Reliable variant calling during runtime of Illumina sequencing

Tobias P. Loka[1], Simon H. Tausch[1,2], Bernhard Y. Renard[1]*

[1] Bioinformatics Division (MF 1), Department for Methods Development and Research Infrastructure
[2] Centre for Biological Threats and Special Pathogens: Highly Pathogenic Viruses (ZBS 1)
* Correspondence to Bernhard Y. Renard, Robert Koch Institute, Nordufer 20, 13353 Berlin; +49 30 18754 2561; RenardB@rki.de

Abstract

The sequential paradigm of data acquisition and analysis in next-generation sequencing leads to high turnaround times for the generation of interpretable results. We combined a novel real-time read mapping algorithm with fast variant calling to obtain reliable variant calls still during the sequencing process. Thereby, our new algorithm allows for accurate read mapping results for intermediate cycles and supports large reference genomes such as the complete human reference. This enables the combination of real-time read mapping results with complex follow-up analysis. In this study, we showed the accuracy and scalability of our approach by applying real-time read mapping and variant calling to seven publicly available human whole exome sequencing datasets. Thereby, up to 89% of all detected SNPs were already identified after 40 sequencing cycles while showing similar precision as at the end of sequencing. Final results showed similar accuracy to those of conventional *post-hoc* analysis methods. When compared to standard routines, our live variant calling approach enables considerably faster interventions in clinical applications and infectious disease outbreaks. Besides variant calling, our approach can be adapted for a plethora of other mapping-based analyses.

Introduction

Common workflows for the analysis of Illumina next-generation sequencing (NGS) data can only be applied after sequencing has finished. In time-critical applications, however, this sequential paradigm of data acquisition and analysis is one of the main bottlenecks leading

to high turnaround times. Examples for such time-critical analyses range from the differential diagnosis of genetic disorders in infants[1, 2, 3, 4], to the determination of *M. tuberculosis* drug resistances[5], and to the identification of pathogens, virulence factors, drug resistances and paths of disease transmission in infectious disease outbreaks[6, 7]. While having considerably higher turnaround times than targeted approaches such as molecular tests, NGS provides a more open view as well as more extensive and reliable results. During bioinformatics analysis of NGS data, read mapping and variant calling are crucial steps to obtain genetic information that is essential for the treatment of a patient, including strain level classification and drug resistances of a pathogen or the presence of genetic disorders that are known to be associated with specific disease characteristics.

While HiLive[8], the predecessor of our new algorithm HiLive2, delivered results at the end of sequencing, HiLive2 can produce read mapping output for arbitrary sequencing cycles while still sequencing. At the same time, the new algorithm is faster, more accurate and enables scalability to large reference genomes such as the complete human reference. The recently published software LiveKraken[9] already gives k-mer based taxonomic classification results for arbitrary sequencing cycles. However, while LiveKraken provides valuable information about the microbial composition of a sample, the results do not allow for complex reference-based follow-up analyses such as variant calling or the analysis of drug resistances. Alternative approaches to obtain read mapping results for Illumina data while still sequencing, such as rapid pulsed whole genome sequencing[4], lack sufficient scalability for high amounts of data and large reference genomes and are therefore only suitable for special use cases. At the same time, the incremental approach of HiLive2 provides higher flexibility in the choice of output cycles which can even be modified during the runtime of the sequencer. When compared to Oxford Nanopore sequencing Technology (ONT), which enables real-time analysis by design, Illumina sequencing provides higher scalability at lower costs and with much lower error rates. Therefore, while being a promising technology for real-time analysis in the future, sequencing technology, protocols and computational analysis

for ONT need to be further established and improved to become a viable alternative for many scenarios.

The workflow described in this study is based on real-time read mapping results with our novel algorithm HiLive2 followed by fast and accurate variant calling with xAtlas[10]. Thereby, live results can be obtained several hours before all data are written by the sequencer and provide increasing insights into the sample over sequencing time. This study describes the application of our workflow to human whole exome sequencing data, showing the scalability of our approach for high amounts of complex data and large reference genomes. However, our general approach can easily be adapted for different types of sequencing methods such as whole genome sequencing (WGS) or amplicon sequencing and a plethora of different mapping-based analysis methods.

Our new real-time read mapping software HiLive2 is publicly available under BSD-3-clause on https://gitlab.com/rki_bioinformatics/hilive2.

Results

Implementation and experimental setup

To allow for faster NGS-based diagnosis and treatment, we developed a new real-time read mapping algorithm that generates high-quality results. We combined our new software with a fast variant caller to produce high-quality variant calls based on intermediate read mapping results, while sequencing is still running. This approach allows reliable and fast variant calling results without reducing the final sequencing coverage or quality. Therefore, we adapted our real-time read mapper HiLive[8] that gives output at the end of sequencing, using a novel algorithm based on an efficient FM-index implementation[11] for continuously analyzing sequencing results during runtime. The new version (HiLive2) achieves scalability to larger indices such as the complete human reference genome. At the same time, the algorithm comes with improved performance in terms of runtime, memory and data storage and overcomes heuristic elements that were present in previous version of HiLive. The high

scalability and accuracy of HiLive2 enable the combination of real-time read mapping results with complex follow-up analyses that have not been possible with the previous version. To demonstrate the power of such analyses, we performed variant calling on seven whole exome sequencing data sets of the human individual NA12878 from the CEPH Utah Reference Collection (cf. Table 1) using the real-time read mapping results of HiLive2 as input data.

Table 1 Summary of data sets evaluated in this study. Information about sequencing platform, exome capture and coverage were adopted from Hwang et al. (2015)[12].

Accession No.	Platform	Exome capture	Exome coverage	Reads[1]	Read length
SRR098401	HiSeq2000	SureSelect v2	116.84x	114M	2 x 76bp
SRR292250	HiSeq2000	SeqCap EZ v2	116.06x	85M	2 x 50bp
SRR515199	HiSeq2000	SureSelect v4	298.45x	167M	2 x 100bp
SRR1611178	HiSeq2000	SeqCap EZ v3	79.93x	45M	2 x 100bp
SRR1611179	HiSeq2000	SeqCap EZ v3	79.84x	45M	2 x 100bp
SRR1611183	HiSeq2500	SeqCap EZ v3	129.94x	74M	2 x 100bp
SRR1611184	HiSeq2500	SeqCap EZ v3	111.90x	64M	2 x 100bp

[1] M:= millions

For variant calling, we used the fast variant caller xAtlas which shows comparable accuracy to established methods at much lower runtime[10]. We compared our results to read mapping with Bowtie 2[13] and variant calling with either xAtlas or GATK HaplotypeCaller[14] for the same data sets. Accuracy was determined by comparing the results to the well-established high-confident variant calls for the human individual NA12878 published by the Genome in a Bottle (GIAB) consortium[15]. As benchmarking method we used the area under the precision-recall curve (APR).

Accuracy of real-time results

In Illumina sequencing, all reads are sequenced in parallel. In each so-called sequencing cycle, sequence information of one additional nucleotide is obtained for all reads. Thus, the current length of a read equals the number of the respective cycle (e.g., 40 nucleotides after

cycle 40). To demonstrate the capability of our approach to provide interpretable results during runtime, we applied our workflow at different stages of sequencing. We expected our live results to show higher accuracy for higher cycles due to the increasing amount of available sequence information. At the same time, we analyzed whether the detected variants in early sequencing cycles are as reliable as variants called at the end of sequencing. This is a crucial criterion for the proposed workflow since interpretation of live results is only meaningful when based on reliable variant calls. Therefore, besides comparing the APR values of different sequencing cycles, we also examined precision and recall separately.

Fig. 1a shows the progression of the APR values for SNP calling in all analyzed data sets with increasing sequencing time. In cycle 30, sequence information was not sufficient to call any variants with the given parameter settings for six of seven data sets. For data set SRR292250, read mapping parameters were adapted by HiLive2 automatically due to the short read length of 50bp. This led to earlier results after 30 cycles, while first results were available after cycle 40 for all other data sets. Results show a continuous increase of the APR values for all cycles of the first read. In cycle 75, an APR larger than 0.9 was achieved for all data sets with sufficient read length. Afterwards, the APR values continue increasing moderately. When regarding the progression of precision (Fig. 1b) and recall (Fig. 1c) over sequencing time separately, it can also be also observed that lower APR values for earlier sequencing cycles are mainly caused by a lower recall while precision changes only slightly with more sequence information available. The same conclusions are supported by the individual precision-recall curves for all data sets which show a large increase of the recall but only minor changes of specificity over sequencing time (cf. Supplementary Fig. 1). This indicates that live results are highly reliable and can therefore serve for early interpretation and problem-specific follow-up analyses.

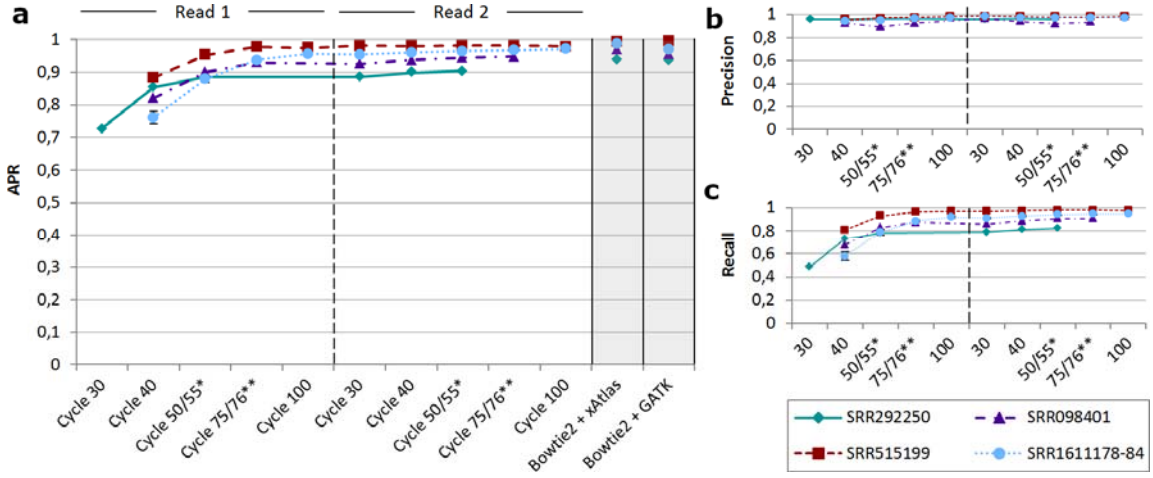

Fig. 1: Area under a precision-recall curve (APR) for SNP calling in seven data sets at different sequencing cycles. SNP calling was performed with xAtlas using real-time read mapping results of HiLive2. Results for the samples SRR1611178, SRR1611179, SRR1611183 and SRR1611184 were combined to a single data series due to their high similarity (SRR1611178-84). Error bars for this data series show the standard deviation. Reads of SRR292250 and SRR098401 were shorter than 2x100bp which leads to missing data points. The vertical, ticked line in the middle of the plot divides the first and second read. **a** The gray columns show APR values using Bowtie 2 for read mapping and xAtlas (left) and GATK-HC (right) for variant calling. The data for Bowtie 2 + GATK were taken from Hwang et al. (2015)[12]. The real-time workflow with HiLive2 and xAtlas provides first results after 40 sequencing cycles (30 cycles for SRR292250). An APR greater than 0.9 is reached after 75 cycles for all data sets with a minimal read length of 75bp. Until end of sequencing, there is a moderate increase of the APR. **b** Precision with a quality threshold of 1 for variant calling with xAtlas. The results show no precision lower than 0.89 for all sequencing cycles. In general, precision increases only slightly over time. This indicates that results in early sequencing cycles are already reliable. **c** Recall with a quality threshold of 1 for variant calling with xAtlas. The results show strong improvements from the first results available until the end of the first read. The progression of all curves is similar to that of the APR curve (cf. Fig. 1a), indicating the correlation between those two measures. *Cycle 50 for SRR292250, cycle 55 for all other data sets. **Cycle 76 for SRR098401, cycle 75 for all other data sets.

The increasing number of SNP calls in subsequent cycles provides additional information for complementing the previous interpretation of the data. However, the final results with HiLive2 show slightly lower maximum recall values than the same workflow applied to read mapping results of Bowtie 2 (cf. Supplementary Fig. 1). This can be explained by the read mapping approach of HiLive2 which tolerates only a specified number of errors for a read. Thus, regions with a high number of variations may be lowly covered which leads to undetected variants. The same effect is somewhat stronger for indels as the mapping algorithm only tolerates indels with a maximum length of three nucleotides by default due to computational costs. While this behavior led to a lower recall than based on read mapping with Bowtie 2, the results showed comparable or higher precision (cf. Supplementary Fig. 1). Thus,

although focussing on SNPs in this study, our workflow can also provide valuable insights about small indels.

Turnaround time of the workflow

Besides the accuracy of results, turnaround time is the second crucial factor for NGS-based real-time analyses. Thereby, live results should be available as soon as possible after the data of the respective sequencing cycle was written without showing significant delay in any stage of sequencing.

We measured the turnaround time of real-time mapping with HiLive2 and subsequent variant calling with xAtlas for the same runs that delivered the accuracy results shown before. All computations were run on a 128-core machine (Intel® Xeon® CPU E5-4667 v4 @ 2.20 GHz, 45 M Cache) with 500GB RAM, using a maximum of 65 threads per data set. Fig. 2 shows an overview for the turnaround time of our workflow for different sequencing

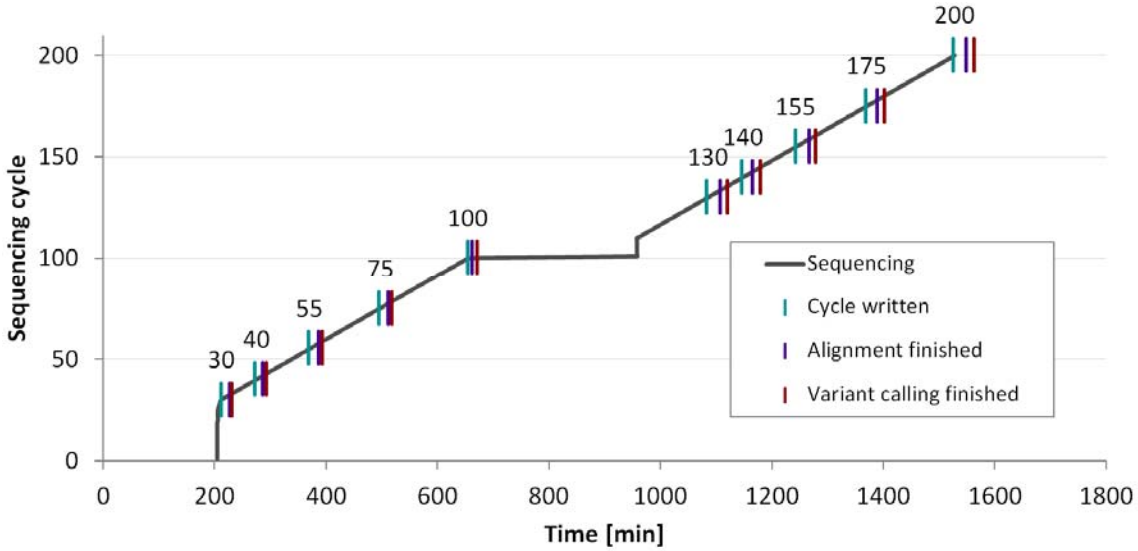

Fig. 2: Turnaround time of our workflow for data set SRR1611178. For each cycle, the first vertical line indicates the time point when the data for the respective cycle was completely written. The second line shows when the alignment output of HiLive2 is written. The third line indicates the end of our workflow resulting in the output of variant calls for the respective cycle. After cycle 100, there is no increase in the sequencing cycle for several hours due to the time spent for sequencing indices and initialization of the second read. In average, the time from data written by the sequencing machine and the output of variant calls is lower than 21 minutes for the first read and around 35 minutes for the second read. Final results are written 36 minutes after sequencing finished.

cycles for data set SRR1611178. On average, variant calling results were available less than half an hour after the data were written by the sequencing machine. The results for cycle 40 were written after 294 minutes, showing that first reliable and interpretable variant calls were available less than five hours after sequencing started. For higher coverage data sets, such as data set SRR515199, the alignment delay of HiLive2 and runtime of xAtlas increase when performed with the same number of threads (cf. Supplementary Fig. 2). However, five of the seven data sets in this study showed a maximum time span of less than one hour from data output to interpretable results for each sequencing cycle. The turnaround times for all data sets are shown in Supplementary Fig. 2.

Discussion

In clinical applications and infectious disease outbreaks, the turnaround time of analyses is a critical factor for an effective treatment of patients. At the same time, a high analysis depth and an open perspective for unexpected findings are further crucial criteria in such scenarios. Therefore, despite its significantly higher turnaround times than alternative methods, NGS presents an established analysis method in several time-critical applications. For example, a comprehensive report of vancomycin resistant *Enterococcus faecium* infections in three patients was created in 48.5 hours including over-night culturing using an Illumina MiSeq benchtop sequencer[16]. Also in the field of acutely ill infants with suspected genetic diseases, there were impressive improvements in the applicability of NGS-based diagnostics including a 26 hours protocol for provisional molecular diagnosis[3]. However, even with a significant speed-up of the computational analysis using faster software or specialized hardware such as field-programmable gate arrays (FPGAs), a decrease of turnaround time is strictly limited due to the sequential paradigm of data creation and analysis. Motivated by this, Miller et al. (2015) introduced the idea to combine bioinformatics analysis using FPGAs with the concept of rapid pulsed whole genome sequencing[4] to achieve near real-time analysis results[3]. However, to the best of our knowledge there is no study describing a proof-of-principle for this idea. Furthermore, such an approach would require specialized hardware while the

method proposed in our study runs on a standard linux machine and allows for easy and flexible adaption of the workflow for different scenarios while being highly scalable and providing high-quality analysis results.

The results of our study demonstrate the enormous potential of our approach to reduce the turnaround time from sample arrival to meaningful analysis output by several hours up to days depending on the used sequencing device. Thereby, live results in very early stages of sequencing can already deliver highly confident results while the quantity of analysis results (i.e. the number of called variants in this study) increases with a growing number of sequenced nucleotides per read. Live analyses can therefore provide first relevant insights into the data while the analysis becomes more comprehensive with ongoing sequencing. To demonstrate the power of our approach, we showed its application to human whole exome sequencing data including real-time alignment of all reads to the full human reference genome hg19. We chose this type of data due to its complexity and computational demands as well as the availability of high-quality and extensively studied gold-standard data sets provided by the GIAB consortium. However, our approach is not restricted to human whole exome sequencing data and the presented use case of variant calling. We rather see an enormous potential of real-time read mapping to provide means for a wide range of complex follow-up analyses for various types of data.

Alternative approaches to NGS for diagnosis can also be highly valuable for different scenarios. Molecular approaches are usually highly reliable and provide answers to specific questions in a very short timeframe and at much lower costs. For example, the detection of 25 genetic mutations in *M. tuberculosis* that confer to drug resistances can be finished in approximately two hours with a variation of the molecular GeneXpert test[5]. Even when providing live results, such short turnaround times are currently not feasible with NGS-based approaches due to the required time for sample preparation and clustering. Another interesting technology for time-critical applications is the Oxford Nanopore sequencing Technology (ONT). It was shown that metagenomic detection of viral pathogens can be

achieved in less than six hours[17]. While ONT shows a high portability as an additional benefit, this and other current long-read technologies are still expensive and limited by their comparatively low coverage and high error rates. It is therefore hard to reliably identify lowly abundant pathogens, genetic variants, parallel infections or the presence of viral quasispecies. Thus, especially when it comes to these or other questions going beyond the identification of highly abundant pathogens in time-critical applications, real-time analyses for Illumina sequencing can be of great benefit.

Concluding, we consider our new real-time workflow for Illumina sequencing to be a complementary method to molecular tests and ultra-portable, long-read sequencing for time-critical analyses. It fills the current gap of short turnaround times, an open-view perspective and high sequencing coverage which is essential for a plethora of applications such as pathogen identification and characterization, identification of acute genetic diseases or epidemiological analyses. Therefore, our approach is an important step for improving the ability for fast interventions in exceptional clinical situations, personalized medicine and infectious disease outbreaks.

Methods

In this chapter, we provide a brief description of our workflow. More detailed information for each step as well as direct links to the data and the executed commands can be found in the supplementary material. All software versions are listed in the *Software versions* section.

Implementation of HiLive2

HiLive2 is based on a novel algorithm based on the efficient FM-index implementation of the SeqAn library[11]. Each alignment starts with an exact match of a *k*-mer. The length of this *k*-mer is set according to the size of the reference genome and the read length or can be manually specified by the user. Starting from the first matching *k*-mer, the alignment is extended in the sequencing direction of the read. Apart from a front softclip due to *k*-mer

mismatches at the beginning of a read, the algorithm guarantees to produce optimal mapping results for non-affine gap costs. HiLive2 is an all-mapper by design meaning that all alignments up to a specified score can be found. However, the default output option of HiLive2 is to return only one best alignment for each read which is the expected behavior for most analyses. During output, temporary files are stored for all output cycles such that new output with the same or different output options can be created by a separate executable. Output is created in the well-established BAM or SAM format.

Data download and conversion

The human reference genome hg19 was obtained from NCBI and only considered chromosomes 1-22, X and Y. Alternative regions were omitted. The sequences were stored in a single multi-FASTA file. For the evaluation of variant calls with RTG Tools[18], the reference genome was converted to SDF format.

The sequencing data sets of the individual NA12878 were downloaded from EBI in FASTQ format. For read mapping with HiLive2, read pairs were converted to Illumina base call file format (BCL), distributed on one lane and 64 tiles. There were four different definitions for exome capture region definition required for the different data sets (cf. Table 1). The regions were obtained in BED format from the respective producer, if available. Whenever multiple definition files were provided, the primary target regions were selected.

Gold standard variants for the individual NA12878 were downloaded from the Genome in a Bottle (GIAB) consortium[15] and regularized with the vcfallelicprimitives tool of VCFtools[19]. SNPs and indels of the gold standard were stored in two separated files and filtered out against the exome capture regions using Bedtools[20] intersect. The resulting files were used as the gold standard for data sets using the respective exome capture definition. During the evaluation of the results, only variant calls in high confidence homozygous regions which were obtained from GIAB were considered.

Real-time read mapping with HiLive2

The index of human reference genome hg19 for HiLive2 was built with default parameters. The creation of base call files by the sequencing machine was simulated using a script for sequencing simulation with a sequencing profile for HiSeq2500 machines in rapid mode and using dual barcodes. As no barcodes were present in our data sets, no data was written by the sequencing simulator for the respective cycles. HiLive2 was run in fast mode allowing faster turnaround times at the expense of slightly lower recall. Technical parameters as lanes, tiles and read length were set according to the data sets. In general, we chose cycles 30, 40, 55, 75 and 100 for each of the two reads as output cycles. For data sets with read lengths other than 2 x 100bp, we adapted the output cycle numbers to 30, 40 and 50 (SRR292250) or 30, 40, 55 and 76 (SRR098401). We used the recommended number of threads (1 thread per tile) for HiLive2 resulting in 64 threads for all data sets.

Read mapping with Bowtie 2

The index of human reference genome hg19 for Bowtie 2 was built with default parameters. Read mapping with Bowtie 2 was performed with default parameters using 10 threads.

Variant calling with xAtlas

Variant calling with xAtlas was performed for each chromosome individually. Therefore, the alignment files of HiLive2 or Bowtie 2 were split in 24 files (one for each chromosome). The resulting files were sorted and indexed using samtools[21]. Afterwards, variants were called with xAtlas for the respective exome capture regions using default parameters. Sorting, indexing and variant calling was performed with 24 threads (1 per chromosome). The resulting VCF files were merged using VCFLIB vcf-concat (https://github.com/vcflib/vcflib) for SNVs and indels separately.

Measure of turnaround time

The sequencing simulation script provides timestamps for each written sequencing cycle. These timestamps were compared to the system time stamps for the last modification of the alignment output files of HiLive2. The time span between both time stamps describes the

alignment delay of HiLive2. Additionally, we measured the clock time of the xAtlas pipeline. The sum of sequencing time until the respective cycle, the alignment delay of HiLive2 and the clock time of xAtlas yields the overall turnaround times of our workflow.

Evaluation with RTG Tools

We used the vcfeval program of RTG Tools for the validation of variant calling results. We used the gold standard for the respective data set (depending on the used exome kit) as baseline and the variant calling output of xAtlas or GATK as call. The human reference hg19 in SDF format was used as reference template. Only variant calls being included in the high-confidence regions for individual NA12878 provided by the GIAB consortium were considered for validation. We ran RTG Tools with 24 threads and used the squash-ploidy and all-records parameters. For variant calls produced by xAtlas, we additionally defined QUAL as the field for variant call quality. For GATK, the GQ field is chosen by default. RTG Tools vcfeval returns a list of statistical measures for different thresholds of the variant call quality field, including precision and recall. These values served as input for the precision-recall curves shown in Supplementary Fig. 1 and used for the calculation of the area under the precision-recall curves (APR).

Statistical Measures

We used precision and recall values for the validation of our approach. True positives (TP) describe the number of correctly detected background-related reads (e.g. human). False negatives (FN) are the number of undetected background-related reads. False positives (FP) are the number of foreground-related reads that are classified as background-related and true negatives (TN) are foreground-related reads that are correctly not detected as background data.

Recall is the relative number of variants of the gold standard that were found by our approach (TP / (TP + FN)). Precision is the fraction of variants called by our approach that are also present in the gold standard (TP / (TP + FP)).

Software versions

Table 2 List of software used in this study. Software with source Bioconda was installed with the environment management software conda (https://conda.io) and obtained from the Bioconda channel[22].

Name	Version	Source	Used for
Bedtools[20]	2.21.0	https://github.com/arq5x/bedtools2	Vcf and Bed file processing
Bowtie 2[13]	2.3.4.1	Bioconda	Read alignment
GATK[14]	3.8	Bioconda	Alignment file processing
HiLive2	2.0	https://gitlab.com/rki_bioinformatics/hilive2	Real-time read alignment
RTG Tools[18]	3.9	Bioconda	Benchmark of variant calls
Samtools[21]	1.8	Bioconda	SAM/BAM file processing
VCFLIB	1.0.0_rc1	Bioconda	Vcf file processing
VCFtools[19]	0.1.12	https://sourceforge.net/projects/vcftools/	Vcf file processing
xAtlas[10]	0.1	Bioconda	Fast variant calling

Code availability

The source code of HiLive2 is available for public download on https://gitlab.com/rki_bioinformatics/hilive2 and comes with extensive documentation and sample data. Scripts that were used for our analyses are provided as Supplementary Software.

Data availability

Sequencing data of the individual NA12878 is publicly available on the NCBI Short Read Archive and on the EBI FTP server. Gold standard variant calls are publicly available from the Genome in a Bottle Consortium. Human reference genome hg19 was obtained from the NCBI FTP server. Exome capture targets are available from the manufacturers or from third party resources.

References

1. Soden SE*, et al.* Effectiveness of exome and genome sequencing guided by acuity of illness for diagnosis of neurodevelopmental disorders. *Science translational medicine* **6**, 265ra168-265ra168 (2014).

2. Saunders CJ*, et al.* Rapid Whole-Genome Sequencing for Genetic Disease Diagnosis in Neonatal Intensive Care Units. *Science Translational Medicine* **4**, 154ra135-154ra135 (2012).

3. Miller NA*, et al.* A 26-hour system of highly sensitive whole genome sequencing for emergency management of genetic diseases. *Genome medicine* **7**, 100-100 (2015).

4. Stranneheim H*, et al.* Rapid pulsed whole genome sequencing for comprehensive acute diagnostics of inborn errors of metabolism. *BMC Genomics* **15**, 1090 (2014).

5. Rubin EJ. TB diagnosis from the Dark Ages to fluorescence. *Nature microbiology* **3**, 268-269 (2018).

6. Quainoo S*, et al.* Whole-Genome Sequencing of Bacterial Pathogens: the Future of Nosocomial Outbreak Analysis. *Clinical microbiology reviews* **30**, 1015-1063 (2017).

7. Gilchrist CA, Turner SD, Riley MF, Petri WA, Jr., Hewlett EL. Whole-genome sequencing in outbreak analysis. *Clinical microbiology reviews* **28**, 541-563 (2015).

8. Lindner MS*, et al.* HiLive: real-time mapping of illumina reads while sequencing. *Bioinformatics* **33**, 917-919 (2017).

9. Tausch SH*, et al.* LiveKraken—real-time metagenomic classification of illumina data. *Bioinformatics* **34**, 3750-3752 (2018).

10. Farek J*, et al.* xAtlas: Scalable small variant calling across heterogeneous next-generation sequencing experiments. Preprint at https://www.biorxiv.org/content/biorxiv/early/2018/04/05/295071 (2018).

11. Doring A, Weese D, Rausch T, Reinert K. SeqAn an efficient, generic C++ library for sequence analysis. *BMC Bioinformatics* **9**, 11 (2008).

12. Hwang S, Kim E, Lee I, Marcotte EM. Systematic comparison of variant calling pipelines using gold standard personal exome variants. *Scientific reports* **5**, 17875 (2015).

13. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**, 357-359 (2012).

14. McKenna A*, et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20**, 1297-1303 (2010).

15. Zook JM*, et al.* Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. *Nature biotechnology* **32**, 246-251 (2014).

16. McGann P*, et al.* Real time application of whole genome sequencing for outbreak investigation - What is an achievable turnaround time? *Diagnostic microbiology and infectious disease* **85**, 277-282 (2016).

17. Greninger AL*, et al.* Rapid metagenomic identification of viral pathogens in clinical samples by real-time nanopore sequencing analysis. *Genome medicine* **7**, 99 (2015).

18. Cleary JG*, et al.* Comparing Variant Call Files for Performance Benchmarking of Next-Generation Sequencing Variant Calling Pipelines. Preprint at https://www.biorxiv.org/content/early/2015/08/03/023754 (2015).

19. Danecek P*, et al.* The variant call format and VCFtools. *Bioinformatics* **27**, 2156-2158 (2011).

20. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841-842 (2010).

21. Li H*, et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079 (2009).

22. Dale R*, et al.* Bioconda: A sustainable and comprehensive software distribution for the life sciences. Preprint at https://www.biorxiv.org/content/early/2017/10/27/207092 (2017).

Acknowledgements

This work was supported by the German Federal Ministry of Health [IIA5-2512-FSB-725 to B.Y.R. and 2515NIK043 to S.H.T.]. We thank Andrea Thürmer and Aleksandar Radonić for their input concerning technical aspects of Illumina sequencing. We further thank Wojciech Dabrowski for infrastructural support and Thilo Muth for critical reading of the manuscript and his highly valuable suggestions. We also thank Martin Lindner, Jakob Schulze, Benjamin Strauch, and Kristina Kirsten for their work on HiLive.

Author information

Simon H. Tausch

Present address: German Federal Institute for Risk Assessment (BfR), Department of Biological Safety, Berlin, Germany

Affiliations

Bioinformatics Division (MF 1), Department for Methods Development and Research Infrastructure, Robert Koch Institute, Berlin, Germany

Tobias P. Loka, Simon H. Tausch, Bernhard Y. Renard

Centre for Biological Threats and Special Pathogens: Highly Pathogenic Viruses (ZBS 1), Robert Koch Institute, Berlin, Germany

Simon H. Tausch

Contributions

B.Y.R. and T.P.L. conceived the study. T.P.L. performed the implementation of HiLive2. S.H.T. continuously supported the development of HiLive2 and evaluated the performance of HiLive2 on different types of data. T.P.L. designed the workflow and performed the analyses. All authors were involved in the preparation of the manuscript and approved the final version.

Competing interests

The authors declare no competing interests.

Corresponding authors

Correspondence to Bernhard Y. Renard.