS and WES protocols proved saliva-derived gDNA as an equivalent to blood-derived gDNA for clinical and population genomic analyses. *BMC Genomics* **25**, 187 (2024). https://doi.org/10.1186/s12864-024-10080-0