Journal of Biotechnology 298 (2019) 64–75

Contents lists available at ScienceDirect

Journal of Biotechnology

journal homepage: www.elsevier.com/locate/jbiotec

On the critical evaluation and confirmation of germline sequence variants identified using massively parallel sequencing

Zuzana Kubiritova[a,b], Marianna Gyuraszova[b,c], Emilia Nagyova[b,d], Michaela Hyblova[b,e], Maria Harsanyova[b,e], Jaroslav Budis[e,f,g], Rastislav Hekel[b,e,g], Juraj Gazdarica[b,e,g], Frantisek Duris[e,g], Ludevit Kadasi[a,b], Tomas Szemes[b,e,h], Jan Radvanszky[a,e,*]

[a] Institute for Clinical and Translational Research, Biomedical Research Center, Slovak Academy of Sciences, Bratislava, Slovakia
[b] Department of Molecular Biology, Faculty of Natural Sciences, Comenius University, Bratislava, Slovakia
[c] Institute of Molecular Biomedicine, Faculty of Medicine, Comenius University, Bratislava, Slovakia
[d] Department of Cardiology, Division Heart & Lungs, UMC Utrecht, University of Utrecht, the Netherlands
[e] Geneton Ltd., Bratislava, Slovakia
[f] Department of Computer Science, Faculty of Mathematics, Physics and Informatics, Comenius University, Bratislava, Slovakia
[g] Slovak Centre of Scientific and Technical Information, Bratislava, Slovakia
[h] Comenius University Science Park, Bratislava, Slovakia

ARTICLE INFO

Keywords:
Massively parallel sequencing
Sanger sequencing
Genomic diagnostics
Sequence variant
Sequence variant confirmation

ABSTRACT

Although massively parallel sequencing (MPS) is becoming common practice in both research and routine clinical care, confirmation requirements of identified DNA variants using alternative methods are still topics of debate. When evaluating variants directly from MPS data, different read depth statistics, together with specialized genotype quality scores are, therefore, of high relevance. Here we report results of our validation study performed in two different ways: 1) confirmation of MPS identified variants using Sanger sequencing; and 2) simultaneous Sanger and MPS analysis of exons of selected genes. Detailed examination of false-positive and false-negative findings revealed typical error sources connected to low read depth/coverage, incomplete reference genome, indel realignment problems, as well as microsatellite associated amplification errors leading to base miss-calling. However, all these error types were identifiable with thorough manual revision of aligned reads according to specific patterns of distributions of variants and their corresponding reads. Moreover, our results point to dependence of both basic quantitative metrics (such as total read counts, alternative allele read counts and allelic balance) together with specific genotype quality scores on the used bioinformatics pipeline, stressing thus the need for establishing of specific thresholds for these metrics in each laboratory and for each involved pipeline independently.

1. Introduction

Since the costs of DNA sequencing are continually dropping (Erlich, 2015; Wetterstrand, 2018), whole-exome or even whole-genome sequencing are becoming a common practice in both research and routine clinical care (van El et al., 2013). Reasonably, methods offering massively parallel sequencing (MPS) matured during the last decade (Green et al., 2017; Shendure et al., 2017), both in technological and data processing aspects, including the bioinformatic pipelines that are generally used for identification of genomic variation causing phenotypic trait of interest. This step, moreover, shifted from simple counting of alleles among sequencing reads and their relative abundance to sophisticated probabilistic measures of uncertainty (Nielsen et al., 2011; Sandmann et al., 2017). These allows to combine prior information from reference data, such as allelic frequencies in the general population and linkage disequilibrium, with information about errors that may have been introduced during the whole process. Posterior probabilities as measure of uncertainty, are usually reported in a form of a genotype quality score. Moreover, several additional tests, such as local realignments and *post hoc* filtering, can be implemented to further improve the accuracy of genotype calls (Nielsen et al., 2011). Although high-quality and accurate single nucleotide variants (SNVs) and small insertion-deletion (indels) calls can be observed even in the level of individual sequencing reads (Budis et al., 2019a), there are still certain

* Corresponding author at: Institute for Clinical and Translational Research, Biomedical Research Center, Slovak Academy of Sciences, Dubravska cesta 9, 845 05, Bratislava, Slovakia.
E-mail address: jradvanszky@gmail.com (J. Radvanszky).

https://doi.org/10.1016/j.jbiotec.2019.04.013
Received 17 January 2019; Received in revised form 12 April 2019; Accepted 13 April 2019
Available online 15 April 2019
0168-1656/

Table 1
Comparison of results of variant calling for true variants using the original and the unified pipeline. Patients 6 and 7 are parents of patient 5 in whom a Sanger confirmed *de novo* deletion was identified. It is interesting that the deletion was identified in a very minority of the reads of the mother (1.1% and 0.8% in the original and unified pipeline, respectively), however, her possible mosaicity was not further studied. # - marked variants were not called automatically, rather identified following a visual inspection by the evaluator as possible variants.

Study Cohort	Patient	Variant description			MPS data – Original pipelines				MPS data – unified pipeline				Sanger validation
		Genomic Position (GRCh37)	Genotype	Gene	Genotype Quality Score	Alt. Allele Read Count	Total Read Count	Alt. Allele Frequency in reads	Genotype Quality Score	Alt. Allele Read Count	Total Read Count	Alt. Allele Frequency in reads	
1	1	chr7: g.143027970C > T	T/T	*CLCN1*	612.47	77	77	100.0	NA	NA	NA	NA	TRUE
1	2	chr7: g.143048771C > T	T/T	*CLCN1*	1123.17	461	463	99.6	NA	NA	NA	NA	TRUE
1	3	chr7: g.143036379 _143036392del TACCCTGCGGAGGC	TACCCTGCGGAGGC/ TACCCTGCGGAGGC	*CLCN1*	1347.2	163	167	97.6	NA	NA	NA	NA	TRUE
1	4	chr7: g.143048771C > T	C/T	*CLCN1*	216.54	75	152	49.3	NA	NA	NA	NA	TRUE
1	5	chr10: g.76789172delA	A/-	*KAT6B*	3498.88	104	237	43.9	236	79	163	48.5	TRUE
1	6	chr10: g.76789172delA	A/A	*KAT6B*	NA	2	186	1.1	0	1	127	0.8	TRUE
1	7	chr10: g.76789172delA	A/A	*KAT6B*	NA	0	143	0.0	0	0	105	0.0	TRUE
1	8	chr9: g.37784950C > G	G/G	*EXOSC3*	11772.77	338	349	96.8	254	171	177	96.6	TRUE
1	9	chr9: g.37784950C > G	G/G	*EXOSC3*	14146.77	384	384	100.0	265	185	185	100.0	TRUE
1	10	chr9: g.37784950C > G	G/G	*EXOSC3*	10646.77	295	296	99.7	250	150	150	100.0	TRUE
1	11	chr9: g.37784950C > G	C/G	*EXOSC3*	7155.77	227	446	50.9	233	104	206	50.5	TRUE
1	12	chr9: g.34635795C > T	C/T	*SIGMAR1*	1765.77	76	139	54.7	187	47	93	50.5	TRUE
1	13	chr9: g.34635795C > T	C/T	*SIGMAR1*	1069.77	48	102	47.1	169	38	72	52.8	TRUE
1	14	chr9: g.34635795C > T	T/T	*SIGMAR1*	4309.77	150	154	97.4	226	92	96	95.8	TRUE
1	15	chr9: g.34635795C > T	C/T	*SIGMAR1*	1423.77	68	164	41.5	175	38	95	40.0	TRUE
1	17	chr2: g.48026687 A > G	A/G	*MSH6*	6450.3	256	467	54.8	280	196	373	52.5	TRUE
1	17	chr13: 32972525C > T	C/T	*BRCA2*	1981.9	66	132	50.0	218	59	120	49.2	TRUE
1	18	chr14: g.23889446 _23889447insG	GG/GGG	*MYH7*	36.0	2	7	28.6	60	3	12	25.0	TRUE
1	19	chr22: g.37333973delC	C/-	*CSF2RB*	4501.73	145	291	49.8	233	95	196	48.5	TRUE
2	19	chr7: g.117199533 G > A	A/A	*CFTR*	2582.77	91	92	98.9	211	68	69	98.6	TRUE
2	19	chr7: g.117235055 T > G	G/G	*CFTR*	1887.77	70	70	100.0	192	52	52	100.0	TRUE
2	19	chr7: g.117307108 G > A	A/A	*CFTR*	5929.77	204	204	100.0	243	122	123	99.2	TRUE
2	19	chr7: g.117175505 A > G	A/G	*CFTR*	#	26	52	50.0	145	18	40	45.0	TRUE
2	19	chr7: g.117267511C > A	A/A	*CFTR*	#	14	14	100.0	116	10	10	100.0	TRUE
2	20	chr3: g.152018102 G > A	G/A	*MBNL1*	814.84	180	351	51.3	NA	NA	NA	NA	TRUE
2	1	chr3: g.152018102 G > A	G/A	*MBNL1*	385.94	128	252	50.8	NA	NA	NA	NA	TRUE

Table 2
Comparison of the results of variant calling for false positive and false negative variants using the original pipeline and the unified pipeline. Relatively loose criteria used in the original pipeline led to the identification of more false-positive variants (unmarked samples), while these were not identified using the more stringent criteria of the unified pipeline which, on the other hand, led to an increased number of false-negatives (marked by bold and outside borders). # - these variants were not called automatically, rather following a visual inspection by the evaluator. * – true homozygous chr7:117235197_117235198delAT variant called as heterozygous by the unified pipeline.

Study Cohort	Patient	Variant description			MPS data – Original pipelines				MPS data – unified pipeline				Sanger validation
		Genomic Position (GRCh37)	Genotype	Gene	Genotype Quality Score	Alt. Allele Read Count	Total Read Count	Alt. Allele Frequency in reads	Genotype Quality Score	Alt. Allele Read Count	Total Read Count	Alt. Allele Frequency in reads	
1	16	chr19: g.15308391 T > G	T/G	*NOTCH3*	#	3	8	37.5	0	1	7	14.3	FALSE
1	17	chr3: 37067125C > A	C/A	*MLH1*	12.66	7	26	26.9	0	1	12	8.3	FALSE
1	17	chr3: 37067129C > T	C/T	*MLH1*	17.75	7	28	25.0	11	2	14	14.3	FALSE
1	19	chr12: g.33049654delG	G/-	*PKP2*	50.73	5	27	18.5	0	1	10	10.0	FALSE
1	19	chr17: g.41197777C > T	C/T	*BRCA1*	61.77	12	70	17.1	36	2	30	6.7	FALSE
2	19	chr7: g.117188736C > A	C/A	*CFTR*	#	29	130	22.3	118	13	106	12.3	FALSE
2	19	chr7: g.117188797 A > G	A/G	*CFTR*	#	18	126	14.3	0	0	111	0.0	FALSE
2	19	chr7: g.117188850 G > T	G/T	*CFTR*	#	11	61	18.0	72	4	52	7.7	FALSE
1	19	chr14: g.23889446_23889447insG	TGG/TGGG	*MYH7*	71.73	4	16	25.0	0	1	11	9.1	TRUE
2	19	chr7: g.117188684 T > G	T/G	*CFTR*	#	32	87	36.8	104	8	28	28.6	TRUE
2	19	chr7: g.117235197_117235198delAT	-/-*	*CFTR*	#	23	23	100.0	123	11	16	68.8	TRUE

inconsistencies about the opinions regarding the necessity to confirm MPS based variant calling results using conventional methods. This is a specifically important question in clinical settings with medically relevant findings, in connection with which the American College of Medical Genetics and Genomics (ACMG) guidelines published in 2013 recommended "*that all disease-focused and/or diagnostic testing include confirmation of the final result using a companion technology*" (Rehm et al., 2013). Few years later, in 2015, the College of American Pathologists (CAP) suggested "*to give laboratories performing NGS-based assays flexibility in determining when confirmatory testing should be performed*" and "*how this testing is performed*" (Aziz et al., 2015). Published larger scale studies either claim the need for independent confirmation (Mu et al., 2016) or suggest that it is not necessary to perform confirmatory testing (Beck et al., 2016; Schenkel et al., 2016), but all of them seems to agree that the necessity and relevant threshold establishments should be defined in each laboratory and for each involved pipeline independently (Li, 2014), further emphasizing the need for specific validation processes (Matthijs et al., 2016). Thorough initial validation processes of entire pipelines are of high importance, especially because of actually existing high degree of variability in how different laboratories establish, combine, configure and validate their bioinformatic pipelines (Roy et al., 2018). In addition, reliability of variant calls seems to gain specific relevance in the upcoming era of complex interpretation of genomic results, going beyond conventional monogenic diagnostics towards evaluation of overall genomic mutational burden of individual patients or tumors (Morganti et al., 2019) and genome-wide polygenic risk scores for complex disorders (Khera et al., 2018). Reasonably, independent evaluation of hundreds or thousands of clinically relevant variants is not feasible for individual patients in routine clinical care.

The aim of this manuscript is, therefore, to report some of our specific findings which could help to get more familiar with some specific aspects of this topic, especially of issues connected to false-positive and false-negative findings having a variety of different sources of errors. These could help those professionals who are intended to perform pipeline validations or who are deciding about to perform variant confirmations using complementary methods.

2. Materials and methods

To characterize the validity of MPS based variant calling we performed several MPS based diagnostic tests in a timeframe of several years in which potentially clinically relevant findings underwent Sanger confirmatory sequencing. For the present study we retrospectively analyzed data primarily generated for these mentioned purposes. The final data set reported here, therefore, represents different data sets in which we simultaneously completed both MPS sequencing: IonTorrent PGM (Thermo Fisher Scientific, Waltham, MA) or Illumina MiSeq/NextSeq (Illumina, Inc. San Diego, CA) and Sanger sequencing (ABI Prism Genetic Analyzer 3130xl; Applied Biosystems, Foster City, CA) for the selected patients, genes and/or variants. In addition, original data analyses were performed using a heterogeneous combination of algorithms and pipelines (Suppl. Table 1), to which we will refer in the further text as original pipeline(s). In contrast to general procedures, to minimize the possibility of losing a relevant variant because of low sequencing quality, confirmatory testing of our variants was performed with no regard to the sequence coverage of the respective variants or specific quality metrics, resulting thus that even ambiguous variant calls having very low coverage or genotype quality scores were considered and evaluated.

2.1. Data set and study cohorts

The first part of our study, marked as study cohort 1, consisted of MPS based diagnostic genomic findings that were evaluated by Sanger sequencing. These findings were represented by 15 different potentially medically relevant DNA variants of 12 different genes in 19 individuals, including a previously reported case of a family trio where a clearly pathogenic *KAT6B* variant was identified in a patient and not in her parents, supporting a *de novo* origin of the variant (Radvanszky et al., 2017). Since some of the variants were validated in several individuals, the total number of validated variant positions in study cohort 1 reached 25 (Table 1; Table 2; Suppl. Table 1).

The second part of our study, marked as study cohort 2, consisted of simultaneous sequencing analysis, using both MPS and Sanger sequencing, of 9 exons of the *MBNL1* gene in 47 patients (having a clinical suspicion of myotonic dystrophy type 1 or 2) and the complete set of 27 exons of the *CFTR* gene in one patient, representing thus altogether 450 exons.

Human Gene Nomenclature Committee (HGNC) approved gene names, reference transcripts, variant descriptions and other additional information for both data cohorts are available in Suppl. Table 1.

2.2. Biological material

Genomic DNA was isolated from peripheral blood leukocytes by Puregene™ DNA Purification Kit (Qiagen, Hilden, Germany), while MPS data were generated using 3 different sequencing platforms and enrichment kits, as specified below, but in each case according to the protocols recommended by the manufacturers. Informed consent consistent with the Helsinki declaration was obtained from each subject before DNA testing. These consents contained an opt-in opt-out checkbox to the possible inclusion of the patient's data and biological samples for a research use in an anonymized form (for purposes including further methods standardizations and population analyses), while each of the included samples belonged strictly to patients giving consent to be included in such a study.

2.3. Massively parallel sequencing and data processing

Massively parallel sequencing was performed either on a MiSeq (Illumina) platform following enrichment and library preparation using a TruSightOne Sequencing Panel Oligos (Illumina), or on a NextSeq (Illumina) platform following enrichment and library preparation using a TruSight Exome library preparation kit (Illumina) (Suppl. Table 1). Patient 18 was sequenced by a commercial vendor (Macrogen Inc., Republic of Korea) using a SureSelectXT exome kit (Agilent Technologies, Santa Clara, CA) on an Illumina HiSeq platform. The variants of interest were identified and selected for this study following original data analysis using the original pipelines implementing different combinations of tools (Suppl. Table 1). Alignment of reads was, in each of the original pipelines, performed to the GRCh37/hg19 reference genome. Since the combinations of the tools were heterogeneous, to perform more detailed analyses of qualitative and quantitative measures all data generated using paired-end sequencing on Illumina platforms were reanalyzed with a unified pipeline. This pipeline consisted of trimming of low quality ends of reads and adapter remnants using Trimmomatic (Bolger et al., 2014), mapping to the reference genome GRCh38/hg38 with Bowtie2 (Langmead and Salzberg, 2012), sorting of reads according to mapped genomic positions with SAMTools Sort (Li et al., 2009), indexing using SAMTools Index (Li et al., 2009), marking PCR duplicates using Picard Tools (http://broadinstitute.github.io/picard/), indel realignment using the Genome Analysis Toolkit (GATK; Broad Institute) (DePristo et al., 2011) and variant calling using VarDict (Lai et al., 2016). Annotation and filtering of called variants from the generated Variant Call Format (VCF) files for Illumina data were performed using the GeneTalk (GeneTalk GmbH) (Kamphans and Krawitz, 2012) and/or Ingenuity Variant Analysis (Qiagen) tools. Variants/exons of interest were visualized and revised from Binary Alignment Map (BAM) files using SAMTools (Li et al., 2009) and/or the Integrative Genomics Viewer (IGV; Broad Institute) (Robinson et al., 2011; Thorvaldsdóttir et al., 2013).

For the *MBNL1* and *CLCN1* genes IonTorrent PGM (Life Technologies) sequencing was performed using an Ion 316 Chip and the corresponding sequencing kits. Enrichment for library preparation was performed using a custom HaloPlex Target Enrichment System (Agilent Technologies) designed through the SureDesign software. The original pipeline of IonTorrent data analysis included mapping to the reference genome GRCh37/hg19 and variant calling by the Torrent Suite Software v.3.2 (Life Technologies), implementing Torrent Mapping Alignment Program (TMAP) and GATK. Annotation and filtering of these data was performed using the SureCall (Agilent Technologies) and/or GeneTalk annotation tools.

2.4. Sanger sequencing and data processing

Genomic regions surrounding the variants of interest were sequenced applying the BigDye Terminator v3.1 Cycle Sequencing kit (Life Technologies). Following preamplification (oligonucleotide primers available in Suppl. Table 2) amplicons were purified using Exonuclease I/Shrimp Alkaline Phosphatase (USB) treatment and sequenced in both forward and reverse direction using the same primers as used for preamplification. After precipitation by sodium acetate (NaOAc) and ethanol the sequencing products were dissolved in Hi-Di formamide (Applied Biosystems), heat-denatured and separated by capillary electrophoresis on the genetic analyzer ABI Prism Genetic Analyzer 3130xl (Applied Biosystems) using a standard protocol. Electrophoretic data were collected and processed by the 3130 Data Collection and Sequencing Analysis v5.4 software (Applied Biosystems). For alignment to the corresponding reference sequences (Suppl. Table 1), variant calling and data visualization/revision the ChromasPro v1.6 (Technelysium Pty Ltd) software was used. To cope with allelic dropout: 1) all false calls were re-analyzed by a second round of Sanger sequencing; 2) primer binding sites were controlled for SNVs in the MPS data directly; 3) for the *PKP2* variant there was a second primer pair designed and evaluated, since one of the original primer binding sites were not covered by any MPS read; 4) the *NOTCH3* gene was independently analyzed using Sanger sequencing in an accredited diagnostic laboratory with no identified pathogenic allele reported (detailed results are, however, not available from this confirmatory analysis).

3. Results

3.1. Confirmatory testing results of the original data sets

The first validation cohort of our study consisted of 15 different suspected DNA variants, which were Sanger validated. In addition to this variant set, 11 different potential DNA variants were identified in the second validation cohort. When considering both cohorts, altogether 26 different DNA variants were validated, including 20 single nucleotide (SNVs) and 6 insertion/deletion variants (indels). However, since some of the variants were validated in several patients and in some patients there were several variants validated, the total number of genotyped and Sanger validated positions reached 37 (25 from the first and 12 from the second validation cohort) (Suppl. Table 1). From these, 29 were found to be true variants (Tables 1 and 2), while 8 were found to be false-positives even after a second round of Sanger confirmation (Table 2). The 8 false-positives consisted from 4 variants identified automatically by one of the originally used pipelines and 4 selected as suspicious positions based on manual evaluation (Table 2).

3.2. Results of the joint data set using the unified pipeline

To allow more detailed statistical evaluations, 31 variant positions, all but those identified in the *CLCN1* and *MBNL1* genes using single-end IonTorrent PGM reads, were re-analyzed from basic FASTQ files using a unified bioinformatics pipeline. From the 31 re-evaluated possible variant positions 20 genotypes were called correctly by the unified pipeline. In addition, each of the 8 previously identified false-positives were correctly ignored by the unified pipeline. On the other hand, we encountered also 3 true variants missed by this pipeline (false-negatives) (Table 2).

In general, all false findings were found to have lower genotype quality scores and total read counts when compared to true variants. On the other hand, both genotype quality scores, total read counts, alternative allele read counts and alternative allele frequencies in reads (allelic balance) showed significant overlapping regions between true and false variants (Fig. 1). In addition, comparisons of values generated by the original pipelines and the unified one revealed differences in data distribution of genotype quality score, total read counts, alternative allele read counts and also of allelic balance (Suppl. Fig. 1).

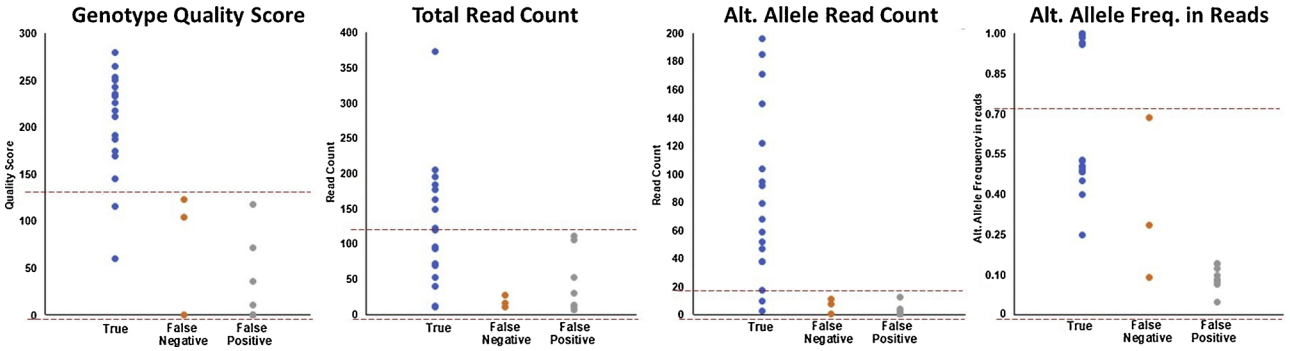


Fig. 1. Distribution of the main qualitative and quantitative indicators of the analyzed variants. Although the values were based on the unified method, the division of variants to groups "true", "false-negative" and "false-positive" reflects the combined results of both pipelines. Empty marks among false-positives show variants manually selected for validation following visual evaluation through a genomic viewer.

3.3. False-positive variants emerging from PCR duplicates

In 3 of them, chr17:41197777C > T (GRCh37; *BRCA1*; patient 19), chr12:33049654delG (GRCh37; *PKP2*; patient 19) and chr19:15,308,391 T > G (GRCh37; *NOTCH3*, patient 16), false positivity was indicated by an allelic disbalance that was further enhanced by the unified pipeline, allowing VarDict to correctly discard their presence. For the *BRCA1* and *PKP2* variants the artefacts originated from single pre-amplification clones, clearly identifiable by visual inspection (Fig. 2). Although the results after re-evaluation did not change on the level of BAM files (Fig. 2), such reads were flagged as PCR duplicates and removed by VarDict allowing to reduce their negative effect on genotype calling.

3.4. False-positive variants emerging from repeat associated errors

Another types of false-positive findings, associated to high error rate, were identified adjacent to the intronic polymorphic $(TA)_n(T)_n$ repeat motif of the *MLH1* gene. Patient 17 had a $(TA)_n(T)_n$ genotype distinct from the reference sequence manifesting in alignments in a form of deletions and substitutions (Fig. 3) which were, however, considered neither for genotyping (because of high indel error rates) nor for Sanger sequencing because of no associated clinical risk. Two nearby variant positions, chr3:37067125C > A and chr3:37067129C > T, however, met criteria to be called by the originally used pipeline as possibly clinically relevant SNVs. Neither re-analysis using the unified pipeline nor Sanger sequencing proved the presence of any of the possible variants. Manual control revealed remarkably high error rates especially in reads running through the repetitive region, while those which did not encompass the repeat motif remained without variants (Fig. 3). When visualized, similarly high error rate and similar error distribution was identified by both the original and the unified pipelines (Fig. 3). In addition, for the one of the most variable 136 bp read, BLAT (Kent, 2002) search did not identify highly identical regions neither in GRCh37 nor in GRCh38.

3.5. False-positive variants emerging from incomplete reference genome

Higher rate of false-positive findings were encountered also in the close proximity of a very similar intronic $(TG)_n(T)_n$ microsatellite motif of the *CFTR* gene. Results of patient 19, in addition to the true variant chr7:117188684 T > G caused directly by the $(TG)_n(T)_n$, contained

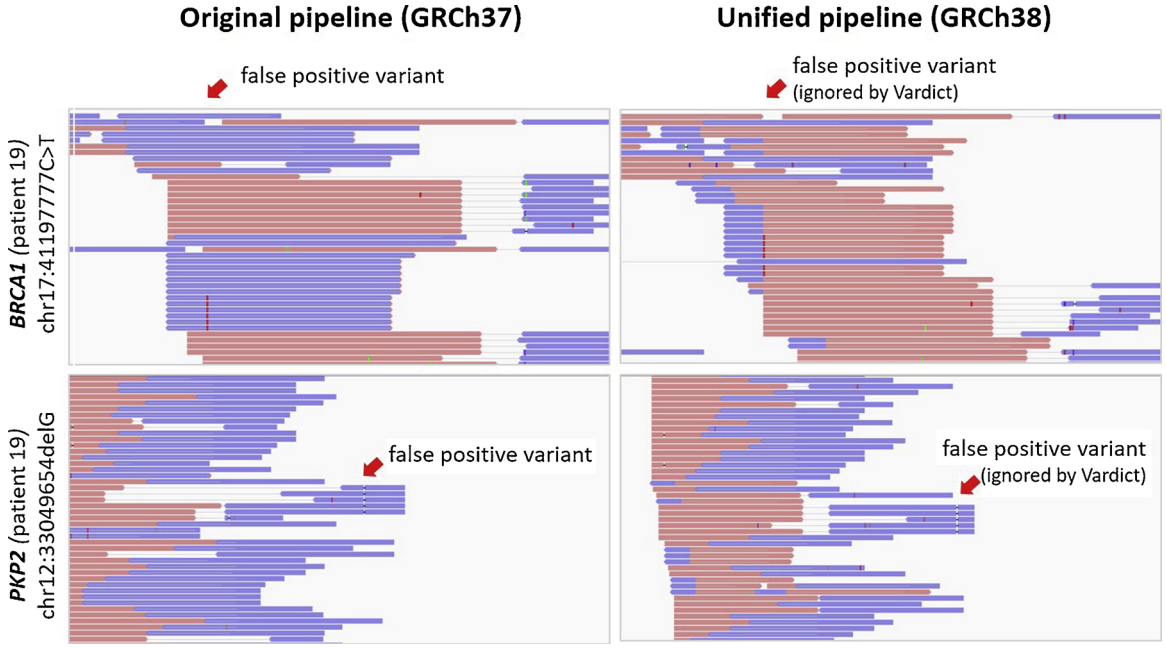


Fig. 2. Aligned reads, visualized in a genomic viewer (IGV), overlapping and surrounding the *BRCA1* and *PKP2* false-positive variants. Despite they were present even after the processing using the unified pipeline, they were correctly ignored by VarDict. Alignments showing reads of the same length, containing the respective variant calls, clearly indicate the PCR duplicate origin of the variant containing reads.

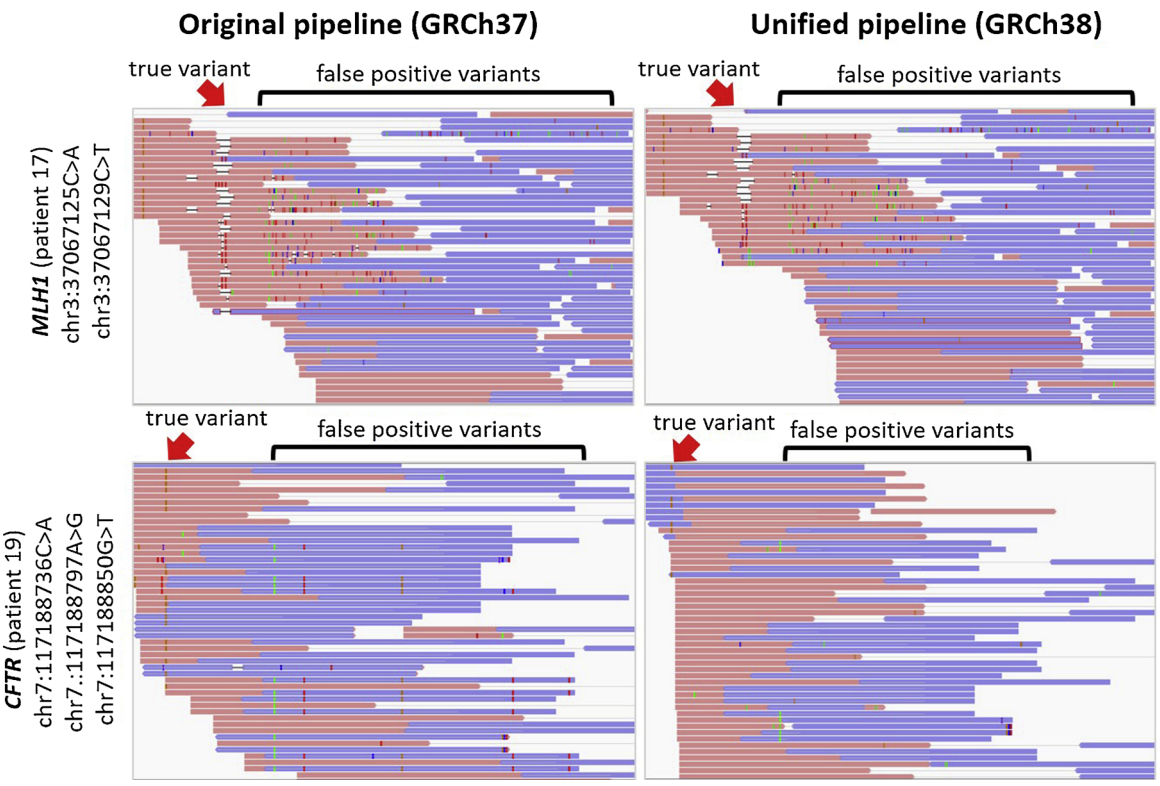


Fig. 3. Aligned reads, visualized in a genomic viewer (IGV), overlapping and surrounding the *CFTR* and *MLH1* false-positive variants. Both erroneous regions are associated, although in different ways, to a highly similar complex microsatellite region, $(TG)_n(T)_n$ in the *CFTR* gene and $(TA)_n(T)_n$ in the *MLH1* gene (in both cases located around the true variant in the left part of the alignment view). The false-positive variant region of the *CFTR* gene get relatively cleared by the unified pipeline, since the origin of the false variants stemmed in differences between the used reference genomes, rather than in the complex repeat motif itself. High error rate of the *MLH1* region, on the other hand, persisted even following alignment to GRCh38, since the false variants emerged, most probably, as amplification/sequencing artifacts directly associated to the complex repeat motif.

several false-positive variant positions from which three reached criteria to be called by the original pipeline, namely chr7:117188736C > A, chr7.:117188797 A > G and chr7:117188850 G > T (each according to GRCh37) (Table 2; Fig. 3).

Interestingly, read depths covering the respective region were found to be sufficient for reliable variant and genotype calling. Paired reads covering the region tended to align in close proximity to each other, as expected for correctly aligned reads, while they supported the presence of the variants when they overlapped each other as expected for true variants. Moreover, false-positives tended to occur in the same sequencing fragments/reads but never on the same fragments/reads with the true one, suggesting they are separate alleles occurring in *trans* phase with the true one. Slight allelic disbalance, relatively dense variant positions for one exon and associated reduced mapping quality of the variant containing reads, on the other hand, raised concerns with regard their true-positive nature that was subsequently negotiated by Sanger sequencing. Thorough revision of chromatograms, however, revealed slightly elevated background signal in the same positions which were found to be false-positive in MPS data (chromatograms not shown).

Re-evaluation of data using the unified pipeline, implementing mapping to GRCh38 instead of GRCh37 used in the original pipeline, led to a significant decrease in reads covering the respective region to 106 (from 130), 111 (from 126) and 52 (from 61), respectively for each of the variant. Interestingly, the most prominent part of this decrease involved reads containing the false-positive variants, lowering alternative allele frequencies to 12% (from 22.3%), 0% (from 14.3%) and 8% (from 18%), respectively (Table 2). These values did not met criteria for variant and genotype calling using the unified pipeline. Particularly, in contrast to the *MLH1* case, visualization revealed significantly cleared alignment pattern when compared to that from the original pipeline (Fig. 3).

Subsequent BLAT (Kent, 2002) search of the sequence of a 120 bp false variants containing read on GRCh37 led to the identification of three separate but highly homologous regions, having a potential to “share” a subset of reads between each other. One on chromosome 7 and two on chromosome 20, while the chromosome 7 and one of the chromosome 20´s regions both revealed 97.5% identity to the sequence of the variant containing read and the third one with a 96.7% identity. In contrast, the same search on GRCh38 found an additional 100% match on chromosome 20, more centromeric from the 97.5% matching region. Using the unified pipeline this GRCh38 unique region attracted most of the variant containing reads from the *CFTR* region to their correct position on chromosome 20, offering a perfect match for them, clearing out the *CFTR* gene sequence on chromosome 7 (Fig. 3).

3.6. False-negative variants

The three-identified false-negatives were also associated to allelic disbalance (Table 2). In contrast with the false-positives mentioned above, these were proved by Sanger sequencing to be true variants but were missed by the unified pipeline. The first one was a heterozygous single nucleotide insertion chr14:23889446_23889447insG (GRCh37) in the *MYH7* gene of patient 19 identified by the originally used pipeline (Fig. 4). The unified pipeline markedly reduced reads supporting the presence of the alternative allele (allelic balance reduced from 25% to 9%) resulting in that the variant did not met criteria to be called and genotyped. Subsequent manual control revealed that some of the original reads were mapped to another positions in GRCh38 while another’s were removed by VarDict as PCR duplicates. Interestingly, the same variant was identified and correctly called by both pipelines in patient 18. In this patient, despite that total read counts were lower

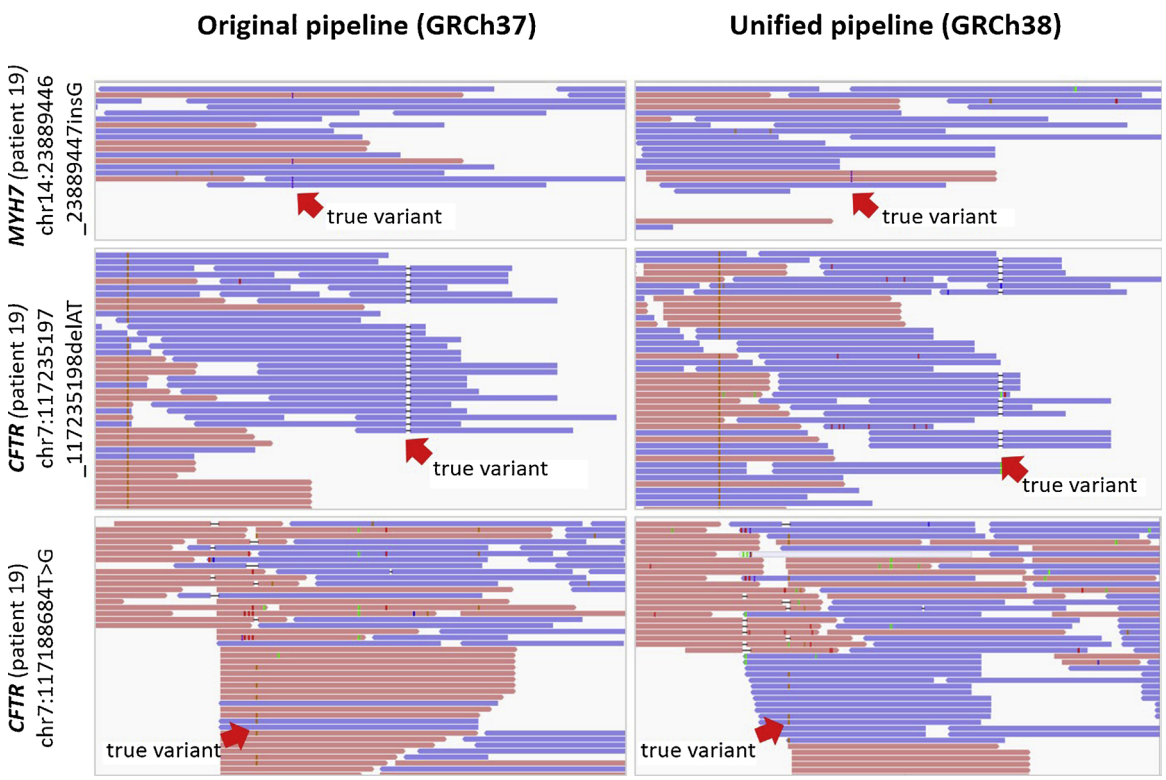


Fig. 4. Aligned reads, visualized in a genomic viewer (IGV), overlapping and surrounding the false-negative variants in the *MYH7* and *CFTR* genes. In case of the G insertion in the *MYH7* gene a reduction of reads, supporting the presence of the alternative allele (inserted), by the unified pipeline is visible even in the level of the alignments (allelic balance reduced from 25% to 9%). In contrast, alignments of the homozygous 2bp AT deletion containing reads of the *CFTR* gene seems to be not affected by the unified pipeline, although an additional built-in re-aligner of VarDict led to a false heterozygous genotype call. Notably, the shown alignment to the GRCh38 is captured before the additional disabling of the VarDict re-aligner. The bottom picture shows the reads containing the potentially pathogenic $(T)_5$ allele in the *CFTR* gene (that is part of the complex $(TG)_n(T)_n$ motif). The variant is represented here as a T to G change because of a specific constellation of the patient´s $(TG)_{10}(T)_7$ and $(TG)_{11}(T)_5$ alleles. The reason in this case was an issue connected to indel artefacts around the repeat region, possibly generated by PCR stutters, leading to an excess variability in reads that in turn posed a problem to VarDict with establishing an anticipated diploid genotype.

than those calculated for patient 19, allelic balance favored variant calling.

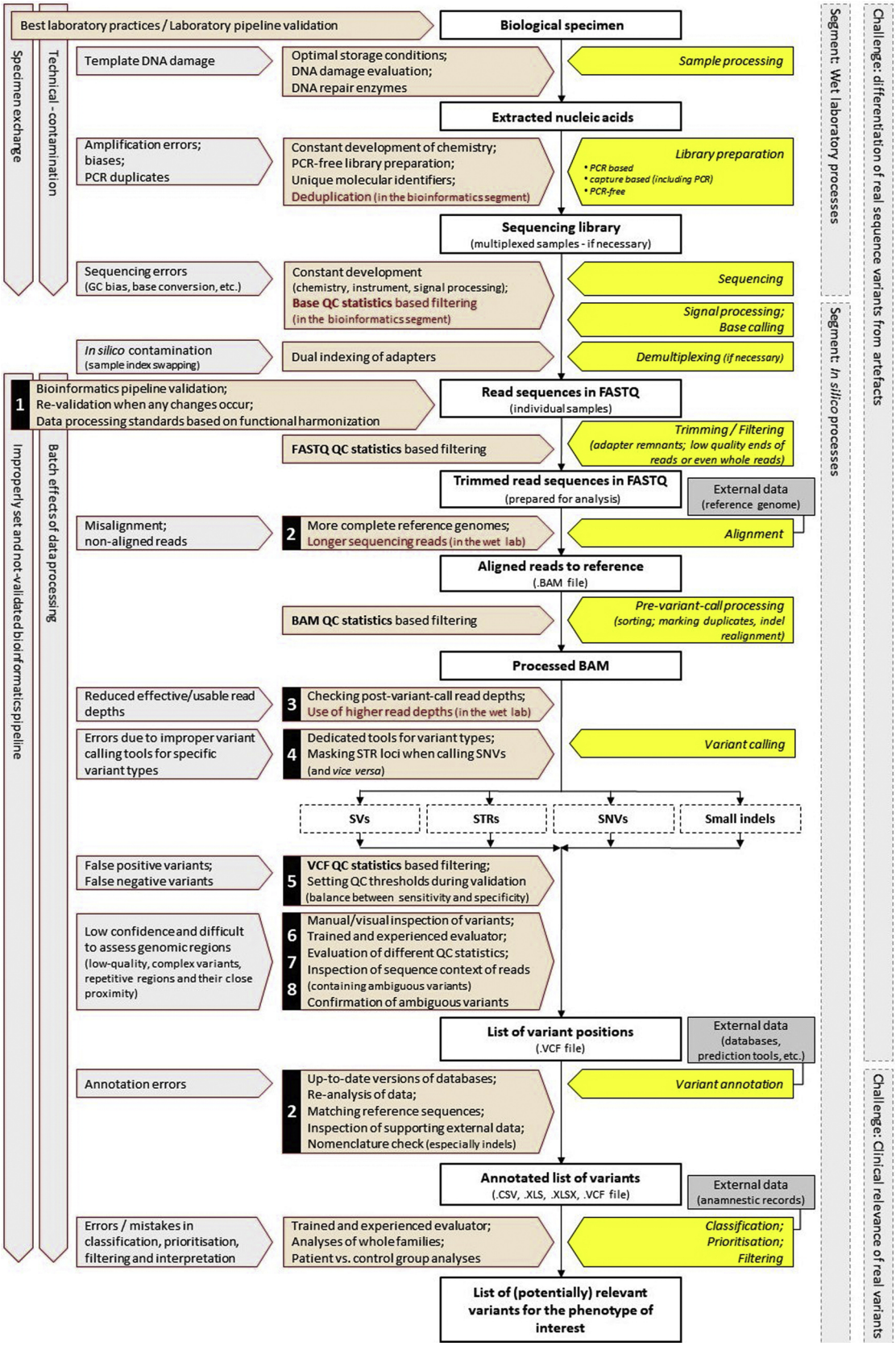
The second false-negative, an allele with a 2bp deletion in the *CFTR* gene (chr7:117235197_117235198delAT according to the GRCh37), was missed by the unified pipeline because of realignment problems resulting in a false heterozygous call instead of homozygous for the deleted allele. In this particular case GATK based realignment correctly placed the deletion containing reads to the reference sequence (Fig. 4), while subsequently a built-in re-aligner of VarDict mixed them up again resulting in an alternative allele frequency of 69% supporting a heterozygous genotype call, while the original pipeline identified a 100% alternative allele frequency. Additional disabling of the VarDict re-aligner corrected allelic balance to 94% allowing a correct homozygous genotype call.

The third false-negative variant was identified again in the *CFTR* gene of patient 19. A potentially pathogenic $(T)_5$ allele, because of a specific constellation of the patient´s $(TG)_{10}(T)_7$ and $(TG)_{11}(T)_5$ alleles manifesting as an SNV chr7:117188684 T > G (GRCh37), was missed. The reason in this case was an issue connected to indel artefacts around the repeat region, possibly generated by PCR stutters, leading to an excess variability in reads virtually creating a vertically complex variant (Fig. 4), that in turn posed a problem to VarDict with establishing an anticipated diploid genotype. Nomenclature errors and false-negative variants were reported to be associated with both horizontal and vertical complex variants (Roy et al., 2018).

4. Discussion

Although several possible error sources are the same for MPS based methods than for earlier generations of genetic analyses, both Sanger sequencing or other methods, the overall need to heavily rely on automated and large-scale evaluation/interpretation of results elevated challenges associated with identification and elimination of errors to completely new dimensions. Despite that there are also MPS method and platform specific error types (actually most commonly used methods are based on sequencing by synthesis and on Illumina platforms), with certain generalization it can be said that both possible error sources and respective prevention/elimination strategies can take place from the very first up to the last step of the whole process (schematically reviewed in Fig. 5). Such as before, specimen exchange and technical or carry-over contaminations can still happen during the whole wet laboratory procedure and are most effectively preventable by adhering with best laboratory practices and laboratory pipeline validations (Matthijs et al., 2016). Similarly, improperly set and not validated bioinformatics pipelines, together with batch effects in data can lead to incorrect results throughout the whole process of the *in silico* segment and can be minimized by thorough bioinformatics pipeline validations (Roy et al., 2018) and implementation of data processing standards (Regier et al., 2018). More specific errors might include template DNA damage, PCR amplification errors and biases, errors in the sequencing process itself (GC bias, base conversions, etc.), *in silico* contaminations (from sample index swapping in multisample analyses) and misalignments, all of which can lead to false-positive or false-negative variant calls. Up to the point of sequence specific signal generation by the sequencing platform, in addition to continuous development of chemistry, instrument and signal processing capabilities (Minoche et al., 2011), possible solutions can range from more optimal sample processing and storage (Aird et al., 2011; Costello et al., 2013),

through the evaluation of DNA damage (Costello et al., 2013; Park et al., 2017) or use of specific DNA repair enzymes (Chen et al., 2017), dual indexing of adapters (Kircher et al., 2012; Costello et al., 2018) and unique molecular identifiers (MacConaill et al., 2018; Kivioja et al., 2011), up to PCR-free library preparations and single molecule sequencing applications (Ameur et al., 2019). Downstream of signal



(caption on next page)

Fig. 5. Schematic representation of a massively parallel DNA sequencing based diagnostic process. Although the process is representative mainly for germline DNA sequence variant identification and interpretation, the majority of the steps is relevant also for other applications, such as somatic variants or transcriptome analyses. Shown are: main processing steps (yellow background, black outline); main products of each processing step (from the wet laboratory or *in silico* segment), such as sequencing libraries or files of certain type (white background, black outline); external data sources used in relevant processing steps (grey background, black outline); variant types (white background, black dashed outline); possible error sources (light grey background, red outline); possible preventive or elimination steps against specific errors (orange background, red outline); and errors having their sources and elimination strategies in different segments (fonts in red colour), such as errors generated in the sequencing process (segment of wet laboratory processes) which can be eliminated by bioinformatics filtering (segment of *in silico* processes). Numbers in boxes with black background refer to the bullet point summarization at the end of the discussion part. QC – quality control; BAM – binary alignment map; VCF – variant call format; SVs – structural variants; STRs – short tandem repeats; SNVs – single nucleotide variants; CSV - comma-separated values; XLS/XLSX - Microsoft Excel spreadsheet formats (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article).

generation and base calling, *in silico* processes need to rely on filtering and read trimming steps based on different quality control statistics (both in FASTQ, BAM and VCF), and on effective and reliable variant calling (Pfeifer, 2017) with optional validation of relevant or ambiguous variants using independent methods (for reference see citations in the Introduction). Moreover, even if real sequence variants are effectively distinguished from artefacts, errors in further downstream *in silico* processes, such as nomenclature generation, annotation, classification, prioritization and final variant filtering for reporting might have also implications on the interpretation of MPS based results. These can be most effectively minimized by nomenclature revisions, exploitation of up-to-date versions of databases and reference data sets, evaluation of possibly relevant findings and connected external metadata by properly trained and qualified personnel (Richards et al., 2015; Matthijs et al., 2016; Roy et al., 2018), as well as by the use of mandatory data elements during results reporting (Swaminathan et al., 2017).

From all the above-mentioned possible error sources the challenge represented by the differentiation of real sequence variants from sequencing artefacts is still one of the mostly debated, especially in clinical settings. An anticipated validity of MPS identified DNA variants is generally based on different qualitative and quantitative measures of reads covering genomic regions under consideration. Different read depth statistics, together with specialized genotype quality scores are, therefore, of high relevance when estimating validity of called variants directly from MPS data. Generally, the validity of an MPS variant with high quality measures is not questioned, especially in case of SNVs (Strom et al., 2014). Comparisons of genotype quality scores across different pipelines are, however, not recommended (Strom et al., 2014), since they are relative measures of the applied variant calling tool. Moreover, our results showed that even basic quantitative metrics such as total read counts, alternative allele read counts and allelic balance tend to depend on the used pipeline. Encountered inconsistencies may result from tools and specific filters implemented in the used pipeline, as well as from different versions of reference genomes (Li, 2014).

Specifically, when considering genotype quality scores, total read counts, alternative read counts and allelic balance in our data set, the distribution pattern of true variants was found to be different from that of false-negatives and false-positives, however, still with considerable overlaps between the groups (Fig. 1). This make it impossible to set defined thresholds to detect all true variants and sort out all false-positive ones, especially in the lower parts of the distribution pattern of true variants, pointing to the importance of taking care when lower-quality variants are evaluated.

To get further insight into the sources of our false results and to define some indications applicable to increase attention about the possible false or true nature of ambiguous variants, we decided to characterize our false results in more details. False-positive variants identified by our original approach were found to have quite different origins. Besides extremely low coverage (*NOTCH3*), conventional causes included also single pre-amplification clones induced artifacts (*BRCA1* and *PKP2*), which were found to be accompanied with certain degree of allelic disbalance in our data set. These latter error sources are generally created during library preparation (Bentley et al., 2008) and are clearly identifiable by visual inspection in aligned reads. Automatic flagging and filtering of such clonal variant containing reads, when library preparation is not amplicon based, may further enhance allelic disbalance that was the case also in our unified pipeline in which VarDict was able to correctly discard the presence of the above mentioned originally identified false-positive variants mitigating thus their effect on the results. Tagging DNA library molecules prior to library amplification by unique molecular identifiers (i.e. molecular barcodes) incorporated into the sequencing adapters can be used post-sequencing to aid in better identification of such PCR duplicates (MacConaill et al., 2018; Kivioja et al., 2011).

Completely different examples of false-positive findings were identified adjacent to two highly similar complex microsatellite loci, namely the $(TA)_n(T)_n$ repeat motif in the *MLH1* gene and the $(TG)_n(T)_n$ motif of the *CFTR*. In the first one, high error rate was identified most probably because of repeat-associated synthesis and subsequent sequencing introduced errors. This is well in line with previous reports describing higher susceptibility of indels to PCR artifacts when compared to SNVs (Li, 2014). Importantly, these did not disappear directly when processed with the unified pipeline, although this pipeline was able to cope with them and ignore false variants generated by them. In the *CFTR* gene, in contrast, error pattern was explained with a less complete version of the human reference genome, to which reads were originally aligned, and largely disappeared when reads were aligned to a newer reference genome version offering decoy sequences for the originally mismapped reads. In general, completing of GRCh38 led to a great reduction of overall and centromere related unassembled regions (N's) of the reference genome when compared to GRCh37 (Li and Freudenberg, 2014) that was previously shown to have impact not only on read mapping but, reasonably, also on variant calling (Budis et al., 2019a; Li, 2014). Our conclusion about the mentioned possible error sources seems to be proven also by findings that the *CFTR* false-positives tend to appear in the majority of samples, while the *MLH1* erroneous pattern did not appeared in samples in which sequencing reads did not run through the repeat region itself (data not shown).

The above mentioned false-positives were, in each of our cases, effectively recognized and sorted out by the unified pipeline, although, the same settings led to a missed variant in the same repeat associated region of the *CFTR* gene to which some of the false positives were connected. In general, reliability of variant calling from MPS data sets tend to increase with increasing stringency of the implemented criteria (Beck et al., 2016). This decreases the number of false-positive variants in the variant list but simultaneously increases the number of missed true-positive ones that is clearly visible in our data set too. Highly stringent criteria therefore tend to reduce sensitivity and diagnostic yield of the assays with a potential to have impact on the patient's health and clinical management. On the other hand, it is significantly easier to manually filter out possibly relevant false-positive variants, generally having signs of false-positivity (in case of lower stringency), than manually identify missed true-variants (in case of higher stringency). Use of different stringency criteria for hard-filtering and soft-filtering during variant calling may help to cope with this dilemma, especially when followed by checking also those variants which did not pass soft-filtering but are still present in the VCF.

With regard to variant types, except of one insertion, false-positives were generally SNVs in our data set, while deeper investigation revealed that 2 of them originated from microsatellite loci induced errors.

Identified false-negatives were exclusively indels, or were at least derived from indels, located in microsatellite repeat regions. It is well known that general methods used for variant calling are not the most suitable tools for microsatellite genotyping, since the reliability of their results depend on correct alignments of reads to a reference genome (McKenna et al., 2010; Tae et al., 2013; Willems et al., 2017). Read mapping with standard methods may not align reads to complex or expanded microsatellite loci in the first place, missing thus the genotype, particularly in complex genomic regions or when the length of an allele highly deviates from the reference number of repeats (Budis et al., 2019b). This issue is generally addressed by specialized tools designed to genotype microsatellite loci using alignment-based (Willems et al., 2017) or alignment-free approaches (Budis et al., 2019b). According to our results, however, in addition to problems with genotyping of the microsatellite loci themselves, these regions can introduce higher error rates even in their close vicinity, as demonstrated in our *MLH1* false-positive variants that should be kept in mind when evaluating such regions.

Finally, whether required or not, when considering to perform Sanger validation there are several points for consideration. Although Sanger sequencing is accounted for a gold standard in DNA sequencing it cannot be considered ultimate mainly due to errors residing in the amplification step, natural variance of the sequence, as well as due to polymerase slippage at low complexity sequences like simple repeats and homopolymers (Kircher and Kelso, 2010). Even if our small scale study did not revealed problems with Sanger results, several studies proved that MPS results are at least as accurate as Sanger based assays, suggesting that they could be used even as stand-alone tests (Baudhuin et al., 2015; McCourt et al., 2013; Sikkema-Raddatz et al., 2013; Strom et al., 2014). Moreover, a single round of Sanger sequencing is more likely to incorrectly refute a true-positive variant identified using MPS than to correctly identify a false-positive variant from an MPS based test, suggesting that Sanger confirmation could not simply make the analyses costly and time-consuming but also less sensitive, especially for variants having robust quality scores (Beck et al., 2016). There are, however, also special cases, which could be ambiguous in technological terms but convincing in biomedical context (known pathogenic variants fitting well into the clinical symptoms) which should be independently confirmed.

Based on our experience and presented results, together with other results published up to date and discussed in this manuscript, we suggest the following minimal points of consideration: 1) data processing standards should be implemented and each bioinformatics pipeline should be thoroughly evaluated following its setup, or even after its small changes or modifications; 2) use of more complete reference genomes, up-to-date versions of databases and newly developed pipelines, in connection with re-analysis of older data when required, might be specifically beneficial for alignment, variant calling and subsequent variant annotation; 3) since even variant callers can remove/ignore reads, coverage statistics calculated following the alignment step may not represent correctly real coverage used for final variant calling and genotyping; 4) microsatellite loci should be masked from conventional variant calling of SNVs and small indels (and vice versa), and genotyped independently using dedicated tools; 5) before deciding about the necessity of independent confirmation of variants, relevant threshold establishments should be defined in each laboratory and for each involved pipeline independently, following uniform re-evaluation of whole validation data sets, since optimal thresholds of quantitative and qualitative metrics may be strongly caller dependent; 6) based on a professional judgement of an appropriately trained and skilled evaluator genotypes with sufficiently high qualitative and quantitative metrics seems to not require independent confirmation, especially if quality thresholds are defined empirically for the used analytical system; 7) in special cases, decreasing the stringency of variant calling and/or manual confirmation of ambiguous regions and calls, with subsequent independent confirmation of low-quality variants, indels, complex variants, and variants detected in their close proximity, seems to be still necessary, and may be beneficial for variant discovery and/or correct genotyping – based on a professional judgement of an appropriately trained and skilled evaluator; 8) reads spanning polymorphic microsatellite loci, or other hard to sequence regions, should be treated with higher attention, possibly also ignored, or at least included and thoroughly evaluated in general validation processes of bioinformatics pipelines.

5. Disclosure statements

The authors declare no conflict of interest. All authors approved the final article. Authors contributions to this work: ZK - performed data summarization and evaluations, validation experiments using conventional methods and wrote selected parts of the manuscript; MG – performed confirmatory testing using Sanger sequencing; EN – performed massively parallel sequencing experiments (IonTorrent platform); MH - performed massively parallel sequencing experiments (MiSeq platform) and selected Sanger verifications; MH - performed massively parallel sequencing experiments (NextSeq platform) and selected Sanger verifications; JB – supervised the bioinformatics team, created the pipelines used, and wrote/revised selected parts of the manuscript; RH – performed bioinformatics analyses and data verifications; JG – performed bioinformatics analyses and data verifications; FD – performed bioinformatics analyses and data verifications; LK – supervised the genetic diagnostics team, involved in the diagnostics and evaluation of variants; TS – supervised the genomics team, involved in the diagnostics and evaluation of variants; JR – coordinated the work, evaluated the results, prepared and supervised the manuscript, involved in the diagnostics and evaluation of variants.

6. Declaration of interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported by projects “REVOGENE – Research centre for molecular genetics” (ITMS 26240220067) (55% of the costs) and ”DIARET_SK - Research Centre for Severe Diseases and Related Complications” (ITMS 26240120038) (40% of the costs), both supported by the Operational Programme Research and Development funded by the European Research and Development Fund (ERDF). Partial support was obtained also from a grant provided by the Scientific Grant Agency of the Ministry of Education, Science, Research and Sport of the Slovak Republic and the Slovak Academy of Sciences (VEGA 2/0115/15) (5% of the costs).

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.jbiotec.2019.04.013.

References

Aird, D., Ross, M.G., Chen, W.-S., Danielsson, M., Fennell, T., Russ, C., Jaffe, D.B., Nusbaum, Ch, Gnirke, A., 2011. Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. Genome Biol. 12 (2), R18. https://doi.org/10.1186/gb-2011-12-2-r18.

Ameur, A., Kloosterman, W.P., Hestand, M.S., 2019. Single-Molecule Sequencing: Towards Clinical Applications. Trends Biotechnol. 37 (1), 72–85. https://doi.org/10.1016/j.tibtech.2018.07.013.

Aziz, N., Zhao, Q., Bry, L., Driscoll, D.K., Funke, B., Gibson, J.S., Grody, W.W., Hegde, M.R., Hoeltge, G.A., Leonard, D.G.B., Merker, J.D., Nagarajan, R., Palicki, L.A., Robetorye, R.S., Schrijver, I., Weck, K.E., Voelkerding, K.V., 2015. College of American Pathologists' laboratory standards for next-generation sequencing clinical

tests. Arch. Pathol. Lab. Med. 139, 481–493. https://doi.org/10.5858/arpa.2014-0250-CP.
Baudhuin, L.M., Lagerstedt, S.A., Klee, E.W., Fadra, N., Oglesbee, D., Ferber, M.J., 2015. Confirming variants in next-generation sequencing panel testing by Sanger sequencing. J. Mol. Diagn. 17, 456–461. https://doi.org/10.1016/j.jmoldx.2015.03.004.
Beck, T.F., Mullikin, J.C., NISC Comparative Sequencing Program, Biesecker, L.G., 2016. Systematic evaluation of Sanger validation of next-generation sequencing variants. Clin. Chem. 62, 647–654. https://doi.org/10.1373/clinchem.2015.249623.
Bentley, D.R., Balasubramanian, S., Swerdlow, H.P., Smith, G.P., Milton, J., Brown, C.G., et al., 2008. Accurate whole human genome sequencing using reversible terminator chemistry. Nature 456, 53–59. https://doi.org/10.1038/nature07517.
Bolger, A.M., Lohse, M., Usadel, B., 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30, 2114–2120. https://doi.org/10.1093/bioinformatics/btu170.
Budis, J., Gazdarica, J., Radvanszky, J., Harsanyova, M., Gazdaricova, I., Strieskova, L., Frno, R., Duris, F., Minarik, G., Sekelska, M., Szemes, T., 2019a. Non-invasive prenatal testing as a valuable source of population specific allelic frequencies. J. Biotechnol Submitted – Accepted with minor revisions.
Budis, J., Kucharik, M., Duris, F., Gazdarica, J., Zrubcova, M., Ficek, A., Szemes, T., Brejová, B., Radvanszky, J., 2019b. Dante: Genotyping of Known Complex and Expanded Short Tandem Repeats. Bioinformatics Accepted - ahead of printhttps://doi.org/10.1093/bioinformatics/bty791.
Chen, L., Liu, P., Evans, T.C., Ettwiller, L.M., 2017. DNA damage is a pervasive cause of sequencing errors, directly confounding variant identification. Science 355 (6326), 752–756. https://doi.org/10.1126/science.aai8690.
Costello, M., Pugh, T.J., Fennell, T.J., Stewart, C., Lichtenstein, L., Meldrim, J.C., Fostel, J.L., Friedrich, D.C., Perrin, D., Dionne, D., Kim, S., Gabriel, S.B., Lander, E.S., Fisher, S., Getz, G., 2013. Discovery and characterization of artifactual mutations in deep coverage targeted capture sequencing data due to oxidative DNA damage during sample preparation. Nucleic Acids Res. 41 (6), e67. https://doi.org/10.1093/nar/gks1443.
Costello, M., Fleharty, M., Abreu, J., Farjoun, Y., Ferriera, S., Holmes, L., Granger, B., Green, L., Howd, T., Mason, T., Vicente, G., Dasilva, M., Brodeur, W., DeSmet, T., Dodge, S., Lennon, N.J., Gabrie, S., 2018. Characterization and remediation of sample index swaps by non-redundant dual indexing on massively parallel sequencing platforms. BMC Genomics 19 (1), 332. https://doi.org/10.1186/s12864-018-4703-0.
DePristo, M.A., Banks, E., Poplin, R., Garimella, K.V., Maguire, J.R., Hartl, C., Philippakis, A.A., del Angel, G., Rivas, M.A., Hanna, M., McKenna, A., Fennell, T.J., Kernytsky, A.M., Sivachenko, A.Y., Cibulskis, K., Gabriel, S.B., Altshuler, D., Daly, M.J., 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat. Genet. 43, 491–498. https://doi.org/10.1038/ng.806.
Erlich, Y., 2015. A vision for ubiquitous sequencing. Genome Res. 25, 1411–1416. https://doi.org/10.1101/gr.191692.115.
Green, E.D., Rubin, E.M., Olson, M.V., 2017. The future of DNA sequencing. Nature 550, 179–181. https://doi.org/10.1038/550179a.
Kamphans, T., Krawitz, P.M., 2012. GeneTalk: an expert exchange platform for assessing rare sequence variants in personal genomes. Bioinformatics 28, 2515–2516. https://doi.org/10.1093/bioinformatics/bts462.
Kent, W.J., 2002. BLAT—The BLAST-Like alignment tool. Genome Res. 12, 656–664. https://doi.org/10.1101/gr.229202.
Khera, A.V., Chaffin, M., Aragam, K.G., Haas, M.E., Roselli, C., Choi, S.H., Natarajan, P., Lander, E.S., Lubitz, S.A., Ellinor, P.T., Kathiresan, S., 2018. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. Nat. Genet. 50 (9), 1219–1224. https://doi.org/10.1038/s41588-018-0183-z.
Kircher, M., Kelso, J., 2010. High-throughput DNA sequencing–concepts and limitations. Bioessays 32, 524–536. https://doi.org/10.1002/bies.200900181.
Kircher, M., Sawyer, S., Meyer, M., 2012. Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. Nucleic Acids Res. 40 (1), e3. https://doi.org/10.1093/nar/gkr771.
Kivioja, T., Vähärautio, A., Karlsson, K., Bonke, M., Enge, M., Linnarsson, S., Taipale, J., 2011. Counting absolute numbers of molecules using unique molecular identifiers. Nat. Methods 9 (1), 72–74. https://doi.org/10.1038/nmeth.1778.
Lai, Z., Markovets, A., Ahdesmaki, M., Chapman, B., Hofmann, O., McEwen, R., Johnson, J., Dougherty, B., Barrett, J.C., Dry, J.R., 2016. VarDict: a novel and versatile variant caller for next-generation sequencing in cancer research. Nucleic Acids Res. 44, e108. https://doi.org/10.1093/nar/gkw227.
Langmead, B., Salzberg, S.L., 2012. Fast gapped-read alignment with Bowtie 2. Nat. Methods 9, 357–359. https://doi.org/10.1038/nmeth.1923.
Li, H., 2014. Toward better understanding of artifacts in variant calling from high-coverage samples. Bioinformatics 30, 2843–2851. https://doi.org/10.1093/bioinformatics/btu356.
Li, W., Freudenberg, J., 2014. Characterizing regions in the human genome unmappable by next-generation-sequencing at the read length of 1000 bases. Comput. Biol. Chem. 53 (Pt A), 108–117. https://doi.org/10.1016/j.compbiolchem.2014.08.015.
Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., 1000 Genome Project Data Processing Subgroup, 2009. The sequence Alignment/Map format and SAMtools. Bioinformatics 25, 2078–2079. https://doi.org/10.1093/bioinformatics/btp352.
MacConaill, L.E., Burns, R.T., Nag, A., Coleman, H.A., Slevin, M.K., Giorda, K., Light, M., Lai, K., Jarosz, M., McNeill, M.S., Ducar, M.D., Meyerson, M., Thorner, A.R., 2018. Unique, dual-indexed sequencing adapters with UMIs effectively eliminate index cross-talk and significantly improve sensitivity of massively parallel sequencing. BMC Genomics 19 (1), 30. https://doi.org/10.1186/s12864-017-4428-5.
Matthijs, G., Souche, E., Alders, M., Corveleyn, A., Eck, S., Feenstra, I., Race, V., Sistermans, E., Sturm, M., Weiss, M., Yntema, H., Bakker, E., Scheffer, H., Bauer, P., 2016. Guidelines for diagnostic next-generation sequencing. Eur. J. Hum. Genet. 24, 1515. https://doi.org/10.1038/ejhg.2015.226.
McCourt, C.M., McArt, D.G., Mills, K., Catherwood, M.A., Maxwell, P., Waugh, D.J., Hamilton, P., O'Sullivan, J.M., Salto-Tellez, M., 2013. Validation of next generation sequencing technologies in comparison to current diagnostic gold standards for BRAF, EGFR and KRAS mutational analysis. PLoS One 8, e69604. https://doi.org/10.1371/journal.pone.0069604.
McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., DePristo, M.A., 2010. The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 20, 1297–1303. https://doi.org/10.1101/gr.107524.110.
Minoche, A.E., Dohm, J.C., Himmelbauer, H., 2011. Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and genome analyzer systems. Genome Biol. 12 (11), R112. https://doi.org/10.1186/gb-2011-12-11-r112.
Morganti, S., Tarantino, P., Ferraro, E., D'Amico, P., Viale, G., Trapani, D., Duso, B.A., Curigliano, G., 2019. Complexity of genome sequencing and reporting: Next generation sequencing (NGS) technologies and implementation of precision medicine in real life. Crit. Rev. Oncol. Hematol. 133, 171–182. https://doi.org/10.1016/j.critrevonc.2018.11.008.
Mu, W., Lu, H.-M., Chen, J., Li, S., Elliott, A.M., 2016. Sanger confirmation is required to achieve optimal sensitivity and specificity in next-generation sequencing panel testing. J. Mol. Diagn. 18, 923–932. https://doi.org/10.1016/j.jmoldx.2016.07.006.
Nielsen, R., Paul, J.S., Albrechtsen, A., Song, Y.S., 2011. Genotype and SNP calling from next-generation sequencing data. Nat. Rev. Genet. 12, 443–451. https://doi.org/10.1038/nrg2986.
Park, G., Park, J.K., Shin, S.-H., Jeon, H.-J., Kim, N.K.D., Kim, Y.J., Shin, H.-T., Lee, H., Lee, K.H., Son, D.-S., Park, W.-Y., Park, D., 2017. Characterization of background noise in capture-based targeted sequencing data. Genome Biol. 18 (1). https://doi.org/10.1186/s13059-017-1275-2.
Pfeifer, S.P., 2017. From next-generation resequencing reads to a high-quality variant data set. Heredity 118, 111–124. https://doi.org/10.1038/hdy.2016.102.
Radvanszky, J., Hyblova, M., Durovcikova, D., Hikkelova, M., Fiedler, E., Kadasi, L., Turna, J., Minarik, G., Szemes, T., 2017. Complex phenotypes blur conventional borders between Say-Barber-Biesecker-Young-Simpson syndrome and genitopatellar syndrome. Clin. Genet. 91, 339–343. https://doi.org/10.1111/cge.12840.
Regier, A.A., Farjoun, Y., Larson, D.E., Krasheninina, O., Kang, H.M., Howrigan, D.P., Chen, B.J., Kher, M., Bank, E., Ames, D.C., English, A.C., Li, H., Xing, J., Zhang, Y., Matise, T., Abecasis, G.R., Salerno, W., Zody, M.C., Neale, B.M., Hall, I.M., 2018. Functional equivalence of genome sequencing analysis pipelines enables harmonized variant calling across human genetics projects. Nat. Commun. 9 (1), 4038. https://doi.org/10.1038/s41467-018-06159-4.
Rehm, H.L., Bale, S.J., Bayrak-Toydemir, P., Berg, J.S., Brown, K.K., Deignan, J.L., Friez, M.J., Funke, B.H., Hegde, M.R., Lyon, E., Working Group of the American College of Medical Genetics and Genomics Laboratory Quality Assurance Commitee, 2013. ACMG clinical laboratory standards for next-generation sequencing. Genet. Med. 15, 733–747. https://doi.org/10.1038/gim.2013.92.
Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., Grody, W.W., Hegde, M., Lyon, E., Spector, E., Voelkerding, K., Rehm, H.L., ACMG Laboratory Quality Assurance Committee, 2015. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genet. Med. 17 (5), 405–424. https://doi.org/10.1038/gim.2015.30.
Robinson, J.T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E.S., Getz, G., Mesirov, J.P., 2011. Integrative genomics viewer. Nat. Biotechnol. 29, 24–26. https://doi.org/10.1038/nbt.1754.
Roy, S., Coldren, C., Karunamurthy, A., Kip, N.S., Klee, E.W., Lincoln, S.E., Leon, A., Pullambhatla, M., Temple-Smolkin, R.L., Voelkerding, K.V., Wang, C., Carter, A.B., 2018. Standards and guidelines for validating next-generation sequencing bioinformatics pipelines: a joint recommendation of the Association for Molecular Pathology and the College of American Pathologists. J. Mol. Diagn. 20, 4–27. https://doi.org/10.1016/j.jmoldx.2017.11.003.
Sandmann, S., de Graaf, A.O., Karimi, M., van der Reijden, B.A., Hellström-Lindberg, E., Jansen, J.H., Dugas, M., 2017. Evaluating variant calling tools for non-matched next-generation sequencing data. Sci. Rep. 7, 43169. https://doi.org/10.1038/srep43169.
Schenkel, L.C., Kerkhof, J., Stuart, A., Reilly, J., Eng, B., Woodside, C., Levstik, A., Howlett, C.J., Rupar, A.C., Knoll, J.H.M., Ainsworth, P., Waye, J.S., Sadikovic, B., 2016. Clinical next-generation sequencing pipeline outperforms a combined approach using sanger sequencing and multiplex ligation-dependent probe amplification in targeted gene panel analysis. J. Mol. Diagn. 18, 657–667. https://doi.org/10.1016/j.jmoldx.2016.04.002.
Shendure, J., Balasubramanian, S., Church, G.M., Gilbert, W., Rogers, J., Schloss, J.A., Waterston, R.H., 2017. DNA sequencing at 40: past, present and future. Nature 550, 345–353. https://doi.org/10.1038/nature24286.
Sikkema-Raddatz, B., Johansson, L.F., de Boer, E.N., Almomani, R., Boven, L.G., van den Berg, M.P., van Spaendonck-Zwarts, K.Y., van Tintelen, J.P., Sijmons, R.H., Jongbloed, J.D.H., Sinke, R.J., 2013. Targeted next-generation sequencing can replace Sanger sequencing in clinical diagnostics. Hum. Mutat. 34, 1035–1042. https://doi.org/10.1002/humu.22332.
Strom, S.P., Lee, H., Das, K., Vilain, E., Nelson, S.F., Grody, W.W., Deignan, J.L., 2014. Assessing the necessity of confirmatory testing for exome-sequencing results in a clinical molecular diagnostic laboratory. Genet. Med. 16, 510–515. https://doi.org/10.1038/gim.2013.183.
Swaminathan, R., Huang, Y., Astbury, C., Fitzgerald-Butt, S., Miller, K., Cole, J., Bartlett, C., Lin, S., 2017. Clinical exome sequencing reports: current informatics practice and future opportunities. J. Am. Med. Inform. Assoc. 24 (6), 1184–1191. https://doi.org/

10.1093/jamia/ocx048.
Tae, H., McMahon, K.W., Settlage, R.E., Bavarva, J.H., Garner, H.R., 2013. ReviSTER: an automated pipeline to revise misaligned reads to simple tandem repeats. Bioinformatics 29, 1734–1741. https://doi.org/10.1093/bioinformatics/btt277.
Thorvaldsdóttir, H., Robinson, J.T., Mesirov, J.P., 2013. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief. Bioinform. 14, 178–192. https://doi.org/10.1093/bib/bbs017.
van El, C.G., Cornel, M.C., Borry, P., Hastings, R.J., Fellmann, F., Hodgson, S.V., Howard, H.C., Cambon-Thomsen, A., Knoppers, B.M., Meijers-Heijboer, H., Scheffer, H., Tranebjaerg, L., Dondorp, W., de Wert, G.M.W.R., ESHG Public and Professional Policy Committee, 2013. Whole-genome sequencing in health care: recommendations of the European Society of Human Genetics. Eur. J. Hum. Genet. 21, 580–584. https://doi.org/10.1038/ejhg.2013.46.
Wetterstrand, K.A., 2018. DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP) [WWW Document]. URL (accessed 9.4.18). https://www.genome.gov/27541954/dna-sequencing-costs-data/.
Willems, T., Zielinski, D., Yuan, J., Gordon, A., Gymrek, M., Erlich, Y., 2017. Genome-wide profiling of heritable and de novo STR variations. Nat. Methods 14, 590–592. https://doi.org/10.1038/nmeth.4267.