Bulgarian Chemical Communications, Volume 55, Special Issue C (pp. 95–102) 2023 *DOI: 10.34049/bcc.55.C.0011*

Bioinformatics pipeline for variant detection in targeted sequencing panel

M. Topalov[1], V Stoyanova[1], H. Ivanov[1], G. Raycheva[2], V. Popov[1], A. Linev[1], N. Miteva-Marcheva[1], D. Dimitrov[1], Zh. Grudeva-Popova[2]

[1] *Department of Pediatrics and Medical Genetics, Medical University Plovdiv*
[2] *Department of Clinical Oncology, Medical University Plovdiv*

Received: November 2023; Revised: December 2023

Bioinformatics pipeline development is fundamental for extracting meaningful results from high-throughput sequencing data in robust manner. This study introduces a meticulously crafted bioinformatics pipeline customized for the analysis of Illumina TruSight Tumor 15 panel data, utilizing the QIAGEN CLC Genomics Workbench software. The panel serves as a comprehensive solution for detecting somatic mutations in genes associated with cancer, proving it applicable for cancer research.

Keywords: next-generation sequencing, bioinformatics, single nucleotide variants, oncology, molecular genetics.

INTRODUCTION

Targeted sequencing, alternatively referred to as "targeted resequencing" or "amplicon sequencing," is a precision-oriented method for DNA sequencing, allowing researchers to delve deep into specifically chosen regions of the genome. In the context of cancer research and diagnostics, targeted sequencing commonly involves the utilization of sequencing panels designed to cover sets of specific genes associated with cancer. By concentrating on these exact genomic areas, targeted sequencing facilitates a more efficient and cost-effective analysis on regions of particular interest for cancer detection and characterization.

The integration of bioinformatics tools into pipelines is crucial in deriving significant insights from high-throughput sequencing data efficiently. As the volume of sequencing data continues to surge, the automation of bioinformatics analysis emerges as an imperative solution. The development of a bioinformatics pipeline tailored for variant detection from targeted sequencing panel data plays a pivotal role in gaining in-depth insights into the genetic intricacies of cancer. This perspective proves invaluable for the customization of personalized treatment strategies, particularly within the realms of Oncology and Molecular Genetics, where precision is paramount. The insights derived from such an adept bioinformatics pipeline designed for variant detection in targeted sequencing panel data offer a valuable foundation for advancing the understanding and treatment of cancer.

MATERIALS AND METHODS

Illumina Trusight Tumor 15 is a panel for targeted NGS sequencing of fifteen genes for which mutations are known to be found in solid tumors. The panel accurately detects low frequency genetic variants from 20 ng of DNA and is optimized for Formalin-Fixed Paraffin-Embedded (FFPE) tissue samples. The list of genes covered by Illumina Trusight Tumor 15 is:

AKT1	AKT serine/threonine kinase 1
BRAF	B-Raf proto-oncogene, serine/threonine kinase
EGFR	epidermal growth factor receptor
ERBB2	erb-b2 receptor tyrosine kinase 2
FOXL2	forkhead box L2
GNA11	G protein subunit alpha 11
GNAQ	G protein subunit alpha q
KIT	KIT proto-oncogene, receptor tyrosine kinase
KRAS	KRAS proto-oncogene, GTPase
MET	MET proto-oncogene, receptor tyrosine kinase

* To whom all correspondence should be sent:
E-mail: momchil.topalov@phd.mu-plovdiv.bg

NRAS	NRAS proto-oncogene, GTPase
PDGFRA	platelet derived growth factor receptor alpha
PIK3CA	phosphatidylinositol-4,5-bisphosphate 3-kinase catalytic subunit alpha
RET	ret proto-oncogene
TP53	tumor protein p53

The panel consists of 1850 target regions, spreading over 370942bp to ensure optimal coverage over these fifteen genes.

Thirty-seven libraries from anonymized breast cancer patients were prepared with the Illumina Trusight Tumor 15 panel at the Department of Pediatrics and Medical Genetics, Medical University Plovdiv.

Illumina MiSeq is an advanced platform for next-generation DNA sequencing, applying sequencing-by-synthesis technology. Illumina MiSeq is suitable for sequencing data from targeted sequencing libraries with the pair-end protocol [1]. This technique carries information of the nucleotide sequence as well as the distance between the reads of the pair. This increases accuracy when assembling new genomes or detecting genetic mutations in known ones [1].

The sequencing of the libraries was conducted according to the protocol [2] on Illumina MiSeq sequencer at the Department of Pediatrics and Medical Genetics, Medical University Plovdiv [3–9].

Illumina BaseSpace is a cloud-based genomics platform offering NGS data management and analysis. The data generated by Illimina MiSeq is archived in BaseSpace. Using the BaseSpace Downloader client, the raw sequencing data is downloaded locally in FASTQ format.

QIAGEN CLC Genomics Workbench v23.0.4 is an easy-to-use graphical interface software for bioinformatic analysis of next-generation sequencing data. The tools in it can be run individually or linked together in workflows, enabling the development of complex bioinformatics pipelines.

Biomedical Genomics Analysis v.23.0.2 is a plugin for QIAGEN CLC Genomics Workbench designed mainly for biomedical and oncological sequencing data [10].

RESULTS AND DISCUSSION

The QIAGEN CLC Genomics Workbench and Biomedical Genomics Analysis were used to develop automated bioinformatics pipeline for the analysis of NGS sequencing data of fifteen breast cancer-associated genes by to extracting meaningful information from the targeted sequencing data. Key steps in the bioinformatics pipeline include:

QC for sequencing reads

Quality control (QC) for sequencing reads is a pivotal process in ensuring the integrity of high-throughput sequencing data. Analyzing various aspects of the data is crucial for reliable downstream analyses. The sequencing yield of the Illumina Trusight Tumor 15 libraries was with 94.9% of bases being ≥ Q30 quality per sample in a 2×150nt pair-end reads protocol.

Trimming the sequencing reads

The trimming step refines the sequence reads before mapping. The trimming consists of adapter trimming, quality trimming, and length trimming altogether (Table 1).

Read mapping

Mapping reads to a human reference genome is a fundamental step in most applications of high-throughput sequencing data. In the current bioinformatics pipeline, the reference genome is Hg38. The parameters are tuned towards adding an extra affine cost associated with opening a such that long contiguous gaps are favored over short gaps, reflecting the mutagenic potential of cancer cells.

Removing ligation artifacts

During the adapter ligation of the Illumina Trusight Tumor 15 library preparation, there can be the case that two different DNA sequences also get ligated together. Such ligation artifacts are prone to occur with higher probability between short DNA fragments, such as the ones generated from FFPE samples. This step of the pipeline removes reads from the read mapping which are likely the result of such an event (Table 2).

Calling for structural variants

A key step of the pipeline is calling structural variants such as deletions, insertions, tandem duplications and inversions by looking for unaligned read ends at each chromosome position. The estimated breakpoints which are instrumental for the downstream analysis as well. In the pool of thirty-

Table 1. Trimming report

Sample name	Reads after trim (%)	Avg. length after trim
LIN_S3_S8	99.9998717	131.3978193
NAK_S4_S9	99.99984623	130.9411136
CAA_S5_S10	99.99984689	130.3090855
NAD_S2_S7	99.9999303	128.9517914
VKI_S1_S6	99.99989479	123.1953821
EKD_S7_S8	99.99965306	129.3947973
LNN_S3_S4	99.99973685	134.1213624
NRG_S5_S6	99.99940513	122.8248149
RCG_S1_S2	99.9997592	134.2346325
RGG_S9_S10	99.99977549	137.2786
LIS_S7_S8	99.99973795	119.6411268
NGT_S3_S4	99.99983826	117.9993056
NPG_S1_S2	99.99984732	121.9701612
PRV_S5_S6	99.99958752	111.1261861
TTG_S9_S10	99.99975481	115.3774664
AAS_S5_S6	99.9998814	117.334696
DHG_S9_S10	99.99957716	95.45574896
EDK_S7_S8	99.99978278	115.2351194
IDD_S1_S2	99.99983687	119.6345178
PGA_S3_S4	99.99988502	117.129644
CDA_SVA_SVB_Sample1	99.9996475	136.5618026
CDA_VGA_VGN_Sample2	99.99969203	133.5521548
CDA_DVA_DVB_Sample3	99.9996645	137.5169503
CDA_AMA_AMB_Sample4	99.99947557	133.2611848
CDA_ZGA_ZGB_Sample5	99.99959254	133.6849961
CDA_MTA_MTB_Sample6	99.9995558	128.814802
CDA_DKA_DKB_Sample7	99.99959025	131.2575002
CDA_SBA_SBB_Sample8	99.99939711	129.5844282
FCD_SVA_SVB_Sample1	99.99981521	136.5594603
FCD_ZGA_ZGB_Sample2	99.99972504	133.8112269
FCD_DKA_DKB_Sample3	99.99976356	131.2795209
FCD_SBA_SBB_Sample4	99.99966349	129.6751804
MD_S2_S7	99.9995275	136.4810005
MS_S3_S8	99.99852922	107.8014028
SA_S4_S9	99.99926349	126.8381724
SG_S5_S10	99.99881518	126.1098424
SK_S1_S6	99.99954313	138.3859788
Minimum	99.99852922	95.45574896
Median	99.99972504	129.5844282
Maximum	99.9999303	138.3859788
Mean	99.99964081	126.6142966
Standard deviation	0.000284196	9.61145427

seven samples, only deletions and tandem duplications are called (Table 3).

Local Realignment

The goal of the local realignment tool is to improve the alignments of the reads in an existing read mapping. An opening for realignment may occur in areas around insertions and deletions in the reads relative to the reference. As a result, an alternative mapping, as good as or better than the original, can be generated.

QC for read mapping

Another QC metric step is included in the pipeline, measuring the performance of the read mapping after the improvements and modifications introduced by remove ligation artifacts and local realignment steps (Table 4).

Table 2. Remove ligation artifacts report

Sample name	Matches in input	Ligation artifacts found and removed	Artifacts trimmed	Artifacts trimmed from single or broken pair reads	Artifacts trimmed from paired read ends
LIN_S3_S8	6247055	22	197335	27988	169347
NAK_S4_S9	5873331	13	171633	27852	143781
CAA_S5_S10	6559845	35	204194	26183	178011
NAD_S2_S7	6476281	4	219959	27687	192272
VKI_S1_S6	6671510	11	221403	32774	188629
EKD_S7_S8	6095406	17	189669	40346	149323
LNN_S3_S4	6694825	44	131072	32817	98255
NRG_S5_S6	6427431	46	195126	44301	150825
RCG_S1_S2	6068593	21	163613	29899	133714
RGG_S9_S10	6278232	19	145073	30773	114300
LIS_S7_S8	6302898	13	137208	15846	121362
NGT_S3_S4	5875878	10	89043	23155	65888
NPG_S1_S2	5246639	5	145398	19301	126097
PRV_S5_S6	6414330	30	173912	21957	151955
TTG_S9_S10	6312553	17	136956	28507	108449
AAS_S5_S6	5508925	20	116398	25556	90842
DHG_S9_S10	6543034	46	167751	46203	121548
EDK_S7_S8	5321912	35	119833	29626	90207
IDD_S1_S2	5553372	33	125841	34886	90955
PGA_S3_S4	5691248	28	125760	33401	92359
CDA_SVA_SVB_Sample1	3979621	8	28224	4748	23476
CDA_VGA_VGN_Sample2	4067005	3	31179	5741	25438
CDA_DVA_DVB_Sample3	3734476	3	28321	5095	23226
CDA_AMA_AMB_Sample4	3821816	2	29450	5571	23879
CDA_ZGA_ZGB_Sample5	3812025	5	32506	6270	26236
CDA_MTA_MTB_Sample6	4847841	6	40336	8822	31514
CDA_DKA_DKB_Sample7	4646628	4	38920	7232	31688
CDA_SBA_SBB_Sample8	4154239	11	34148	6557	27591
FCD_SVA_SVB_Sample1	7867491	13	63113	11537	51576
FCD_ZGA_ZGB_Sample2	7475663	7	77211	14483	62728
FCD_DKA_DKB_Sample3	9330096	12	85711	16722	68989
FCD_SBA_SBB_Sample4	8342277	6	82185	15867	66318
MD_S2_S7	3920462	12	17166	2030	15136
MS_S3_S8	3908374	12	25353	3257	22096
SA_S4_S9	4144416	17	21205	2707	18498
SG_S5_S10	3166411	3	17614	2240	15374
SK_S1_S6	3508904	13	13807	2207	11600
Minimum	3166411	2	13807	2030	11600
Median	5873331	13	116398	19301	90207
Maximum	9330096	46	221403	46203	192272
Mean	5591650	16.3783784	103881.8	19463.3514	84418.4324
Standard deviation	1448724	12.6564682	68025.84	13187.7593	56832.79

Target region coverage

Measuring the read coverage over target regions is instrumental for evaluating the overall quality of the sample and determining if the variant calling results are reliable. In the pool of thirty-seven samples, above 98.7 of all targets were covered by 160 reads or more, securing high sensitivity for the variant calling of single nucleotide polymorphisms (SNPs) (Table 5).

Low Frequency Variant Detection

Variant calling is the primary step in deciphering the genetic code, involving the identification of variations such as single SNPs, small insertions, and deletions. A step for calling variants with low frequency is a necessary attribute in the pipeline for targeted sequencing analysis of samples of mixed tissue types such as cancer samples. In such samples, low frequent variants are likely to be present,

Table 3. Structural variant caller report

Sample name	Left breakpoints	Right breakpoints	Deletions	Tandem Duplications
LIN_S3_S8	45	47	1	1
NAK_S4_S9	47	47	2	1
CAA_S5_S10	53	49	1	1
NAD_S2_S7	49	35	1	1
VKI_S1_S6	50	58	1	2
EKD_S7_S8	34	57	2	0
LNN_S3_S4	43	54	2	0
NRG_S5_S6	55	70	2	0
RCG_S1_S2	46	48	1	1
RGG_S9_S10	31	45	2	0
LIS_S7_S8	48	53	2	0
NGT_S3_S4	73	97	1	1
NPG_S1_S2	50	69	0	0
PRV_S5_S6	60	62	1	0
TTG_S9_S10	39	59	0	1
AAS_S5_S6	32	52	1	0
DHG_S9_S10	55	67	1	0
EDK_S7_S8	48	49	0	0
IDD_S1_S2	37	40	2	0
PGA_S3_S4	59	54	1	1
CDA_SVA_SVB_Sample1	21	39	1	0
CDA_VGA_VGN_Sample2	17	42	1	0
CDA_DVA_DVB_Sample3	22	35	0	0
CDA_AMA_AMB_Sample4	26	43	1	0
CDA_ZGA_ZGB_Sample5	21	39	1	1
CDA_MTA_MTB_Sample6	25	38	1	0
CDA_DKA_DKB_Sample7	27	38	1	0
CDA_SBA_SBB_Sample8	20	45	1	1
FCD_SVA_SVB_Sample1	31	52	1	1
FCD_ZGA_ZGB_Sample2	28	53	2	0
FCD_DKA_DKB_Sample3	30	54	1	0
FCD_SBA_SBB_Sample4	28	66	1	0
MD_S2_S7	22	64	0	2
MS_S3_S8	24	47	1	0
SA_S4_S9	31	57	3	0
SG_S5_S10	24	51	0	0
SK_S1_S6	20	61	1	1
Minimum	17	35	0	0
Median	32	52	1	0
Maximum	73	97	3	2
Mean	37.0540541	52.324324	1.108108	0.43243243
Standard deviation	14.1596962	12.213504	0.698561	0.60279629

as well as for samples for which the ploidy is unknown or not well defined. The step allows for calling variants with minimum frequency starting from 0.4%, calculated as 'count of reads supporting the variant'/'the overall coverage in that region'.

This low minimum frequency of 0.4% is less than the industry standard of 0.5% as it aims to detect significantly low frequency variants that have cancer origin but are not represented definitively in the sample.

Variant filtering cascade

The increased the risk of introducing false positive variants, requires for the pipeline to provide a stricter variant filtering cascade of steps that filters out marginal variants, variants in regions of insufficient read depth, sequencing errors, alignment artifacts and random noise in the data. The filtering cascade applies criteria that balances between sensitivity and specificity, keeping only the true variants.

Table 4. QC for read mapping report

Sample name	Reads (#)	Mapped reads (#)	Mapped reads (%)	Not mapped reads (#)	Not mapped reads (%)
LIN_S3_S8	12470686	12281843	98.4857	188843	1.514295
NAK_S4_S9	11706112	11500925	98.24718	205187	1.752819
CAA_S5_S10	13062798	12790260	97.91363	272538	2.086368
NAD_S2_S7	12911622	12692645	98.30403	218977	1.695968
VKI_S1_S6	13307256	13041130	98.00014	266126	1.999856
EKD_S7_S8	12105606	11926284	98.51869	179322	1.481314
LNN_S3_S4	13300456	13119071	98.63625	181385	1.36375
NRG_S5_S6	12775740	12514999	97.95909	260741	2.040907
RCG_S1_S2	12043094	11849744	98.39452	193350	1.605484
RGG_S9_S10	12471656	12294052	98.57594	177604	1.424061
LIS_S7_S8	12592868	12381540	98.32184	211328	1.678156
NGT_S3_S4	11747372	11562761	98.42849	184611	1.571509
NPG_S1_S2	10479330	10299609	98.285	179721	1.715005
PRV_S5_S6	12849046	12566297	97.79946	282749	2.200545
TTG_S9_S10	12642988	12379329	97.91458	263659	2.085417
AAS_S5_S6	10961534	10725781	97.84927	235753	2.15073
DHG_S9_S10	13007040	12590283	96.79591	416757	3.204088
EDK_S7_S8	10588342	10348217	97.73218	240125	2.267824
IDD_S1_S2	11034058	10817898	98.04097	216160	1.959025
PGA_S3_S4	11306646	11055941	97.78268	250705	2.217324
CDA_SVA_SVB_Sample1	7943146	7916378	99.66301	26768	0.336995
CDA_VGA_VGN_Sample2	8117528	8081209	99.55259	36319	0.447415
CDA_DVA_DVB_Sample3	7451546	7420524	99.58368	31022	0.416316
CDA_AMA_AMB_Sample4	7627260	7594159	99.56602	33101	0.433983
CDA_ZGA_ZGB_Sample5	7608106	7571658	99.52093	36448	0.479068
CDA_MTA_MTB_Sample6	9680330	9636050	99.54258	44280	0.457422
CDA_DKA_DKB_Sample7	9273790	9224873	99.47252	48917	0.527476
CDA_SBA_SBB_Sample8	8293230	8247843	99.45272	45387	0.547278
FCD_SVA_SVB_Sample1	15693192	15630866	99.60285	62326	0.397153
FCD_ZGA_ZGB_Sample2	14911090	14824909	99.42203	86181	0.577966
FCD_DKA_DKB_Sample3	18609030	18499235	99.40999	109795	0.590009
FCD_SBA_SBB_Sample4	16641354	16533409	99.35134	107945	0.648655
MD_S2_S7	7830650	7812792	99.77195	17858	0.228053
MS_S3_S8	7818742	7789944	99.63168	28798	0.36832
SA_S4_S9	8282230	8256896	99.69412	25334	0.305884
SG_S5_S10	6329902	6311764	99.71346	18138	0.286545
SK_S1_S6	7004062	6985695	99.73777	18367	0.262234
Minimum	6329902	6311764	96.79591	17858	0.228053
Median	11706112	11500925	98.57594	179322	1.424061
Maximum	18609030	18499235	99.77195	416757	3.204088
Mean	11148093	11002076	98.77499	146016.9	1.225006
Standard deviation	2883528.5	2828923	0.807734	104113.8	0.807734

Functional Annotation

The filtered variants are annotated with ClinVar and dbSNP[11]. The variants are categorized based on their location within coding regions, splice sites, or regulatory elements. Predictive algorithms were employed to assess the deleteriousness of the variants, prioritizing those with potential clinical relevance which are further scrutinized for their potential implications in cancer development.

Since the target regions are covering both strands of DNA, the result of annotated variants is split into two groups.

The first group contains only variants overlapping with the genes from Illumina Trusight Tumor 15 list. These variants are the primary focus on the pipeline. In all thirty-seven samples, an average of 43 mutations associated with cancer were detected, 97.98% of which were single nucleotide polymorphisms (SNV) and 2.02% were deletions [12–16].

Table 5. Target region coverage report

Sample name	Max coverage	Avg. coverage	Target regions with low coverage (%)	Length of target region positions with low coverage (%)
LIN_S3_S8	86027	333.717	98.7027027	98.87260003
NAK_S4_S9	74626	291.8734	98.7027027	98.87206086
CAA_S5_S10	85181	337.4652	98.7027027	98.87233045
NAD_S2_S7	163235	409.5384	98.7027027	98.87233045
VKI_S1_S6	105854	385.5133	98.7027027	98.87206086
EKD_S7_S8	80433	331.3737	98.7027027	98.87260003
LNN_S3_S4	92358	357.5401	98.7027027	98.87260003
NRG_S5_S6	81814	310.9182	98.7027027	98.87260003
RCG_S1_S2	90000	324.8757	98.7027027	98.87260003
RGG_S9_S10	155637	318.3564	98.7027027	98.87206086
LIS_S7_S8	83242	294.9421	98.7027027	98.87260003
NGT_S3_S4	107295	338.3943	98.7027027	98.87233045
NPG_S1_S2	89869	291.8255	98.7027027	98.87233045
PRV_S5_S6	101883	271.4018	98.7027027	98.87260003
TTG_S9_S10	71662	273.177	98.7027027	98.87260003
AAS_S5_S6	70122	257.7923	98.7027027	98.87260003
DHG_S9_S10	144870	220.1225	98.7027027	98.87260003
EDK_S7_S8	66832	246.0914	98.7027027	98.87260003
IDD_S1_S2	69096	258.6306	98.7027027	98.87260003
PGA_S3_S4	75957	246.8681	98.7027027	98.87260003
CDA_SVA_SVB_Sample1	62177	219.9835	98.7027027	98.87260003
CDA_VGA_VGN_Sample2	51875	216.3667	98.7027027	98.87260003
CDA_DVA_DVB_Sample3	52585	212.0567	98.7027027	98.87260003
CDA_AMA_AMB_Sample4	52292	216.2776	98.7027027	98.87260003
CDA_ZGA_ZGB_Sample5	52648	216.4168	98.7027027	98.87260003
CDA_MTA_MTB_Sample6	72336	238.4238	98.7027027	98.87260003
CDA_DKA_DKB_Sample7	71213	249.8491	98.7027027	98.87260003
CDA_SBA_SBB_Sample8	67409	226.2924	98.7027027	98.87260003
FCD_SVA_SVB_Sample1	123303	435.3392	98.7027027	98.87260003
FCD_ZGA_ZGB_Sample2	104546	427.1888	98.7027027	98.87260003
FCD_DKA_DKB_Sample3	139052	501.2281	98.7027027	98.87233045
FCD_SBA_SBB_Sample4	135785	452.3595	98.7027027	98.87260003
MD_S2_S7	57901	219.9873	98.7027027	98.87260003
MS_S3_S8	73860	174.4307	98.75675676	98.88742714
SA_S4_S9	52779	208.149	98.7027027	98.87260003
SG_S5_S10	40647	161.1073	98.7027027	98.87260003
SK_S1_S6	54232	209.2885	98.7027027	98.87260003
Minimum	40647	161.1073	98.7027027	98.87206086
Median	75957	271.4018	98.7027027	98.87260003
Maximum	163235	501.2281	98.75675676	98.88742714
Mean	85422.51	288.7882	98.70416362	98.87292062
Standard deviation	31100.4	82.41751	0.008886432	0.002456715

The second group contains variants that passed all filtering criteria and belong to the opposite strand of DNA for the respective gene. Such variants may overlap with another genes, pseudogenes, or long non-coding RNA. This is an extra piece of information available from the targeted sequencing panel that is usually is overlooked by the standard pipelines. The current pipeline records such variants as they have the potential of additional insights and may have application in populational genetics [17] if the pipeline is run for a larger cohort of patients.

Analysis of the thirty-seven samples using this pipeline reveals a comprehensive landscape of somatic SNVs and indel mutations within cancer-related genes [18]. The pipeline effectively identifies putative mutations, thus providing valuable insights into their significance within cancer research, contributing to the development of personalized treatment strategies [19].

The pipeline finishes with an export step that is preparing the lists with variants in format suitable for further interpretation by medical genetics and oncology experts.

CONCLUSION

In conclusion, this bioinformatics pipeline demonstrates its effectiveness in systematically analyzing Illumina TruSight Tumor 15 panel data across thirty-seven samples. It presents an in-depth overview of the performance of the sample during the library preparation and sequencing by generating detailed reports with high precision metrics. It serves as a promising resource for advancing cancer research and clinical care. The pipeline provides a descriptive grouping of the variants. Further validation and seamless integration with clinical data and functional annotation with more external resources are imperative next steps in realizing the full potential of this pipeline in oncology research. The application of this bioinformatics pipeline for variant detection in targeted sequencing panel expands our knowledge of these specific genes but also paves the way for further research into personalized medicine and targeted therapies.

Acknowledgments: *Project BG05M2OP001-1.002 -0005 – Competence Center "Personalized Innovative medicine (PERIMED)", financed by Operational Program "Science and Education for Smart Growth", EU, ESIF.*

REFERENCES

1. J. J. Kozich, S. L. Westcott, N. T. Baxter, S. K. Highlander, P. D. Schloss, *Appl. Environ. Microbiol.*, **79**, 5112 (2013).
2. Illumina, MiSeq System Guide (15027617), www.illumina.com (2021).
3. Liu, L. et al., *J. Biomed. Biotechnol.*, **2012**, 251364 (2012).
4. J. Podnar, H. Deiderick, G. Huerta, S. Hunicke-Smith, *Curr. Protoc. Mol. Biol.*, **106**, 4.21.1 (2014).
5. R. H. Deurenberg et al., *J. Biotechnol.*, **243**, 16 (2017).
6. J. Plitnick et al., *J. Clin. Microbiol.*, **59**(12), e0064921 (2021).
7. D. A. Read, G. Pietersen, *Methods Mol. Biol*., **2015**, 179 (2019).
8. T. Unno, *J. Microbiol. Biotechnol.*, **25**(6), 765 (2015).
9. V. Valentini et al., *Front. Oncol.*, **12**, 1092201 (2022).
10. R. K. Ravi, K. Walton, M. Khosroheidari, *Methods Mol. Biol.*, **1706**, 223 (2018).
11. S. T. Sherry et al., *Nucleic Acids Res*., **29**, 308 (2001).
12. A. Z. Dayem Ullah, N.R. Lemoine, C. Chelala, *Brief Bioinform*., **14**, 437 (2013).
13. J. Oscanoa et al., *Nucleic Acids Res.*, **48**, W185 (2020).
14. A. Z. Dayem Ullah et al., *Nucleic Acids Res.*, **46**, W109 (2018).
15. A. Z. Dayem Ullah, N. R. Lemoine, C. Chelala, *Nucleic Acids Res.*, **40**, W65 (2012).
16. C. Chelala, A. Khan, N. R. Lemoine, *Bioinformatics*, **25**, 655 (2009).
17. G. Ribas et al., *Hum. Genet.*, **118**, 669 (2006).
18. I. Adzhubei, D. M. Jordan, S. R. Sunyaev, *Curr. Protoc. Hum. Genet*., **7**, 7.20.1 (2013).
19. P. C. Ng, S. Henikoff, *Nucleic Acids Res.*, **31**, 3812 (2003).