Reliable variant calling during runtime of Illumina sequencing

Tobias P. Loka[1], Simon H. Tausch[1,2,3], 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)
[3] Present address: German Federal Institute for Risk Assessment (BfR), Department of Biological Safety, Berlin, Germany
* To whom correspondence should be addressed

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 designed a workflow using an advanced real-time read mapping approach to obtain reliable variant calls for human whole-exome data still during the sequencing process. When compared to standard routines, our live variant calling approach enables considerably faster interventions in clinical applications such as pathogen characterization and the determination of drug resistances in disease outbreaks or the design of individually tailored vaccines in precision medicine. Besides variant calling, our approach can be adapted for a plethora of other mapping-based analyses.

Keywords

Next Generation Sequencing, variant calling, real-time analysis, short read alignment, Illumina sequencing, pathogen identification, personalized medicine, drug resistance

Background

Common workflows for the analysis of 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 production of

individually tailored vaccines for cancer immunotherapy [2], to the determination of *M. tuberculosis* drug resistances [3], and to the identification of pathogens, virulence factors, drug resistances and paths of disease transmission in infectious disease outbreaks [4, 5]. While having considerably higher turnaround times than alternative 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, variant calling is a crucial step to find differences in the genomic sequence of the investigated sample when compared to a reference genome. Thereby, valuable information for the treatment of a patient can be obtained such as strain level classification and drug resistances of a pathogen or individual characteristics of healthy and defected tissue. To reduce the turnaround time for the generation of NGS-based results and to enable fast and accurate treatment of patients, we designed a workflow to obtain variant calling results before sequencing has finished. The workflow is based on real-time read mapping results with HiLive2 followed by fast and accurate variant calling with xAtlas. In doing so, 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.

Methods

In this chapter, we provide a description of our workflow and the methods for evaluation. More detailed information for each step such as direct links to the data and the executed commands can be found in the supplementary material. All software versions used in this study are listed in the *Software versions* section (Table 1).

Implementation of HiLive2

HiLive2 is based on a novel algorithm using the efficient FM-index implementation of the SeqAn library [6]. Each alignment starts with a short exact match sequence of length *k*, referred to as seed *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 seed *k*-mer, the alignment is extended in the sequencing direction of the read. Apart from a front soft clip that occurs 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 [7], the reference genome was converted to SDF format.

Seven whole-exome sequencing (WES) data sets of the human 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 2). 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 [8] and regularized with the vcfallelicprimitives tool of VCFtools [9]. SNPs and indels of the gold standard were stored in two separated files and filtered out against the exome capture regions using Bedtools [10] intersect. The resulting files were used as the gold standard for data sets using the respective exome capture definition. During

78 the evaluation of the results, only variant calls in high confidence homozygous regions which
79 were obtained from GIAB were considered.

80 **Real-time read mapping with HiLive2**

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

93 **Read mapping with Bowtie 2**

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

97 **Variant calling with xAtlas**

98 Variant calling with xAtlas was performed for each chromosome individually. Therefore, the
99 alignment files of HiLive2 or Bowtie 2 were split in 24 files (one for each chromosome). The
100 resulting files were sorted and indexed using samtools [11]. Afterwards, variants were called
101 with xAtlas for the respective exome capture regions using default parameters. Sorting,
102 indexing and variant calling was performed with 24 threads (one thread per chromosome).

The resulting VCF files were merged using VCFLIB vcf-concat [12] 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).

Software versions

All analyses performed in this study were done with the software versions listed in Table 1.

Table 1 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 [13].

Name	Version	Source	Used for
Bedtools [10]	2.21.0	https://github.com/arq5x/bedtools2	Vcf and Bed file processing
Bowtie 2 [14]	2.3.4.1	Bioconda	Read alignment
GATK [15]	3.8	Bioconda	Alignment file processing
HiLive2	2.0	https://gitlab.com/rki_bioinformatics/hilive2	Real-time read alignment
RTG Tools [7]	3.9	Bioconda	Benchmark of variant calls
Samtools [11]	1.8	Bioconda	SAM/BAM file processing
VCFLIB [12]	1.0.0_rc1	Bioconda	Vcf file processing
VCFtools [9]	0.1.12	https://sourceforge.net/projects/vcftools/	Vcf file processing
xAtlas [16]	0.1	Bioconda	Fast variant calling

Results

Implementation and experimental setup

To allow for faster NGS-based diagnosis and treatment, we developed a workflow 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 the real-time read mapper HiLive [17] that gave output at the end of sequencing, using a novel algorithm based on an efficient FM-index implementation 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 enable the combination of real-time read mapping results with complex follow-up analyses. To demonstrate the power of such analyses, we performed variant calling on seven WES data sets of the human individual NA12878 from the CEPH Utah Reference Collection (cf. Table 2) using the real-time read mapping results of HiLive2 as input data.

Table 2 Summary of data sets evaluated in this study. Information about sequencing platform, exome capture and coverage were adopted from Hwang et al. (2015) [1].

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 [16] which shows comparable accuracy to established methods at much lower runtime. We compared our results to read mapping with Bowtie 2 [14] and variant calling with either xAtlas or GATK HaplotypeCaller [15] 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 [8]. 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 our real-time workflow since interpretation of live results is only meaningful when based on reliable variant calls. Therefore, besides comparing

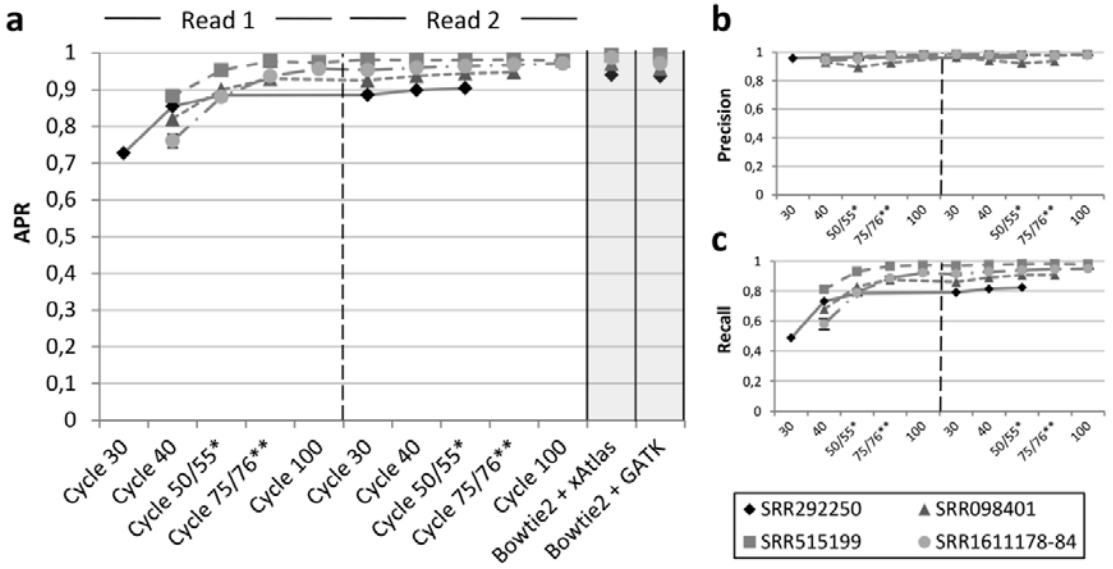

Fig. 1: Area under a precision-recall curve (APR) for SNP calling in seven data sets at different sequencing cycles. SNP calling was performed with xAtlas using real-time read mapping results of HiLive2. Results for the samples SRR1611178, SRR1611179, SRR1611183 and SRR1611184 were combined to a single data series (SRR1611178-84). Error bars for this data series show the standard deviation. 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) [1]. The real-time workflow with HiLive2 and xAtlas provides first results after 40 sequencing cycles (30 cycles for SRR292250). An APR value 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. 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 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

175 increasing moderately. When regarding the progression of precision (Fig. 1b) and recall (Fig.
176 1c) over sequencing time separately, it can also be also observed that lower APR values for
177 earlier sequencing cycles are mainly caused by a lower recall while precision changes only
178 slightly with more sequence information. The same conclusions are supported by the
179 individual precision-recall curves for all data sets which show a large increase of the recall
180 but only minor changes of specificity over sequencing time (cf. Supplementary Fig. 1). This
181 indicates that live results are highly reliable and can therefore serve for early interpretation
182 and problem-specific follow-up analyses. The increasing number of SNP calls in subsequent
183 cycles provides additional information for complementing the previous interpretation of the
184 data. However, the final results with HiLive2 show slightly lower maximum recall values than
185 the same workflow applied to read mapping results of Bowtie 2 (cf. Supplementary Fig. 1).
186 This can be explained by the read mapping approach of HiLive2 which tolerates only a
187 specified number of errors for a read. Thus, regions with a high number of variations may be
188 lowly covered which leads to undetected variants. The same effect is somewhat stronger for
189 indels as the algorithm only tolerates indels with a maximum length of three nucleotides by
190 default. While this behavior led to a lower recall than based on read mapping with Bowtie 2,
191 the results showed comparable or higher precision (cf. Supplementary Fig. 1). Thus,
192 although focussing on SNPs in this study, our workflow can also provide valuable insights
193 about small indels.

194 **Turnaround time of the workflow**

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

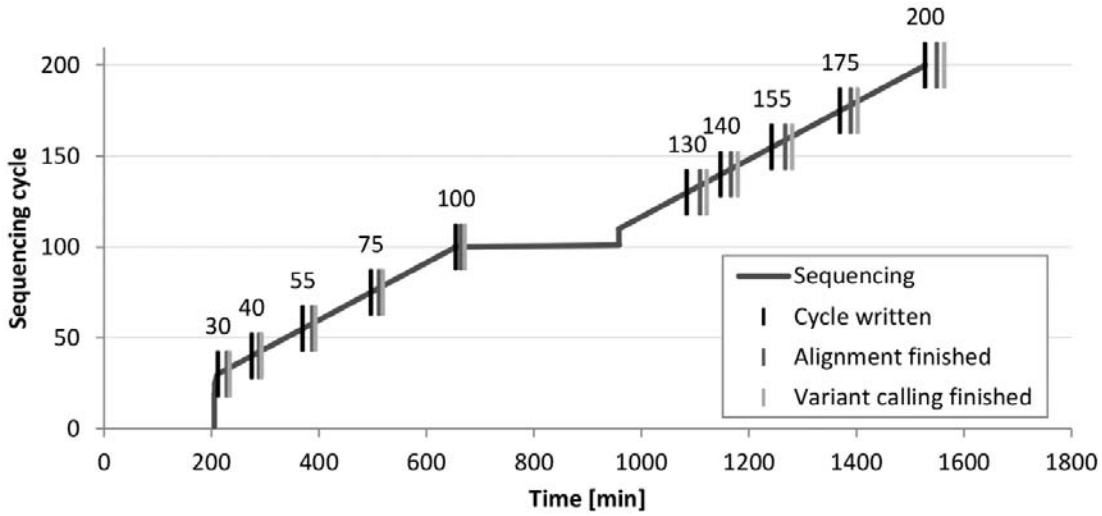

Fig. 2: Turnaround time of our workflow for data set SRR1611178. For each cycle, the first vertical line indicates the time point when the data for the respective cycle was completely written. The second line shows when the alignment output of HiLive2 is written. The third line indicates the end of our 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.

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 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

Our results show that real-time read mapping results in very early stages of sequencing can already serve as input for variant calling and deliver confident results. However, 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. Thereby, our approach does not only apply to the presented use case of variant calling. We rather see the enormous potential of real-time read mapping to provide means for a wide range of complex follow-up analyses.

In clinical applications and infectious disease outbreaks, the turnaround time of analyses is a critical factor for an effective treatment of patients. However, a high depth of analysis and an open perspective for unexpected findings are further crucial criteria in such scenarios. Despite its significantly higher turnaround times than alternative methods, NGS presents an established analysis method in several time-critical applications due to its high sensitivity. 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 [18]. A different study reports a workflow for *M. tuberculosis* diagnostics including phylogenetic placement being finished in 44h on an Illumina MiSeq or 16h on an Illumina MiniSeq [19]. To further accelerate such analyses, we showed that live results can deliver a major proportion of the full analysis depth already in a fraction of the final sequencing time. These results demonstrate the enormous potential of our approach to reduce the turnaround time from sample arrival to meaningful analysis output by several hours. However, alternative approaches 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 [3]. 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. However, NGS enables more detailed and unbiased analyses ranging from strain level identification to the determination of infection chains. Another interesting technology for time-critical applications is nanopore sequencing. It was shown that metagenomic detection of viral pathogens can be achieved in less than six hours [20]. While nanopore sequencing 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.

Conclusion

We consider our new real-time workflow for NGS 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, personalized vaccine design or epidemiological analyses. Therefore, we are convinced that our approach will improve the ability for fast interventions in exceptional clinical situations, personalized medicine and infectious disease outbreaks.

Abbreviations

APR: area under the precision-recall curve

NGS: next-generation sequencing

WES: whole exome sequencing

Declarations

Acknowledgements

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.

Funding

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.].

Availability of data and materials

The source code of HiLive2 is published under BSD-3-clause license and available for public download on https://gitlab.com/rki_bioinformatics/hilive2. It comes with extensive documentation and sample data. Scripts that were used for our analyses are provided as Supplementary Software.

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.

Authors' 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.

289 **Ethics approval and consent to participate**

290 Not applicable.

291 **Consent for publication**

292 Not applicable.

293 **Competing interests**

294 The authors declare no competing interests.

295 ## References

296 1. Hwang S, Kim E, Lee I, Marcotte EM. Systematic comparison of variant calling
297 pipelines using gold standard personal exome variants. Scientific reports
298 2015;5:17875.

299
300 2. Sahin U, Tureci O. Personalized vaccines for cancer immunotherapy. Science
301 2018;359:1355-1360.

302
303 3. Rubin EJ. TB diagnosis from the Dark Ages to fluorescence. Nature microbiology
304 2018;3:268-269.

305
306 4. Quainoo S, Coolen JPM, van Hijum S, Huynen MA, Melchers WJG, van Schaik W,
307 Wertheim HFL. Whole-Genome Sequencing of Bacterial Pathogens: the Future of
308 Nosocomial Outbreak Analysis. Clinical microbiology reviews 2017;30:1015-1063.

309
310 5. Gilchrist CA, Turner SD, Riley MF, Petri WA, Jr., Hewlett EL. Whole-genome
311 sequencing in outbreak analysis. Clinical microbiology reviews 2015;28:541-563.

312
313 6. Doring A, Weese D, Rausch T, Reinert K. SeqAn an efficient, generic C++ library for
314 sequence analysis. BMC Bioinformatics 2008;9:11.

315
316 7. Cleary JG, Braithwaite R, Gaastra K, Hilbush BS, Inglis S, Irvine SA, Jackson A, Littin
317 R, Rathod M, Ware D *et al*. Comparing Variant Call Files for Performance
318 Benchmarking of Next-Generation Sequencing Variant Calling Pipelines. 2015;
319 doi:10.1101/023754.

320
321 8. Zook JM, Chapman B, Wang J, Mittelman D, Hofmann O, Hide W, Salit M. Integrating
322 human sequence data sets provides a resource of benchmark SNP and indel
323 genotype calls. Nature biotechnology 2014;32:246-251.

324

9. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST *et al*. The variant call format and VCFtools. Bioinformatics 2011;27:2156-2158.

10. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 2010;26:841-842.

11. 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.

12. Garrison E, Yandell M, Shapiro M, Marth G, Durbin R, Eichler EE, Consortium TV, Kronenberg ZN. VCFLIB: an ensemble of methods for variant manipulation and population genetics. 2016. https://github.com/vcflib/vcflib. Accessed 19.05.2018.

13. Dale R, Grüning B, Sjödin A, Rowe J, Chapman BA, Tomkins-Tinch CH, Valieris R, Köster J. Bioconda: A sustainable and comprehensive software distribution for the life sciences. bioRxiv 2017; doi:10.1101/207092.

14. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods 2012;9:357-359.

15. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M *et al*. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 2010;20:1297-1303.

16. Farek J, Hughes D, Mansfield A, Krasheninina O, Nasser W, Sedlazeck FJ, Khan Z, Venner E, Metcalf G, Boerwinkle E *et al*. xAtlas: Scalable small variant calling across heterogeneous next-generation sequencing experiments. bioRxiv 2018; doi:10.1101/295071.

17. Lindner MS, Strauch B, Schulze JM, Tausch SH, Dabrowski PW, Nitsche A, Renard BY. HiLive: real-time mapping of illumina reads while sequencing. Bioinformatics 2017;33:917-919.

18. McGann P, Bunin JL, Snesrud E, Singh S, Maybank R, Ong AC, Kwak YI, Seronello S, Clifford RJ, Hinkle M *et al*. Real time application of whole genome sequencing for outbreak investigation - What is an achievable turnaround time? Diagnostic microbiology and infectious disease 2016;85:277-282.

19. Votintseva AA, Bradley P, Pankhurst L, Del Ojo Elias C, Loose M, Nilgiriwala K, Chatterjee A, Smith EG, Sanderson N, Walker TM *et al*. Same-Day Diagnostic and Surveillance Data for Tuberculosis via Whole-Genome Sequencing of Direct Respiratory Samples. Journal of clinical microbiology 2017;55:1285-1298.

20. Greninger AL, Naccache SN, Federman S, Yu G, Mbala P, Bres V, Stryke D, Bouquet J, Somasekar S, Linnen JM *et al*. Rapid metagenomic identification of viral pathogens in clinical samples by real-time nanopore sequencing analysis. Genome medicine 2015;7:99.