European Review for Medical and Pharmacological Sciences 2019; 23: 6753-6765

PipeMAGI: an integrated and validated workflow for analysis of NGS data for clinical diagnostics

G. MARCEDDU[1], T. DALLAVILLA[2], G. GUERRI[2], E. MANARA[1], P. CHIURAZZI[3,4], M. BERTELLI[1]

[1]MAGI EUREGIO, Bolzano, Italy
[2]MAGI'S LAB, Rovereto (TN), Bolzano, Italy.
[3]Istituto di Medicina Genomica, Università Cattolica del Sacro Cuore, Rome, Italy
[4]UOC Genetica Medica, Fondazione Policlinico Universitario "A. Gemelli" IRCCS, Rome, Italy

Abstract. – OBJECTIVE: We describe how to set up a custom workflow for the analysis of next generation sequencing (NGS) data suitable for the diagnosis of genetic disorders and that meets the strictest standards of quality and accuracy. Our method goes from DNA extraction to data analysis with a computational in-house pipeline. The system was extensively validated using three publicly available Coriell samples, estimating accuracy, sensitivity and specificity. Multiple runs were also made to assess repeatability and reproducibility.

MATERIALS AND METHODS: Three different Coriell samples were analyzed in a single run to perform coverage, sensitivity, specificity, accuracy, reproducibility and repeatability analysis. The three samples were analyzed with a custom-made oligonucleotide probe library using Nextera Rapid Capture enrichment technique and subsequently quantified using the Qubit method. Sample quality was verified using a 4200 TapeStation and sequenced on a MiSeq personal sequencer. Analysis of NGS data was then performed with a custom pipeline.

RESULTS: The workflow enabled an accurate and precise analysis of NGS data that meets all the requirements of quality and accuracy required by international standards such as ISO15189 and the Association of Molecular Pathology.

CONCLUSIONS: The proposed analysis/validation workflow has high assay accuracy, precision and robustness and can, therefore, be used for clinical diagnostic applications.

Key Words:
NGS, Validation, Diagnostics, Genome analysis.

Introduction

In the last 10 years, next-generation sequencing (NGS) technologies have acquired an increasingly important role in disease research and diagnosis. The fast and affordable simultaneous interrogation of thousands of target regions for genetic variants is making a striking contribution to the discovery of gene-disease associations[1], increasing our understanding of the molecular foundations of diseases in many different fields of medicine[2-6]. NGS is also responsible for incredible advances in the diagnosis of genetic disorders[1]. Unlike previous diagnostic sequencing technologies, NGS can deliver a full qualitative and quantitative analysis of the DNA sequences of a sample in a single test, thus giving a better idea of the diagnosis. The ability to analyze multiple regions of the genome in a short time and its low costs and accuracy make NGS an excellent substitute for Sanger technology, reducing the time and consistently increasing the probability of a diagnosis[7].

While NGS provides better methods for the diagnosis of genetic disorders, setting up an NGS workflow for clinical diagnosis involves various challenges. First of all, an NGS workflow is a multi-step procedure from DNA extraction to the clinical report, as shown in Figure 1. DNA extraction, library preparation, sequencing and data analysis can be done with a great variety of technologies and tools[8]. Choosing the right set of tools is not always straightforward, and the analysis pipeline can vary significantly depending on the final objective of the analysis. Moreover, in its transition from a new and experimental technology to a standard procedure for diagnosis, guidelines for validating NGS pipelines had to be designed to prevent inaccurate results that could be detrimental for patient management[9]. When using NGS for diagnostic purposes, the entire workflow therefore needs to be properly validated and well documented. Analysis of NGS data for

Corresponding Author: Tiziano Dallavilla, Ph.D; e-mail: tiziano.dallavilla@assomagi.org

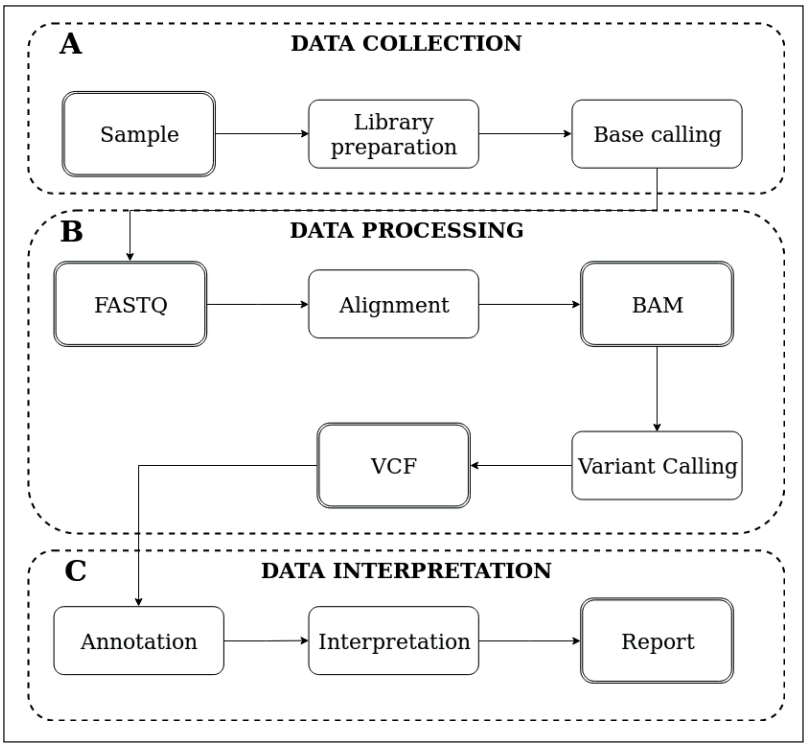

Figure 1. Typical workflow for analysis of NGS data. The procedure can be divided into three main steps: ***A,*** data collection in which the sample is prepared and sequenced with the chosen technology; ***B,*** data processing in which the raw output of the sequencer is used to determine variants in sequenced samples; ***C,*** data interpretation associating clinical significance with the variants called.

clinical diagnosis requires higher performance, quality standards, reliability, reproducibility and output robustness with respect to NGS for research. Being able to generate a diagnostic report is strictly linked to the possibility of providing an NGS workflow that meets the highest quality standards. When validating an NGS workflow it is essential to estimate many quality parameters, like sensitivity, specificity and accuracy, along with extensive coverage analysis. Regarding the analytical part of the workflow, different checkpoints need to be set up to ensure that the analytical part of the framework gives the expected results.

Here we describe a reliable, accurate and fast NGS analytical and computational approach for diagnosis of rare genetic diseases. It complies with the strictest quality standards. We demonstrate our NGS workflow by analyzing multiple reference Coriell samples with a large custom panel used for the detection of causative germline genetic variants. The samples are analyzed with a panel for the diagnosis of eye disorders, but the pipeline can be used for the analysis of a multitude of panels related to different rare genetic disorders such as cardiovascular, infertility and lymphatic conditions. We also illustrate the steps designed to validate the entire process of NGS analysis, from DNA extraction to variant annotation. The results show that our method is accurate, reliable and designed to provide safe patient care according to the recommendations and standards of the Association of Molecular Pathology[9]; it also meets the latest international quality standards for diagnostics, like ISO15189.

Materials and Methods

Custom Panel Design

For panel design and library preparation we used a method already described in our previous paper[10]. A custom-made oligonucleotide probe library was designed to capture all coding exons and flanking exon/intron boundaries (±20 bp) of 290 genes and some intronic positions (hg19 chr12:88494955-88494965 *CEP290*; chr4:15989855-15989865 *PROM1*; chr1:216247471-216247481 *USH2A*; chr1:216064535-216064545 *USH2A*; chr1:216039716-216039726 *USH2A*; chr1:215967778-215967788 *USH2A*) known from the literature or databases [Human Gene Mutation Database (HGMD Professional)[11-13], Online Mendelian Inheritance in Man (OMIM)[14], Orphanet NCBI GeneReviews, NCBI PubMed and specific databases] to be associated with a large group of eye diseases. The DNA probe set complementary to the target regions (GRCh38/hg38)

was designed using a specific online tool provided by Illumina DesignStudio (http://designstudio.illumina.com/Home/SelectAssay/) with dense probe spacing and at least two probes per target. To improve coverage of low-performance target regions, the design was optimized with the support of design experts (Illumina Concierge). The first design produced a total of 8688 probes. The final design optimized by the Illumina Concierge service generated a total of 11065 capture probes over 3921 targets and 861029 bp of cumulative target design size.

Library Preparation, Targeted Capture and Sequencing

As described in our previous paper[10], in-solution target enrichment was performed according to the manufacturer's protocol "Nextera Rapid Capture Enrichment Guide, September 2014 (Illumina Inc., San Diego, CA, USA)", except for the quantity of Tagment DNA Enzyme (5 µl instead of 15 µl specified in the protocol). 5 ng of genomic DNA was simultaneously fragmented and tagged by Nextera transposon-based shearing technology. Limited cycle PCR was carried out to incorporate specific index adaptors in each sample library. 500 nanograms of each indexed DNA library was combined with a 12-plex library pool and then hybridized with target-specific biotinylated probes and subsequently captured using streptavidin magnetic beads. A second round of hybridization, capture, PCR amplification and PCR clean-up was performed. The final enriched pooled libraries, with sizes mainly between 500 and 600 bp, were quantified using the Qubit method (Invitrogen, Carlsbad, CA, USA) and sample quality was verified using a 4200 TapeStation (Agilent Technologies, Palo Alto, CA, USA). Each pool (12-plex library) was sequenced on a MiSeq personal sequencer (Illumina Inc, San Diego, CA, USA) according to the manufacturer's instructions (150 bp paired-end (PE) reads sequencing, kit MiSeq V3).

Pipeline Environment

Analysis of NGS data was performed with a custom pipeline (PipeMagi) that runs in Ubuntu 16.04. The pipeline is written in Python 2.7[15,16] and requires several modules. The main packages are: 1) Numpy[17]; 2) Pandas[18]; 3) Matplotlib[19]; 4) Mygene[20,21]; 5) Biopython[16]. PipeMagi also requires the following software: 1) BWA[22,23]; 2) Fastx-toolkit[24]; 3) FastQC[25]; 4) Samtools[26]; 5) GATK[27,28]; 6) Variant Effect Predictor (VEP)[29] and several datasets such as the latest assembling GrCh38 for alignment; APPRIS[30] for detection of the principal transcripts and HGMD professional 2014 to match for deleterious variants[13].

Validation Algorithm

The workflow validation process was completely automatized. For this purpose, we developed an algorithm, written in Python 3.6 and running in IPython[31] with Jupyter Notebook[32], that use the following packages. Numerical computations are performed with Numpy[17] and Pandas[18] and the statistic module. Graphs are generated by Matplotlib[19]. MD5 hashes for file identification are generated with the hashlib module. The final reports are generated with openpyxl and pixiedust modules. All numeric data is stored in binary files with the help of pickle library.

Data Processing

Our pipeline accepts raw-read data in fastq format, generated by the Ilumina MiSeq reporter software (version 2.5), as input. The process of data analysis can be divided in two main parts.

In the first part we transform raw reads from the sequencing platform into data that can be used for variant calling. A scheme illustrating all the steps of the process is shown in Figure 2.

First the raw reads undergo a series of quality controls to check if the overall process of base calling went well. For this purpose, we use a common tool named FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/). This program creates a report that can be inspected to ensure that base calling went as expected and that there are no issues with sequencer output, such as quality drops or bias in the GC content length distribution and many other parameters. Next, we discard all reads that did not reach a certain quality score. Reads that do not reach a minimum Phred score of 20 on at least 97% of reads length are filtered using the FASTX toolkit[24]. Quality filtering of data reduces the number of error-prone reads, improving alignment results, accuracy of variant calling and throughput. This is essential to reduce alignment artifacts and incorrect data coverage when using enrichment systems like Nextera that do not provide stringent design capture analysis of low complexity and repeated regions. In fact, in these regions we can have up to 100% of reads with quality below the threshold[33]. Reads filtering is followed by trimming of Illumina adapters. Since these adapters are synthetic, they do not occur in the human genome,

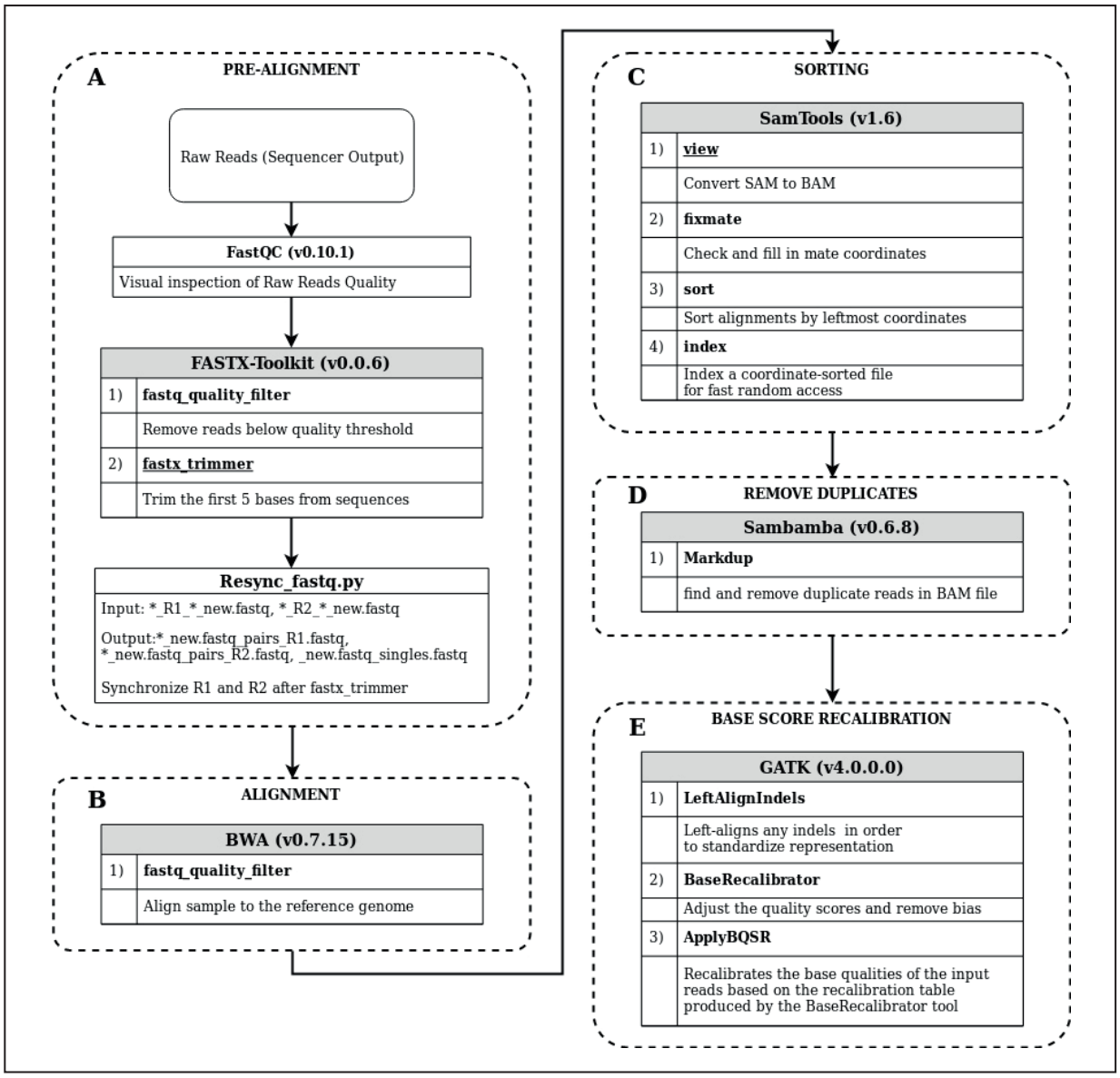

Figure 2. Scheme of the first part of the NGS data processing pipeline. ***A,*** In the pre-alignment phase we perform quality control and prepare data for alignment by trimming adapters and performing read re-synchronization. ***B,*** Reads are aligned with the genome by BWA. ***C,*** Files are sorted to improve performance. ***D,*** Duplications are removed to avoid bias during variant calling. ***E,*** To correct sequencer errors in the assignment of scores to bases, we perform base score recalibration.

so they need to be removed before mapping our reads. Both filtering and trimming are performed with the FASTX-toolkit (http://hannonlab.cshl.edu/fastx_toolkit/).

Since we opted for paired end sequencing, we perform a process called reads re-synchronization. It can happen that during quality score analysis, one read of a pair does not reach the threshold score, and as a consequence is discarded, disrupting the pair. Thus there may be regions of the genome covered only by single reads. This may create problems during alignment, since it is more difficult to map single-end than double-ends reads to the genome. This is especially true in structural rearrangements like gene insertions, deletions, duplications or low complexity and repetitive regions. Since correct mapping is fundamental for variant calling, to minimize the probability of misplaced reads we chose to exclude from analysis all reads that did not have both ends. To do so, read re-synchronization checks that every read is present in two copies, forward and reverse. Before alignment

we also checked that all read-pairing information was still intact after mapping. This was done with the samtools function "fixmate".

We then proceed with the alignment phase where we map our reads on the human genome in order to reconstruct the genotype of the samples. For the alignment (build hg38) we use the Burrows-Wheeler Aligner (BWA)[22]. The SAM file with all the reads mapped onto a region of the human genome is then converted into a binary compressed format (BAM) with SamTools[26]. We then proceed to sort our data according to position on a reference sequence. The final step of the sorting phase is duplicate removal to mitigate potential bias during variant calling: this is done with the sambamba markdup function[34]. Base quality scores are corrected to mitigate errors made by the sequencer when estimating the quality score of each base call. This is achieved using GATK BaseRecalibrator and applyBQSR[27].

The second part of the workflow consists in variant annotation and interpretation. A scheme of the procedure is shown in Figure 3.

Before proceeding with variant calling, we have two intermediate steps. First we generate a BED file containing the target regions of the diagnostic suspicions for which the sample needs to be analyzed. As required by the guidelines, the BED file is composed of the whole coding region of a selected list of genes and all neighboring bases (15 for each coding region). This phase ensure that we only analyze the regions of interest, minimizing the possibility of incidental findings, a current ethical problem for which a clear solution still needs to be found[35,36]. After generation of the BED file, extensive coverage analysis is performed on the selected regions. In the diagnostic environment, it is critical to know whether a region of a gene has been sequenced with sufficient quality and depth, so that geneticists can be confident about the results of variant calling. To do this, Samtools has a convenient command that generates a text file indicating the depth measured on each base. This file is, then, processed to estimate per-base coverage.

Once the base scores are recalibrated, we perform variant calling and annotation (Figure 3). For variant calling we use two main tools. The first is Samtools mpileup[26] in association with bcftools[37]. Mpileup collects summary information in the filtered input BAM, computes the likelihood of the reference genome and stores them in a BCF file. The samtools mpileup call command is used to make the calls. The second tool for variant calling is GATK HaplotypeCaller. The two VCF files produced are united in a single file. To determine the effect of each variant we annotate it with a tool called variant effect predictor (VEP)[29]. For each variant we annotate the gene and transcript affected, along with their location. We also retrieve the predicted protein sequence, SIFT and POLYPHEN scores, Minor allele frequency (MAF), and known variants. The APPRIS[30] database is used to filter annotations on each variant. Based on RefSeq 107, we use APPRIS to keep the annotation that is only associated with the principal transcript, excluding all the other annotations (i.e. annotations on model organisms).

Sanger Validation and Sequencing of Poorly Covered Target Regions

Each predicted pathogenic variant is confirmed by conventional Sanger sequencing using genomic DNA from different aliquots of the sample. Target regions with coverage less than 10 reads were additionally analyzed by Sanger sequencing according to the manufacture's protocols (CEQ8800 Sequencer, Beckman Coulter)[10].

Validation Process

Coverage analysis

Coverage analysis provides information on the sequencing performance of each gene in the panel, highlighting regions of the panel that perform poorly. Identification of these regions is fundamental in diagnostics, since no variant analysis can be performed on low coverage regions. Identifying holes is also useful since the panel can be re-designed during a revision phase to improve performance. Coverage analysis starts by analyzing the overall coverage of each sample, estimating the following parameters:

- mean coverage
- standard deviation
- median
- 5th percentile
- 95th percentile

Since under optimal condition coverage will have a Poisson-like distribution, a comparison of the mean and median values is a quick way to check whether coverage is correctly distributed. Very different values indicate a skewed distribution, and are therefore a symptom of problems during the run. It is also important to ensure that there is low variability between samples of the same run, and those of the second run. Too much variability may indicate a problem during library preparation or base calling.

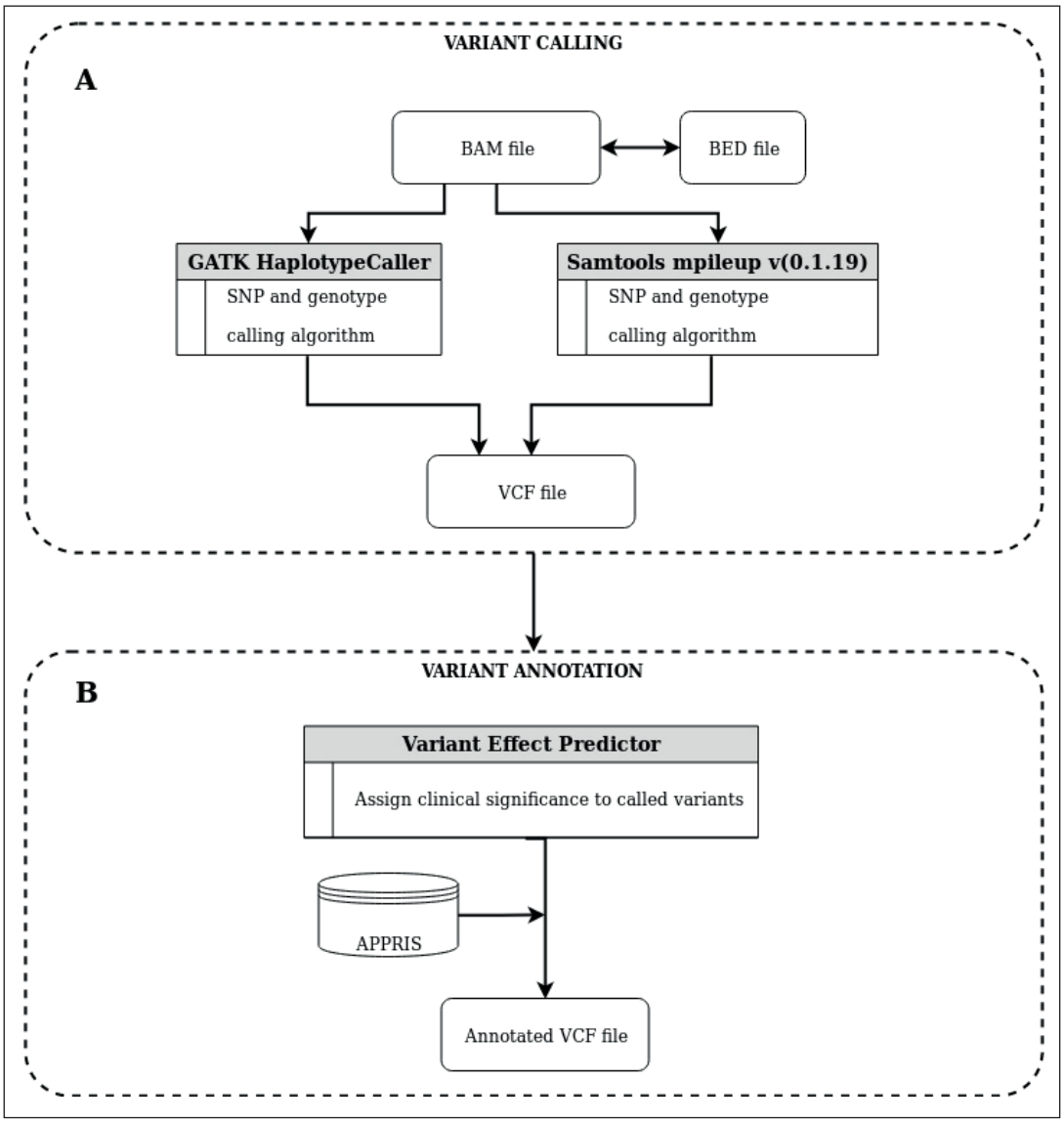

Figure 3. Scheme of the second part of the NGS data processing pipeline. ***A,*** First we perform variant calling using Samtools mpileup call and GATK UnifiedGenotyper. The calling is only focused on target regions and not on the entire genome. To filter called variants, we design a BED file containing the coordinates of target regions. Prior to variant calling the BAM file is filtered with the BED file so as only to report findings in regions of interest. ***B,*** The called variants are then annotated by VEP. Since a variant can overlap multiple transcripts, it may end up with multiple annotations. This is why we use the APPRIS database to determine the principal transcript and we filter all annotations that do not belong to the principal transcript.

Assay accuracy

Three different Coriell samples were analyzed in a single run to perform coverage, sensitivity, specificity and accuracy analysis. Sample NA20509 was loaded twice during the first run to assess reproducibility, while the same sample was analyzed on a separate run to address repeatability. To perform validation, we wrote a custom algorithm that requires three files as input for each sample to analyze:

- the BED file of the panel
- the depth file generated with the Samtools function "depth"
- the VCF file produced by the pipeline.

The output of the analysis consists of two reports *(supplementary material S1 and S2):* a validation report, which is a summary of all the validation parameters, and a technical report containing detailed information about every step of the validation, which can also be used for troubleshooting.

Panel design validation

During validation we perform a quality check on the panel design to ensure that what we designed matches what we are really sequencing. In this section we check:

- cumulative target length
- number of targeted regions
- numbers of probes
- number of genes.

All these statistics need to match what was defined during the panel design phase.

Sensitivity, specificity, accuracy

To validate the workflow, we calculated the following parameters:

1. Sensitivity: a measure of the capacity to correctly identify variants that really exist in the samples. In our analysis sensitivity is calculated as:
   True Positive (TP) / TP+False Negative
2. Specificity: quantification of the capacity to avoid calling variants that are not really present in the samples. Specificity is measured as:
   True Negative (TN) / TN+False Positive (FP)
3. Accuracy: measure of how well the pipeline correctly identifies or excludes a variant. It is calculated as:
   TP+TN / TP+TN+FP+FN
4. Reproducibility: measure of the capacity of the workflow to reproduce the same result on a sample in the same run.
5. Repeatability: measure of the capacity of the workflow to reproduce the same result on a sample in successive runs.

Results

Validation was needed to ensure that the performance of the NGS approach was in line with the standard required for clinical diagnosis. The following tests confirmed that the pipeline was robust and minimized the probability of errors during analysis.

Assay Robustness

During the validation process, we defined the acceptability thresholds of the NGS analytical performance parameters. Limitations on critical parameters ensured assay success and the desired level of precision (Internal Quality Control – Q/C).

1. Genomic DNA sample quality threshold: Total genomic DNA (gDNA) of the samples analyzed for genetic testing was extracted using Magpurix Extraction Kit (Resnova) by ZINEXTS MagPurix 12 System. DNA extraction from peripheral blood leucocytes and tissue samples was performed according to the manufacturer's recommendations. MagPurix Forensic DNA Extraction Kit was used for saliva samples according to the manufacturer's instructions with only a few exceptions in the sample preparation step to adapt the protocol to the sample collection method used (Isohelix Genefix Saliva DNA Collection kit – GFX-01). The collectors are designed to collect 2 ml saliva in 2 ml lysis buffer pre-filled in 10 ml collection tubes. The optimized protocol included the following changes: 3:1 saliva and BL2 Buffer volume ratio instead of the 1:4 ratio specified in the manufacturer's instructions; incubation time at 56°C in the lysis step, modified to 45 min instead of 15 min. Note that 300 µl of sample solution collected (saliva + lysis buffer) was mixed with 200 µl Buffer BL2 and 20 µl proteinase K and incubated at 56°C for 45 min.
2. gDNA input quality threshold
   - gDNA ≥50 ng
   - 260/280 absorbance ratio of 1.7-2.0
   - Only for DNA from saliva samples (Isohelix Genefix Collection): 260/230 absorbance ratio ≥1. DNA samples that did not meet these criteria showed poor clustering.
   - DNA integrity. Degraded DNA samples resulted in library preparation failure (they generated inserts of a shorter length) and poor clustering.
3. Enriched library quality threshold: Library DNA fragments in the size range ~200 bp to ~1 kbp. Average fragment size not less than 300 bp. Incorrect DNA fragmentation resulted in suboptimal data.
4. Post-run/pre-analysis read quality threshold: more than 80% of bases higher than Q30 at 2x 150 bp
5. Post-analysis targeted region coverage threshold: Optimal coverage of target regions: at least 98% of bases with at least 10x coverage, while 95% of bases with at least 10x coverage is considered the minimum acceptable threshold of the subpanel of genes associated with the sub-phenotype, except for problematic regions (e.g. GC-rich or repetitive regions).

Sample, algorithm and diagnostic suspicion identification

While this is not part of the validation procedure itself, one of the points that we address when doing validation is that the analysis needs to be repeatable at any time. In the report it is therefore essential to keep track of the location of all the files used for validation, which are normally stored on a proprietary server. Unfortunately, knowing the location of the files does not ensure that we are always able to find and identify them unequivocally. Files may change location, migrating during systems updates or simply by accident. We therefore implemented a second step of identification to recognize the files used in a particular validation run. This step consists in calculating the MD5 hash for each file used (Figure 4).

Coverage analysis

In this study, we performed NGS analysis of 290 panel genes encompassing multiple eye disorders, including diseases affecting the retina, cornea and macula, as well as IOP and myopia (see **Supplementary material S1**). Ten gigabytes

Samples information

MAGI APP ID	INTERNAL ACCEPTANCE ID	CATOLOG ID DNA CORIELL	Fastq link ; Files results link (VCF; BED; annotation file); MAGI APP results link	Run Data; Pipeline Analysis Data
ONA20509.2018	RE047	NA20509	\\Storage\ngs\RAW\BOLZANO\01_Oct_2018_OCULARE ; \ \Storage\ngs\result\BOLZANO\01_Oct_2018_OCULARE; http://192.168.1.220/ngs/ngsresult/ONA20509.2018	28/09/2018; 01/10/2018

Description: INTERNATIONAL HAPMAP PROJECT - TOSCANI IN ITALIA (TUSCANS IN ITALY) INTERNATIONAL HAPMAP PROJECT - PANEL OF 90 TOSCANI IN ITALIA (TUSCANS IN ITALY) 1000 GENOMES PROJECT - PANEL OF 114 TOSCANI IN ITALIA Gender: Male https://catalog.coriell.org/0/Sections/Search/Sample_Detail.aspx?Ref=GM20509&PgId=166

Samples coverage files:

Sample n.	Sample ID	Coverage file name	MD5 coverage file
1	ONA20763.2018	ONA20763.2018_all	e73b5dc073e50ed9c96b167673651c8c
2	ONA20828.2018	ONA20828.2018_all	29fd7b3be3587bc971fbcb7752ae81c8
3	ONA20509.2018	ONA20509.2018_all	8f8e2dca8dbc9da80c8e6f19deaf4066
4	bis2ONA20509.2018	bis2ONA20509.2018_all	d8889a5e0a9ccc80e8f15cec1645e6ee
5	bisONA20509.2018	bisONA20509.2018_all	44dd5d3c3b6ea845090c21ace449d09d

Samples vcf files:

Sample n.	Sample ID	VCF file name	MD5 VCF file
1	ONA20509.2018	ONA20509.2018_pheno_predict.csv	274599ba3edfb71110c51b12642ba6a5
2	ONA20828.2018	ONA20828.2018_pheno_predict.csv	6b61a8a2787812fcd24f8473f887e86e
3	ONA20763.2018	ONA20763.2018_pheno_predict.csv	e03ced862c441d54cffdb4e59feffa2f
4	bis2ONA20509.2018	bis2ONA20509.2018_pheno_predict.csv	d2889ce4fbcb8a2760023615ff22b87e
5	bisONA20509.2018	bisONA20509.2018_pheno_predict.csv	5c0b9d33eb15cbc1788b7fac13f71582

Reference vcf files:

Reference n.	Reference ID	VCF file name	MD5 VCF file
1	refVCF	refVCF_coriell_ocular.csv	9c9df97c2082590b72cb69e8128e3fba

Figure 4. Identification of files needed for validation. We report the location of files used in validation on our internal server along with the MD5 hash. In this way, it is always possible to identify the files used in the validation, in case we need to repeat it.

Table I. Coverage analysis results of each sample.

	Average depth				
Sample	**Mean depth**	**Standard deviation**	**5th percentile**	**Median**	**95th percentile**
NA20509	312.9x	154.9	74	303	587
NA20763	267.3x	131.7	60	262	495
NA20828	293.2x	144.3	67	286	546
Run	291.1±18.7	143.6±11.6	67.0±7.0	283.7±20.6	542.7±46.1

(±0.6) of sequenced bases were generated per run, producing 2.5 (±0.2) million mappable reads per sample. The two runs were of high quality with a mean of 91.8% (±1.8) of sequenced bases and a Phred Q score ≥ 30. The mean coverage of targeted bases for the five reference samples (including the two replicates) was 294.9× (±19.1) per sample, while 98.8% (±0.04), 98.3% (±0.1) and 97% (±0.2) of all bases were covered at least 10×, 20× and 40×, respectively (see **Supplementary material S1**). Our results (Tables I and II) show that the mean and median for each sample were indeed very close, with the biggest variation being NA20509, where mean and median differed by 3.16%. This is indicative of a Poisson-like distribution of all samples. The mean coverage gives us an overview of how well the targets in the panel were covered. The results with the samples indicated that we were well above the threshold required to perform variant calling. Moreover, the variability between samples was much reduced and sequencing data quality was high, indicating a robust analysis workflow.

Next, we analyzed the percentage of samples covered above a certain threshold (Table II). In our case we checked the percentage of the panel covered at 10x, 20x, 40x, 100x, considering 10x the minimum coverage necessary for variant calling. The sub-panel needs to reach 10x coverage on at least 95% of the targets in order to proceed with the analysis, otherwise we repeat sample analysis. The results show that we were well above the threshold, since more than 98% of the targets were above the threshold of 20x in all samples. We also did analysis of coverage for each diagnostic suspicion (DS) (**Supplementary material S1**) to ensure that they were all above the coverage threshold of 10x for at least 95% of DSs, since a clinical report for a particular DS is only reliable if that DS reaches the minimum coverage on most of the genes, irrespective of the performance of the whole panel. Our results indicate very good results for almost all DSs: 91% were above the threshold of 10x. The target regions of the DS that did not reach the fixed thresholds are reported in **Supplementary material S1** and **Supplementary material S2** and used during panel revision to improve performance on poorly performing targets. Since coverage analysis is done not only during validation but on every sample analyzed, whenever the coverage of a region falls below the threshold, if the region contains variants that are significant for the DS for which the sample is analyzed, the clinician can have those regions sequenced by Sanger technology. Coverage analysis is concluded by two graphs that summarize the results (**Supplementary material S1**). One histogram shows the mean coverage per gene. The other shows the percentage of each gene covered at 10× and 40×. These graphs make it easy to identify genes that perform poorly during sequencing. These genes become the focus of panel revision. If we have genes that perform poorly, we need to check whether performance is poor in all samples. If so, it indicates a problem in panel design. If the genes that do not perform well differ between samples, it may be symptomatic of a problem during the data collection phase.

Table II. Percentage of panel covered at different thresholds.

Sample	**Mean depth**	**Target covered at 10x**	**Target covered at 20x**	**Target covered at 40x**	**Target covered at 100x**
NA20509	312.9x	98.9%	98.4%	97.3%	92.6%
NA20763	267.3x	98.8%	98.2%	96.8%	89.8%
NA20828	293.2x	98.8%	98.3%	97.2%	91.5%
Run	291.1±18.7	98.9%±0.1	98.3%±0.1	97.1%±0.3	91.3%±1.4

Assay Accuracy

All our workflow was developed with the aim of achieving elevated specificity, sensitivity and accuracy. Variant calling accuracy was assessed by in-house pipeline using reference samples NA20509, NA20763 and NA20828. The reference sample data was downloaded from the 1000 genomes phase III project (http://www.internationalgenome.org/category/phase-3/). The datasets were filtered for SNVs and small indel variants in regions of interest. Since the genomes in phase III are in Hg19 version, and we are working with Hg38, prior to comparison between the reference genotype and what was generated by the pipeline, we had to do a liftover to convert the Hg19 reference genotype to Hg38 in order to work with the same assembly. The result of the liftover was hand curated since the passage from one version to another introduces different problems that need to be addressed, for example many variants are changed from REF to ALT between the two versions of the genotype. Once we obtained the reference genotypes they were used to make a comparison with our results from the pipeline, so that for each sample we were able to estimate the parameters needed for calculation of sensitivity, specificity and accuracy. For each sample we calculated (Table III):

- true positives: variants in the reference sample that were correctly called by the pipeline.
- true negatives: variants not detected by the pipeline and not expected in the reference sample.
- false positives: variants detected by the pipeline that were not present in the reference sample.
- false negatives: variants not called by the pipeline but present in the reference sample.

Variants within the 800 kbp target region were compared between our in-house pipeline and the reference variant list of SNVs and indels. Discrepancy between reference data sets was resolved by further examining the quality score of our pipeline data, variant context, and BAM file alignment.

Comparison of the reference data set with the in-house pipeline data showed on average 39 (±5.50) putative false positives in the pipeline data set. Of these, six (±3.3) indel samtools calls were artifacts, showing misalignment in the .bam file, and 9.7 (±2.9) were variants in low complexity regions, such as homopolymer and repeat regions, indicating that the variants were unlikely to be true positives (see "FP" variants in Supplementary material S2, sheet FP).

Some variants in the public datasets of the three reference samples were putative false negatives. Of these, SNVs not correctly called showed misalignment calls for reads falling in indel variant positions and indel calls in low complexity regions, indicated as true negatives. Intronic indel variants were correctly called by the pipeline but not selected by target filtering (indel first position off-target), resulting as false negatives (see "FN" variants in Supplementary material S2 sheet FN). In summary, more than 799,204 (±18) positions in the targeted regions were correctly called as true negatives and 533 (±14) variants were correctly called as true positives, resulting in 99% average analytical sensitivity and specificity for SNV and indel detection. The detailed results are reported in Supplementary material S1 and Table IV.

Assay Precision

Precision refers to the reproducibility or "robustness" of the assay, meaning the ability to obtain the same results from the same sample when the assay is performed repeatedly. For reproducibility, both intra-run and inter-run reproducibility should be assessed. To evaluate intra-run precision (repeatability), two parallel libraries were prepared from reference sample NA20509, each with a unique index. An equimolar amount of each library was pooled and sequenced on the same Miseq flow cell. To evaluate inter-run precision (reproducibility), NA20509 DNA was captured and sequenced in another independent run. Variants called for each sam-

Table III. Summary of variant calling for each sample. This table reports the number of false positives (FP), true positives (TP), false negatives (FN) and true negatives (TN) estimated in each sample.

NA20509		NA20763		NA20828	
FN	***TN***	***FN***	***TN***	***FN***	***TN***
4	799222	2	799203	2	799186
FP	***TP***	***FP***	***TP***	***FP***	***TP***
10	543	14	559	22	567

Table IV. Summary of specificity, sensitivity and accuracy scores with mean and 95% confidence interval (CI).

Sample	Specificity	Sensitivity	Accuracy
NA20509	0.99	0.99	0.99
NA20763	0.99	0.99	0.99
NA20828	0.99	0.99	0.99
Mean and CI	0.99 (0.99-1.00)	0.98 (0.98-0.99)	0.99 (0.99-1.00)

ple/run were compared among intra-run library samples (ONA20509, bisONA20509) to assess intra-run repeatability, and between the inter-run library sample (ONA20509) and the intra-run sample (bis2ONA20509) to assess inter-run reproducibility. Reproducibility was calculated by dividing the number of discordant calls by total variants in the reference sample; the results are shown in Table V. Regarding reproducibility (Table V and Supplementary material S1) (sample bisONA20509.2018), we observed close agreement between the two samples, with an identical percentage of target covered above 10x in the two samples. The target coverage was in close agreement between NA20509 sample replicas (see Table V and Supplementary material S1), confirming high robustness of the method. For variant calling, we observed 99.28% identity between intra-run replicas (repeatability) and 99.28% identity between inter-run replicas (reproducibility), excluding indel misalignments.

Conclusions

The validation method presented in this paper was applied to a large panel of ~801 kb, and on all our panels used for diagnostic purposes. The NGS workflow was tested using three control samples from the Coriell Institute, in order to measure with high fidelity the detection rate performance of our method and to collect information useful for improving its analytical performance. The results showed a high mean coverage (291.1±18.7), and most of the target (98.8%±0.06) was covered by at least 10 independent reads, our estimated minimum threshold for variant calling in germline samples. The quality of our workflow was also highlighted by the fact that we exceeded the chosen standard of coverage: in fact 98.3%±0.1 and 97%±0.26 of the bases were covered at least 20x and 40x, and we observed very close results between samples, indicating good coverage uniformity and high robustness of the method.

While an optimal coverage ensures a more reliable variant calling phase, it is known that low coverage increases the risk of false negatives, and can lead to assignment of wrong allelic states (zygosity). It also lowers the probability of effectively filtering sequencing artifacts, increasing the number of false positives. The workflow was set up accurately in order to minimize false positives/negatives. In particular, we tried to minimize the number of reads misaligned to homologous regions by adopting strategies like paired end sequencing, long reads and local realignment strategies after global alignment. Concerning the removal of artifacts, we underline the importance in the diagnostic context of applying quality filtering of data, so that variant calling is only performed on high quality reads, thus improving the alignment, accuracy and data coverage results. In our case, application of an appropriate filter increased the specificity by about 2% without affecting assay sensitivity (data not shown).

We confirmed the effectiveness of our strategy by elevated assay accuracy, precision and robustness and good target coverage and uniformity. Since the analytical parameters reported reflect the overall quality of the entire method, we conclude that the strategies used to develop our NGS targeted panel produced an in-house pipeline with very high quality standards and reproducibility. The pipeline can therefore be used confidently for clinical diagnostic applications.

Table V. Summary of reproducibility (rep, sample bisONA20509) and repeatability (repe, sample bis2ONA20509) analysis.

	Variants in reference sample	Variants in rep/repe samples	Variants unique to rep/repe samples	Variants unique to reference sample	% overlap in variant calling
Repeatability	553	553	2	2	99.28
Reproducibility	553	547	1	3	99.28

Supplementary material

Supplementary file S1. The validation report generated by the algorithm which gives an overview of the main parameters estimated during validation.
Supplementary file S2. Technical report used for trouble-shooting during validation.

Data availability

The data used to obtain the findings of this study can be found in the article and the supplementary information files.

Conflict of Interests

The authors declare that they have no conflict of interest regarding publication of this paper.

References

1) Su Z, Ning B, Fang H, Hong H, Perkins R, Tong W, Shi L. Next-generation sequencing and its applications in molecular diagnostics. Expert Rev Mol Diagn 2011; 11: 333-343.
2) Sommariva E, Pappone C, Martinelli Boneschi F, Di Resta C, Rosaria Carbone M, Salvi E, Vergara P, Sala S, Cusi D, Ferrari M, Benedetti S. Genetics can contribute to the prognosis of Brugada syndrome: a pilot model for risk stratification. Eur J Hum Genet 2013; 21: 911-917.
3) Daoud H, Luco SM, Li R, Bareke E, Beaulieu C, Jarinova O, Carson N, Nikkel SM, Graham GE, Richer J, Armour C, Bulman DE, Chakraborty P, Geraghty M, Lines MA, Lacaze-Masmonteil T, Majewski J, Boycott KM, Dyment DA. Next-generation sequencing for diagnosis of rare diseases in the neonatal intensive care unit. CMAJ 2016; 188: 254-260.
4) Rehm HL. Disease-targeted sequencing: a cornerstone in the clinic. Nat Rev Genet 2013; 14: 295-300.
5) LePichon JB, Saunders CJ, Soden SE. The future of next-generation sequencing in neurology. JAMA Neurol 2015; 72: 971-972.
6) Peters DG, Yatsenko SA, Surti U, Rajkovic A. Recent advances of genomic testing in perinatal medicine. Semin Perinatol 2015; 39: 44-54.
7) Schrijver I, Aziz N, Farkas D, Furtado M, Ferreira-Gonzalez A, Greiner TC, Grody WW, Hambuch T, Kalman L, Kant JA, Klein RD, Leonard D, Lubin I, Mao Rg, Nagan N, Pratt V, Sobel M, Voelkerding KV, Gibson JS. Opportunities and challenges associated with clinical diagnostic genome sequencing: a report of the association for molecular pathology. J Mol Diagn 2012; 14: 525-540.
8) Levy SE, Myers RM. Advancements in next-generation sequencing. Annu Rev Genomics Hum Genet 2016; 17: 95-115.
9) Roy S, Coldren C, Karunamurthy A, Kip NS, Klee EW, Lincoln SE, Leon A, Pullambhatla M, Temple-Smolkin RL, Voelkerding KV, Wang C, Carter AB. 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 2018; 20: 4-27.
10) Mattassi R, Manara E, Colombo PG, Manara S, Porcella A, Bruno G, Bruson A, Bertelli M. Variant discovery in patients with Mendelian vascular anomalies by next-generation sequencing and their use in patient clinical management. J Vasc Surg 2018; 67: 922-932.
11) Stenson PD, Mort M, Ball EV, Shaw K, Phillips A, Cooper DN. The human gene mutation database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. Hum Genet 2014; 133: 1-9.
12) Stenson PD, Mort M, Ball EV, Evans K, Hayden M, Heywood S, Hussain M, Phillips AD, Cooper DN. The human gene mutation database: towards a comprehensive repository of inherited mutation data for medical research, genetic diagnosis and next-generation sequencing studies. Hum Genet 2017; 136: 665-677.
13) Stenson PD, Ball EV, Mort M, Phillips AD, Shaw K, Cooper DN. The Human Gene Mutation Database (HGMD) and its exploitation in the fields of personalized genomics and molecular evolution. Curr Protoc Bioinformatics 2012; 1. doi: 10.1002/0471250953.bi0113s39.
14) Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Res 2002; 30: 52-55.
15) van Rossum G. Python tutorial. Technical Report CS-R9526, Centrum voor Wiskundeen Informatica (CWI), Amsterdam, 1995. Available online at https://ir.cwi.nl/pub/5007/05007D.pdf
16) Cock PJ, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, Friedberg I, Hamelryck T, Kauff F, Wilczynski B, de Hoon MJ. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics 2009; 25: 1422-1423.
17) Travis E. Oliphant. Guide to NumPy. CreateSpace Independent Publishing Platform, USA, 2nd edition, 2015. ISBN: 151730007X.
18) McKinney W. Data structures for statistical computing in Python. Proc of the 9th Python in Science Conf 2010; 51-56. Available online at: https://conference.scipy.org/proceedings/scipy2010/pdfs/mckinney.pdf.
19) Hunter JD. Matplotlib: a 2d graphics environment. Comput Sci Eng 2007; 9: 90-95. doi:10.1109/MCSE.2007.55
20) Xin J, Mark A, Afrasiabi C, Tsueng G, Juchler M, Gopal N, Stupp GS, Putman TE, Ainscough BJ, Griffith OL, Torkamani A, Whetzel PL, Mungall CJ, Mooney SD, Su AI, Wu C. High-performance web services for querying gene and variant annotation. Genome Biol 2016; 17: 91.
21) Wu CL, MacLeod I, Su AI. BioGPS and MyGene.info: organizing online, gene-centric information. Nucleic Acids Res 2013; 41: 561-565.

22) Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. ArXiv 2013; 00: 1-3. Available online at: https://arxiv.org/abs/1303.3997.
23) Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics 2010; 26: 589-595.
24) Gordon A, Hannon GJ. Fastx-toolkit. FASTQ/A short-reads pre-processing tools. 2010. Available online at: http://hannonlab.cshl.edu/fastx_toolkit/index.html.
25) Andrews S. FastQC a quality control tool for high throughput sequence data. 2010. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc
26) Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. The sequence alignment/map format and SAMtools. Bioinformatics 2009; 25: 2078-2079.
27) McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 2010; 20: 1297-1303.
28) Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, Del Angel G, Levy-Moonshine A, Jordan T, Shakir K, Roazen D, Thibault J, Banks E, Garimella KV, Altshuler D, Gabriel S, DePristo MA. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Curr Protoc Bioinformatics 2013; 43: 1-33.
29) McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GR, Thormann A, Flicek P, Cunningham F. The Ensembl variant effect predictor. Genome Biol 2016; 17: 122.
30) Rodriguez JM, Maietta P, Ezkurdia I, Pietrelli A, Wesselink JJ, Lopez G, Valencia A, Tress ML. APPRIS: annotation of principal and alternative splice isoforms. Nucleic Acids Res 2013; 41: 110-117.
31) Pérez F, Granger BE. IPython: a system for interactive scientific computing. Comput Sci Eng 2007; 9: 21-29.
32) Kluyver T, Ragan-Kelley B, Perez F, Granger B, Bussonnier M, Frederic J, Kelley K, Hamrick J, Grout J, Corlay S, Ivanov P, Avila D, Abdalla S, Willing C. Jupyter notebooks – a publishing format for reproducible computational workflows. Positioning and power in academic publishing: players, agents and agendas. 2016. Available online at: https://eprints.soton.ac.uk/403913/.
33) Dozmorov MG, Adrianto I, Giles CB, Glass E, Glenn SB, Montgomery C, Sivils KL, Olson LE, Iwayama T, Freeman WM, Lessard CJ, Wren JD. Detrimental effects of duplicate reads and low complexity regions on RNA- and ChIP-seq data. BMC Bioinformatics 2015; 16: 10.
34) Tarasov A, Vilella AJ, Cuppen E, Nijman IJ, Prins P. Sambamba: fast processing of NGS alignment formats. Bioinformatics 2015; 31: 2032-2034.
35) Kalia SS, Adelman K, Bale SJ, Chung WK, Eng C, Evans JP, Herman GE, Hufnagel SB, Klein TE, Korf BR, McKelvey KD, Ormond KE, Richards CS, Vlangos CN, Watson M, Martin CL, Miller DT. Recommendations for reporting of secondary findings in clinical exome and genome sequencing, 2016 update (ACMG SF v2.0): a policy statement of the American College of Medical Genetics and Genomics. Genet Med 2017; 19: 249-255.
36) Green RC, Berg JS, Grody WW, Kalia SS, Korf BR, Martin CL, McGuire AL, Nussbaum RL, O'Daniel JM, Ormond KE, Rehm HL, Watson MS, Williams MS, Biesecker LG. ACMG recommendations for reporting of incidental findings in clinical exome and genome sequencing. Genet Med 2013; 15: 565-574.
37) Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics 2011; 27: 2987-2993.