RESEARCH ARTICLE

REVISED Evaluating the performance of tools used to call minority variants from whole genome short-read data [version 2; referees: 2 approved]

Khadija Said Mohammed[1,2], Nelson Kibinge [2], Pjotr Prins[2,3], Charles N. Agoti [1,2], Matthew Cotten [4], D.J. Nokes [2,5], Samuel Brand [5], George Githinji [2]

[1]Pwani University, Kilifi, Kenya
[2]KEMRI-Wellcome Trust Research Programme, KEMRI Centre for Geographic Medicine Research – Coast, Kilifi, Kenya
[3]University Medical Center Utrecht, Utrecht, The Netherlands
[4]Virosciences Department, Erasmus Medical Centre, Rotterdam, The Netherlands
[5]School of Life Sciences and Zeeman Institute (SBIDER), University of Warwick, Coventry, UK

v2 **First published:** 05 Mar 2018, **3**:21 (https://doi.org/10.12688/wellcomeopenres.13538.1)
Latest published: 13 Sep 2018, **3**:21 (https://doi.org/10.12688/wellcomeopenres.13538.2)

Abstract

Background: High-throughput whole genome sequencing facilitates investigation of minority virus sub-populations from virus positive samples. Minority variants are useful in understanding within and between host diversity, population dynamics and can potentially assist in elucidating person-person transmission pathways. Several minority variant callers have been developed to describe low frequency sub-populations from whole genome sequence data. These callers differ based on bioinformatics and statistical methods used to discriminate sequencing errors from low-frequency variants.
Methods: We evaluated the diagnostic performance and concordance between published minority variant callers used in identifying minority variants from whole-genome sequence data from virus samples. We used the ART-Illumina read simulation tool to generate three artificial short-read datasets of varying coverage and error profiles from an RSV reference genome. The datasets were spiked with nucleotide variants at predetermined positions and frequencies. Variants were called using FreeBayes, LoFreq, Vardict, and VarScan2. The variant callers' agreement in identifying known variants was quantified using two measures; concordance accuracy and the inter-caller concordance.
Results: The variant callers reported differences in identifying minority variants from the datasets. Concordance accuracy and inter-caller concordance were positively correlated with sample coverage. FreeBayes identified the majority of variants although it was characterised by variable sensitivity and precision in addition to a high false positive rate relative to the other minority variant callers and which varied with sample coverage. LoFreq was the most conservative caller.
Conclusions: We conducted a performance and concordance evaluation of four minority variant calling tools used to identify and quantify low frequency variants. Inconsistency in the quality of sequenced samples impacts on sensitivity and accuracy of minority variant callers. Our study suggests that

Open Peer Review

Referee Status: ✓ ✓

	Invited Referees 1	2
REVISED version 2 published 13 Sep 2018	✓ report	✓ report
version 1 published 05 Mar 2018	? report	? report

1 **Brad A. Chapman**, Harvard Public School of Health, USA

2 **Micha M. Bayer**, James Hutton Institute, UK

Discuss this article

Comments (0)

combining at least three tools when identifying minority variants is useful in filtering errors when calling low frequency variants.

Keywords
variant calling, minority variants, concordance, performance, RSV

This article is included in the KEMRI | Wellcome Trust gateway.

Corresponding author: George Githinji (GGithinji@kemri-wellcome.org)

Author roles: Said Mohammed K: Conceptualization, Data Curation, Methodology, Software, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Kibinge N**: Conceptualization, Data Curation, Methodology, Writing – Review & Editing; **Prins P**: Investigation, Writing – Review & Editing; **Agoti CN**: Investigation, Writing – Review & Editing; **Cotten M**: Investigation, Writing – Review & Editing; **Nokes DJ**: Conceptualization, Funding Acquisition, Investigation, Methodology, Project Administration, Writing – Review & Editing; **Brand S**: Formal Analysis, Methodology, Writing – Review & Editing; **Githinji G**: Conceptualization, Data Curation, Investigation, Methodology, Project Administration, Software, Supervision, Visualization, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: The work was funded by the Wellcome Trust Senior Investigator Award to Prof D. James Nokes [102975] in addition, this work was supported through the DELTAS Africa Initiative [DEL-15-003]. The DELTAS Africa Initiative is an independent funding scheme of the African Academy of Sciences (AAS)'s Alliance for Accelerating Excellence in Science in Africa (AESA) and supported by the New Partnership for Africa's Development Planning and Coordinating Agency (NEPAD Agency) with funding from the Wellcome Trust [107769] and the UK government. The views expressed in this publication are those of the author(s) and not necessarily those of AAS, NEPAD Agency, Wellcome Trust or the UK government
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

How to cite this article: Said Mohammed K, Kibinge N, Prins P *et al.* **Evaluating the performance of tools used to call minority variants from whole genome short-read data [version 2; referees: 2 approved]** Wellcome Open Research 2018, **3**:21 (https://doi.org/10.12688/wellcomeopenres.13538.2)

First published: 05 Mar 2018, **3**:21 (https://doi.org/10.12688/wellcomeopenres.13538.1)

REVISED **Amendments from Version 1**

This version of the manuscript has been revised to consider the reviewers' comments and suggestions based on the initial version 1 of the paper. We have modified the main text, the figures and removed two tables that contained information or data that could be described in the main text.

The main changes are:

- Removed table 1 and table 2 from the main text and provided description of the same information in the main text.
- Revised table 3 to reflect proposed corrections.
- Revised Figure 4 to reflect amended changes.
- Added a new figure, Figure 5 as per reviewers' suggestion.
- Revised the initial R code to correct a bug that resulted in miscalculation of TP and TN for FreeBayes estimates.
- Revised the supplementary materials to include explicit methods used in the variant calling procedures.

These changes have been clarified further in responses to the reviewers' specific comments.

See referee reports

Introduction

RNA viruses have been described as a population of closely related sequences that arise from rapid genomic evolution coupled with a high replication and mutation rates (Domingo *et al.*, 2012; Eigen *et al.*, 1988; Holland *et al.*, 1992). Genetic changes in RNA viruses result from genetic drift, erroneous replication processes, mutagenic agents and upon which natural selection acts (Moya *et al.*, 2004). Rapid replication and mutations generate an ensemble of mutant genomes that are comprised of both dominant and low frequency variants. This diversity has been shown to affect virus fitness landscape, transmission, colonization and replication (Henn *et al.*, 2012; Stack *et al.*, 2013; Vignuzzi *et al.*, 2006).

Many recent studies (Henn *et al.*, 2012; Poon *et al.*, 2016; Stack *et al.*, 2013) have demonstrated the potential application of virus diversity to inform person-to-person transmission during virus outbreaks. A number of methods that incorporate both genomic and epidemiologic data to infer pathogen transmission have recently been developed (Worby *et al.*, 2017). These approaches rely partly on the accurate detection and quantification of minority variant populations from genomic samples.

Several tools have been developed to identify and quantify minority variants from short-read data (Koboldt *et al.*, 2009; Koboldt *et al.*, 2012; Lai *et al.*, 2016; Macalalad *et al.*, 2012; Wilm *et al.*, 2012; Yang *et al.*, 2013). Nonetheless, these tools do not fully account for discrepancies that arise from sample collection, pre-processing and sequencing in addition to errors that are introduced during downstream bioinformatic analysis. Rigorous quality control in sample processing and analysis is often suggested to distinguish true biological variants from artefactual variants (Zhang & Flaherty, 2017). In some cases, sequencing errors can be reduced by developing high-fidelity protocols and laboratory quality control measures (Kinde *et al.*, 2011; McCrone & Lauring, 2016; Watson *et al.*, 2013). Additionally, the uncertainty resulting from random sequencing errors can be countered by sequencing larger populations at higher coverage (Zukurov *et al.*, 2016). A number of studies have extensively explored variants from somatic or tumour samples (Hofmann *et al.*, 2017; Koboldt *et al.*, 2012; Krøigård *et al.*, 2016; Lai *et al.*, 2016; Pabinger *et al.*, 2014) and their application in clinical genomics, but only a limited number of studies have explored the nature of variants from patient-derived samples that target viral populations (Henn *et al.*, 2012; Macalalad *et al.*, 2012; Wilm *et al.*, 2012; Yang *et al.*, 2013; Zukurov *et al.*, 2016) and especially when calling variants from respiratory viruses such as the respiratory syncytial virus (RSV).

In this study, we evaluated four published minority variant detection tools using artificial short-read data with different error profiles. We explored the tools' ability to detect and quantify minority variants and assessed their overall agreement which we defined using two metrics, concordance accuracy, which measures the combined accuracy of the variant callers, and inter-caller concordance, which is the size of the largest set of variant callers that agree at each position. We show that concordance metrics are dependent on sample coverage and are influenced by the quality of input data.

Methods

Overall, we considered ten published, open-source tools with presumed ability to call minority variants from virus deep sequence data. A number of callers were excluded from the analysis for various reasons, for example, the GATK HaplotypeCaller primarily targets germline calling from human and not variant calling from viral samples. We experienced technical difficulties in setting up the Platypus caller and even after setup, Platypus did not provide calls across all levels of coverage in our datasets. SAMtools mpileup did not provide direct allele frequencies while V-Phaser was superseded by V-Phaser 2 which has reported bugs and could not handle reads aligned with BWA-MEM. Therefore, the following four tools were evaluated, FreeBayes version 1.1.0-3-g961e5f3, LoFreq version 2.1.2, VarDict version 30.3.17 and VarScan version 2.4.2. A schematic diagram showing the overall approach is shown in Figure 1.

Artificial datasets

Artificial datasets were generated based on an RSV reference sequence (GenBank accession number KX510245.1) using ART-Illumina version 2.5.8 (Huang *et al.*, 2012). ART-Illumina was took the reference RSV genome sequence as input and generated artificial reads using data derived error models to mimic sequence data. Each dataset comprised of eight samples with varying depth of coverage (20, 50, 100, 500, 1000, 2000, 5000, 10000) and was generated using the methods described in Supplementary File 1, section S1.1. The first dataset did not incorporate an error profile. Error profile models (empirical error

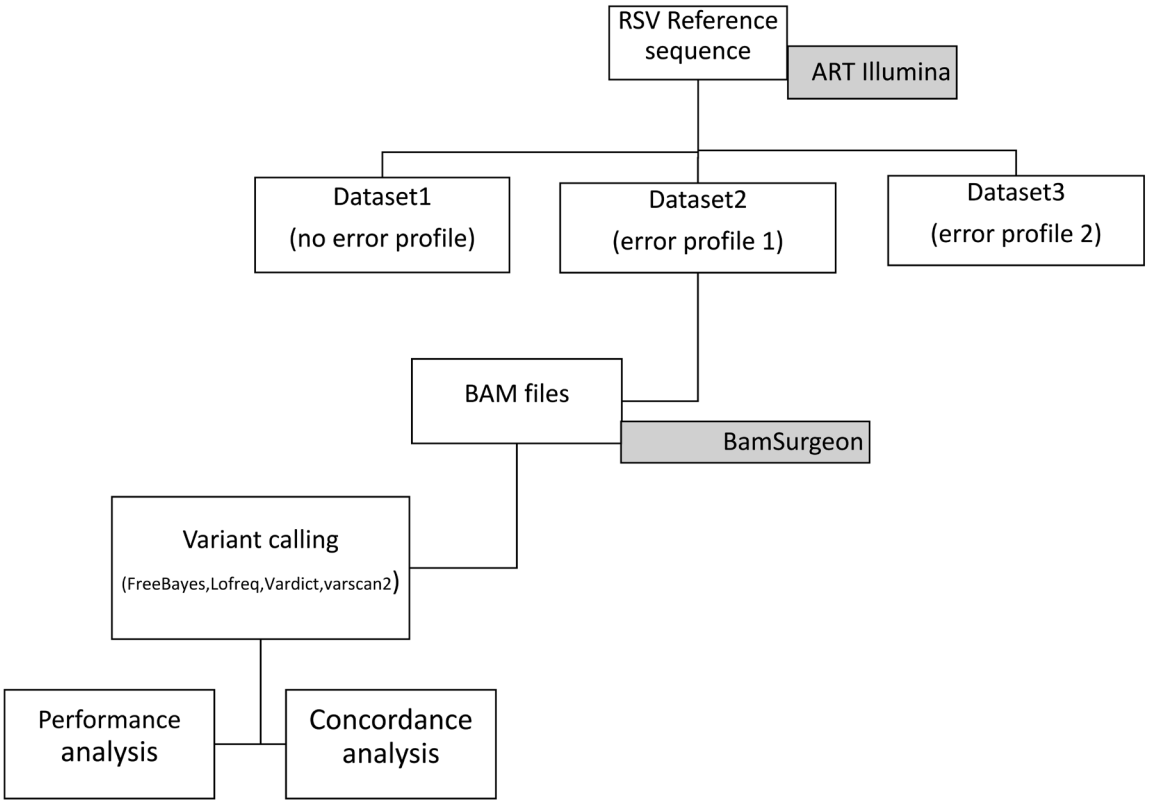

Figure 1. A schematic diagram showing the variant calling workflow. The artificial datasets (BAM files) were generated using ART-Illumina based on an RSV reference genome. BAMSurgeon was used to spike the resulting BAM files by inserting known variants at known locations across the artificial BAM file.

models based on the distribution of base quality scores) were generated from uncompressed FastQC raw reads derived from sequenced RSV whole genome samples, referred to as "good" and a "bad" sample based on FastQC metrics (Supplementary File 2 and Supplementary File 3) and used to generate artificial reads in dataset 2 and 3. ART-Illumina-generated artificial SAM files were converted to the BAM format, sorted and indexed using SAMtools version 1.3.1 for each dataset.

166 randomly and uniformly generated nucleotide mutations at different frequencies (Supplementary Table 1) were inserted into each of the artificial datasets using BamSurgeon (Ewing *et al.*, 2015), such that a base change was made amongst the reads at each alignment position as described in Supplementary File 1, section S1.2. This process was repeated with a separate set of 155 positions that comprised a set of mutations with frequencies below 0.5 (Supplementary Table 1).

Variant calling

The BAM files from each of the three datasets were used as input to each of the four variant callers (FreeBayes, Lofreq, VarDict and VarScan2). The default parameter options used in each tool are explicitly provided in Supplementary File 4. All output files were provided in the variant call format (VCF) or as a tabular file for the case of VarDict. The output from the VCF and tabular file was parsed and written as a comma separated (CSV) file.

Performance measures

To evaluate the performance of the variant calling algorithms, we compared the sequence generated by each variant caller *vc*, denoted $S^{vc} = (S_i^{vc} \in \{A,C,T,G\})_{i=1,\ldots,N}$, to the gold standard "spiked" sequence, denoted S^{true}, at each of N=15205 nucleotide positions. The accuracy of each variant caller is the normalized Hamming distance from the gold standard sequence, $\frac{1}{N}\sum_{i=1}^{N} d(S_i^{true}, S_i^{v})$, where d(x,y) is the standard discrete metric giving 1 when x=y, and 0 otherwise. By distinguishing between the sets of positions where variants did and did not occur in the gold standard sequence we calculated sensitivity, specificity, precision and accuracy (Table 1).

Concordance analysis

We defined two concordance metrics to present the level of agreement between different callers in detecting the same variant positions in the sequence. The first concordance metric is concordance accuracy, which measures the combined accuracy of the variant callers. At the true variant position i we then $C_{acc}(i) = \sum_{vc=1}^{4} d(S_i^{true}, S_i^{vc})$, which can be either 0, 1, 2, 3, 4

Table 1. A breakdown of performance metrics of variant callers evaluated from first dataset that did not incorporate an error profile. The samples represent simulated datasets of varying depth of coverage. True positive (TP), true negative (TN), false positive (FP) and false negatives (FN) were used to calculate performance metrics of each caller. FPR – False positive rate.

Sample	Caller	TP	TN	FP	FN	Sensitivity	Specificity	Precision	FPR	Accuracy
1 (20X)	freebayes	87	15018	21	79	0.5241	0.9986	0.8056	0.0014	0.9934
	lofreq	29	15039	0	137	0.1747	1	1	0	0.991
	vardict	72	15039	0	94	0.4337	1	1	0	0.9938
	varscan	11	15039	0	155	0.0663	1	1	0	0.9898
2 (50X)	freebayes	118	14901	138	47	0.7152	0.9908	0.4609	0.00918	0.9878
	lofreq	67	15039	0	99	0.40361	1	1	0	0.9935
	vardict	108	15039	0	58	0.6506	1	1	0	0.9962
	varscan	22	15039	0	144	0.13253	1	1	0	0.9905
3 (100X)	freebayes	127	14454	585	38	0.7697	0.9611	0.1784	0.0389	0.959
	lofreq	57	15039	0	109	0.3434	1	1	0	0.9928
	vardict	104	15038	1	62	0.6265	0.9999	0.9905	6.65E-05	0.9959
	varscan	40	15039	0	126	0.241	1	1	0	0.9917
4 (500X)	freebayes	131	12559	2480	30	0.8137	0.8351	0.0502	0.1649	0.8349
	lofreq	60	15039	0	106	0.3614	1	1	0	0.993
	vardict	110	15029	10	56	0.6627	0.9993	0.9167	6.65E-04	0.9957
	varscan	73	15039	0	93	0.4398	1	1	0	0.9939
5 (1000X)	freebayes	146	14414	625	20	0.8795	0.9584	0.1894	0.0416	0.9576
	lofreq	57	15039	0	109	0.3434	1	1	0	0.9928
	vardict	109	15036	3	57	0.6567	0.9998	0.9732	1.99E-04	0.9961
	varscan	79	15039	0	87	0.4759	1	1	0	0.9943
6 (2000X)	freebayes	146	14923	116	20	0.8795	0.9923	0.5571	0.0077	0.9911
	lofreq	70	15039	0	96	0.4217	1	1	0	0.9937
	vardict	120	15039	0	46	0.7229	1	1	0	0.997
	varscan	83	15039	0	83	0.5	1	1	0	0.9945
7 (5000X)	freebayes	149	15020	19	17	0.8976	0.9987	0.8869	0.0013	0.9976
	lofreq	67	15039	0	99	0.40366	1	1	0	0.9935
	vardict	117	15036	3	49	0.7048	0.9998	0.975	1.99E-04	0.9966
	varscan	78	15039	0	88	0.4699	1	1	0	0.9942
8 (10000X)	freebayes	145	15022	17	21	0.8735	0.9989	0.8951	0.0011	0.9975
	lofreq	72	15039	0	94	0.4337	1	1	0	0.9938
	vardict	118	15038	1	48	0.7108	0.9999	0.9916	6.65E-05	0.9968
	varscan	97	15039	0	69	0.5843	1	1	0	0.9955

for each true variant position. The second concordance metric is inter-caller concordance, which is the size of the largest set of variant callers that agree at each position *i*, without reference to any gold standard sequence. We used both bar plots and heat maps to visualize the effect of coverage on C_{acc}. Visualization of inter-caller concordance for variant sets was achieved using a bar plot and expounded by UpSet plots (Lex *et al.*, 2014) in R version 3.4.2.

Results

We used three artificial datasets of varying coverage and error profile to assess the concordance accuracy and inter-caller concordance for four minority variant callers. The first dataset comprised of artificial reads based on an RSV genome, the second dataset comprised of the similar simulated set of reads whilst incorporating an error profile from the set of reads used to assemble the reference genome, the third dataset was

generated using an error profile from a poorly sequenced sample. Overall, concordance accuracy improved with increase in sample coverage (Figure 2), and the proportion of positions that could not be identified by any variant caller decreased with increase in coverage (Supplementary Figure 1). For all the three datasets, and at each coverage level, fully concordant variants were below 50% of the total variants suggesting that considering only fully concordant positions eliminated a substantial number of variant positions. There were marginal improvements in the number of concordant variants in the second dataset compared to the first and the third error profile. Across all datasets, there was little improvement at detecting fully concordant positions after a coverage of 2000 (Figure 2). We utilized UpSetR plots to provide a visual summary of the combination of variant callers that contributed to the observed concordance accuracy (Supplementary Figure 4).

FreeBayes identified the majority of variants (Figure 3) across all the datasets although it was characterised by a substantial trade-off between sensitivity and precision in artificial dataset 1 (Figure 4) in addition to a high false positive rate relative to the other minority variant callers observed in datasets 1 and 3. Regardless, Freebayes reported comparatively better sensitivity relative to the rest of the tools. Lofreq was the most conservative of the evaluated callers and it missed majority of variants across all the three datasets. In addition, Lofreq's sensitivity in coverages above 100 did not differ in a substantial way (Figure 4). Vardict performance increased with read coverage but not by a great magnitude compared to FreeBayes and Varscan. Its performance across different datasets was more consistent relative to other tools. Overall, we observed lower reported frequency in called minority variants compared with spiked frequencies (Figure 5). This observation was consistent in all the datasets from all the variant callers.

Discussion

Detecting and reporting minority variant calls is challenging, given that low frequency calls occur at a frequency that is the same as error generated from sequencing and PCR reactions. Recent studies have linked the sharing of minority variants with transmission patterns, and hence it is important to distinguish actual minor variants from spurious variant calls. Several minority variants callers use different detection algorithms and statistics, each of which attempt to optimize an aspect of the variant calling process. Therefore, there could be disparities between what is reported by a given minority variant caller, given datasets of varying sequencing depths and error profiles.

This study aimed to identify the proportion of positions that were recognized as variants by a set of tools using three artificial datasets of varying coverage and error profiles. Concordance accuracy and inter-caller concordance measures were dependent on the sample coverage and error profile.

Sensitivity for the majority of the tools was positively correlated with depth of coverage and similarly observed previously (Spencer *et al.*, 2014) in a study that investigated performance in methods used to detect low-frequency variants. It is important to note that the tools provide different performance metrics depending on the variant's threshold (Supplementary File 5).

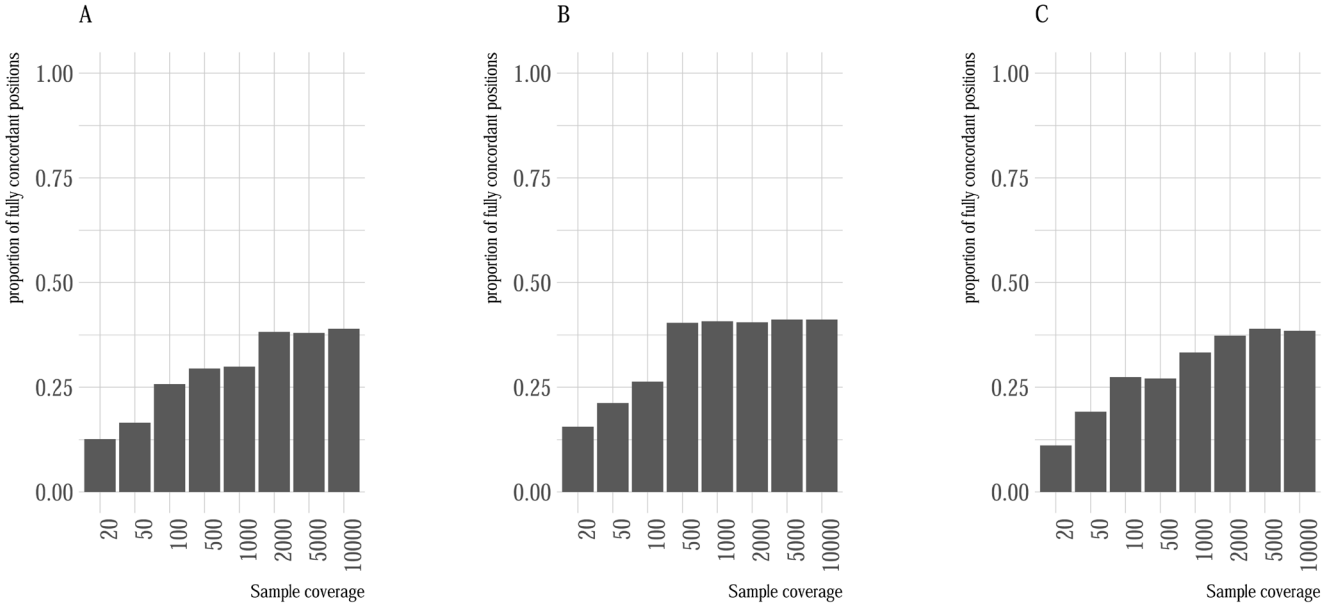

Figure 2. Proportion of fully concordant positions with respect to sample coverage. Each plot **A**–**C** represents the proportion (y-axis) of fully concordant variants with respect to read coverage (x-axis) for the first, second and third dataset. Concordant positions were defined as positions that were identified by all the four variant callers.

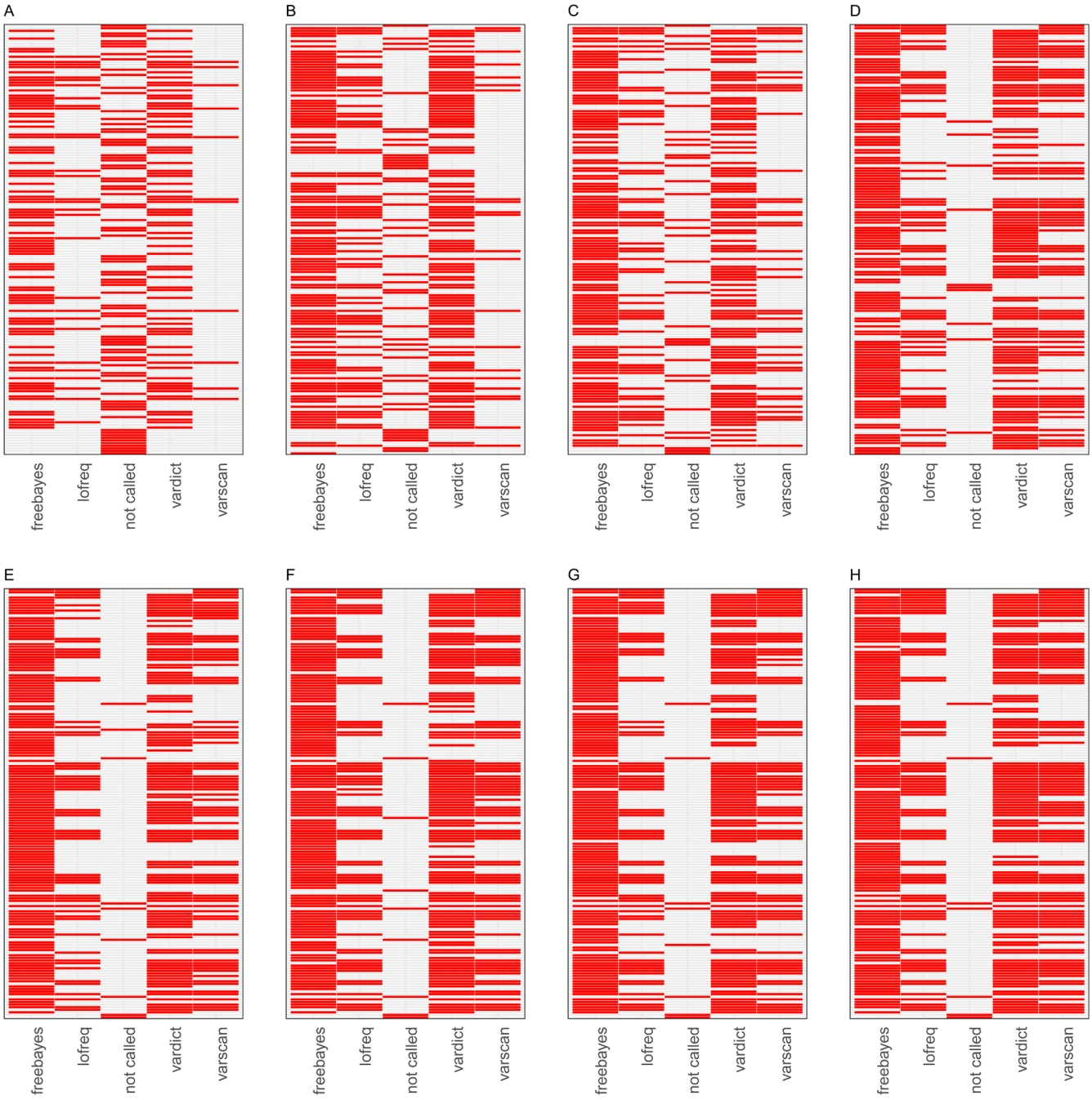

Figure 3. Heat maps illustrating tool specific concordance for the first artificial dataset. The red tiles represent variants detected by each caller from the list of 166 variant positions. The panels are arranged left to right **A**–**H** in the order of increasing sample coverage (20,50,100,500,1000,2000,5000 and 10,000). The "not called" column in each panel represents the variants that were not identified by any of the variant callers.

In the first artificial dataset, VarDict detected true positive variants with comparably good performance (sensitivity 43.4% – 72.3%), though it was marginally invariant to changes in average coverage above 20. VarDict has in-built features that could contribute to its efficient performance. It is able to activate an "amplicon calling mode" that filters out amplicon biased variants and mispaired primers as PCR artefacts. A similar pattern was observed with LoFreq, where sensitivity was not significantly affected by depth of coverage. VarScan2 was more affected by coverage and maintained average sensitivity

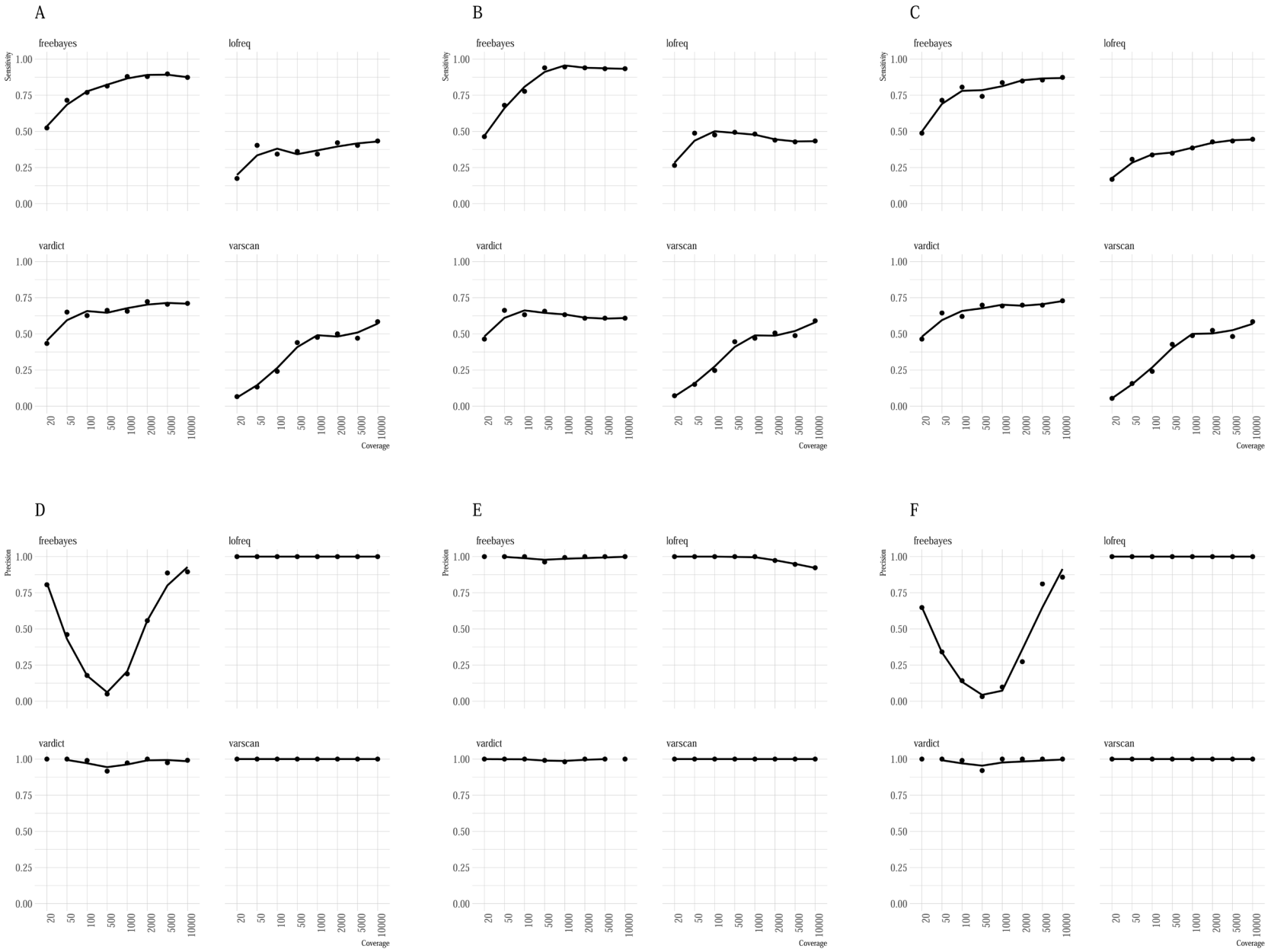

Figure 4. A summary of the relationship between sample coverage, sensitivity (**A**–**C**) and sample coverage and precision (**D**–**F**). The x-axis shows the sample coverage and the y-axis represents the sensitivity and precision respectively. Sensitivity of the callers rose gradually from low to high coverage samples. Again, precision was variable for FreeBayes calls while relatively high for the rest of the three callers.

Wellcome Open Research 2018, 3:21 Last updated: 13 NOV 2018

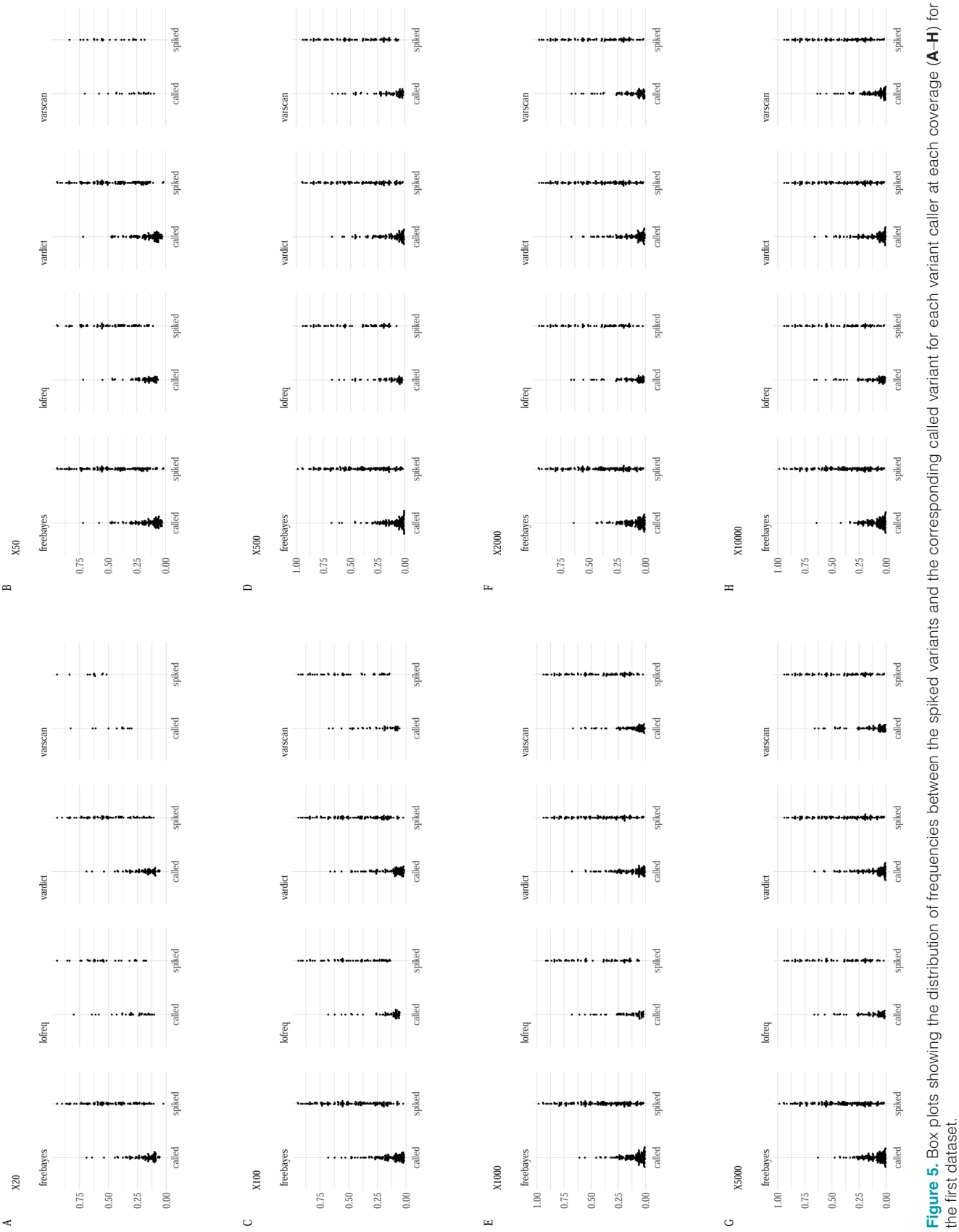

Figure 5. Box plots showing the distribution of frequencies between the spiked variants and the corresponding called variant for each variant caller at each coverage (**A**–**H**) for the first dataset.

(6.6% – 58.4%). Applying filters in VarScan has been reported to improve sensitivity by reducing number of false positives (Hofmann *et al.*, 2017; Koboldt *et al.*, 2013). FreeBayes' trade-off between sensitivity and precision was also reported by other studies (Hwang *et al.*, 2015; Sandmann *et al.*, 2017). The use of caller-specified filters could have enhanced the sensitivity of some callers, but the option to adopt default parameters allows equivalent assessment of tool performance. Moreover, all possible combinations of tuning parameters are challenging, time-consuming and sometimes impractical.

Based on artificial reads from the second dataset, FreeBayes performed comparatively better than the other tools with a very low false positive rate and better sensitivity (46.4% – 94.6%). This suggests that FreeBayes is potentially useful in identifying minority variants when sample data comes from reads with a low error profile. This implies the error rate results are outcome of tool performance.

Specificity of a caller is its ability to correctly predict the absence of a variant. The variant callers make use of a high specificity to minimize the number of false positive calls thereby reducing post-call filtering and consequently filter out true low-frequency variants. Moreover, high accuracy measures demonstrate the reliability of the variant caller in correctly identifying true variants.

All the minority variant callers reported slightly lower frequencies in called variants compared to the frequencies in the original spiked variants (Figure 5). This could be explained by the fact that many of the callers are tuned to report lower differences in the calls owing to stringent pre-processing criteria. A thorough investigation of this observation is therefore required.

In absence of an explicit error model from samples of heterogeneous sequencing quality, combining at least three tools when identifying minority variants could potentially assist in filtering out errors from low frequency variants. Given that there are no definitive data and next generation sequencing pipeline standards for variant calling approaches that are specific for viruses, there are opportunities to develop robust methods and tools that strike a balance between detecting errors and true minority variants from field virus samples that present with different sequencing quality.

Data availability

The data analysis scripts and datasets used in analysis are available from our institutional Dataverse repository: http://dx.doi.org/10.7910/DVN/ZIO43M (Mohammed & Githinji, 2018)

Grant information

The work was funded by the Wellcome Trust Senior Investigator Award to Prof D. James Nokes [102975] in addition, this work was supported through the DELTAS Africa Initiative [DEL-15-003]. The DELTAS Africa Initiative is an independent funding scheme of the African Academy of Sciences (AAS)'s Alliance for Accelerating Excellence in Science in Africa (AESA) and supported by the New Partnership for Africa's Development Planning and Coordinating Agency (NEPAD Agency) with funding from the Wellcome Trust [107769] and the UK government. The views expressed in this publication are those of the author(s) and not necessarily those of AAS, NEPAD Agency, Wellcome Trust or the UK government.

The funders had no role in the study design, data generation and analysis, decision to publish, or preparation of the manuscript.

Supplementary material

Supplementary File 1: A description of data generation methods and command line tools used to create the artificial datasets.

Click here to access the data.

Supplementary File 2: FastQC metrics for sample used to generate simulated reads for the second dataset.

Click here to access the data.

Supplementary File 3: FastQC quality profile for sample used to generate simulated reads for the third dataset.

Click here to access the data.

Supplementary File 4: Tool-specific parameter settings for variant calling

Click here to access the data.

Supplementary File 5: Figures showing the performance of variant calling for minority variants that were inserted at different thresholds, standard (>25%), moderate (5-25%) and low (<5%).

Click here to access the data.

Supplementary Figure 1: Proportion of fully concordant positions with respect to sample coverage in dataset1. (A) Represents the proportion (y-axis) of fully concordant variants with respect to read coverage (x-axis) and (B) shows the proportion of variants positions that could not be identified by any minority caller at each level of sample coverage.

Click here to access the data.

Supplementary Figure 2: UpSetR plots showing the concordance between called variants and the respective variant callers using the first artificial dataset. The intersection size illustrates the number of variants in each intersection set. The horizontal axis shows the combination matrix identifying the intersections. A single filled circle represents a unique set of variants. Connected lines depict shared variants (intersections) among the variant callers. Intersection size was positively correlated with average coverage at (a) 20X, (b) 50X, (c) 100X, (d) 500X, (e) 1000X, (f) 2000X, (g) 5000X and (h) 10,000X.

Click here to access the data.

Supplementary Tables 1-4.

Supplementary Table 1: 166 artificially generated nucleotide mutations with frequencies from 0-1.

Supplementary Table 2: 156 artificially generated nucleotide mutations with frequencies below 0.5.

Supplementary Table 3: A breakdown of performance metrics of variant callers evaluated using the second dataset incorporated with an error profile derived from the set of reads used to assemble the reference genome. The samples represent simulated dataset of varying average depth of coverage. True positive (TP), true negative (TN), false positive (FP) and false negatives (FN) were used to calculate performance metrics of each caller. FPR – False positive rate.

Supplementary Table 4: A breakdown of performance metrics of variant callers evaluated using the third dataset generated with an error profile from a poorly sequenced sample. The samples represent simulated dataset of varying average depth of coverage. True positive (TP), true negative (TN), false positive (FP) and false negatives (FN) were used to calculate performance metrics of each caller. FPR – False positive rate.

Click here to access the data.

References

Domingo E, Sheldon J, Perales C: **Viral quasispecies evolution.** *Microbiol Mol Biol Rev.* 2012; **76**(2): 159–216.
PubMed Abstract | Publisher Full Text | Free Full Text

Eigen M, McCaskill J, Schuster P: **Molecular Quasi-Species.** *J Phys Chem.* 1988; **92**(24): 6881–6891.
Publisher Full Text

Ewing AD, Houlahan KE, Hu Y, *et al.*: **Combining tumor genome simulation with crowdsourcing to benchmark somatic single-nucleotide-variant detection.** *Nat Methods.* 2015; **12**(7): 623–630.
PubMed Abstract | Publisher Full Text | Free Full Text

Henn MR, Boutwell CL, Charlebois P, *et al.*: **Whole genome deep sequencing of HIV-1 reveals the impact of early minor variants upon immune recognition during acute infection.** *PLoS Pathog.* 2012; **8**(3): e1002529.
PubMed Abstract | Publisher Full Text | Free Full Text

Hofmann AL, Behr J, Singer J, *et al.*: **Detailed simulation of cancer exome sequencing data reveals differences and common limitations of variant callers.** *BMC Bioinformatics.* 2017; **18**(1): 8.
PubMed Abstract | Publisher Full Text | Free Full Text

Holland JJ, De La Torre JC, Steinhauer DA: **RNA virus populations as quasispecies.** *Curr Top Microbiol Immunol.* 1992; **176**: 1–20.
PubMed Abstract | Publisher Full Text

Huang HW, NISC Comparative Sequencing Program, Mullikin JC, *et al.*: **Evaluation of variant detection software for pooled next-generation sequence data.** *BMC Bioinformatics.* 2015; **16**: 235.
PubMed Abstract | Publisher Full Text | Free Full Text

Huang W, Li L, Myers JR, *et al.*: **ART: a next-generation sequencing read simulator.** *Bioinformatics.* 2012; **28**(4): 593–594.
PubMed Abstract | Publisher Full Text | Free Full Text

Hwang S, Kim E, Lee I, *et al.*: **Systematic comparison of variant calling pipelines using gold standard personal exome variants.** *Sci Rep.* 2015; **5**: 17875.
PubMed Abstract | Publisher Full Text | Free Full Text

Kinde I, Wu J, Papadopoulos N, *et al.*: **Detection and quantification of rare mutations with massively parallel sequencing.** *Proc Natl Acad Sci U S A.* 2011; **108**(23): 9530–9535.
PubMed Abstract | Publisher Full Text | Free Full Text

Koboldt DC, Chen K, Wylie T, *et al.*: **VarScan: variant detection in massively parallel sequencing of individual and pooled samples.** *Bioinformatics.* 2009; **25**(17): 2283–2285.
PubMed Abstract | Publisher Full Text | Free Full Text

Koboldt DC, Larson DE, Wilson RK: **Using VarScan 2 for Germline Variant Calling and Somatic Mutation Detection.** *Curr Protoc Bioinformatics.* 2013; **44**: 15.4.1–17.
PubMed Abstract | Publisher Full Text | Free Full Text

Koboldt DC, Zhang Q, Larson DE, *et al.*: **VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing.** *Genome Res.* 2012; **22**(3): 568–576.
PubMed Abstract | Publisher Full Text | Free Full Text

Krøigård AB, Thomassen M, Laenkholm AV, *et al.*: **Evaluation of Nine Somatic Variant Callers for Detection of Somatic Mutations in Exome and Targeted Deep Sequencing Data.** *PLoS One.* 2016; **11**(3): e0151664.
PubMed Abstract | Publisher Full Text | Free Full Text

Lai Z, Markovets A, Ahdesmaki M, *et al.*: **VarDict: a novel and versatile variant caller for next-generation sequencing in cancer research.** *Nucleic Acids Res.* 2016; **44**(11): e108.
PubMed Abstract | Publisher Full Text | Free Full Text

Lex A, Gehlenborg N, Strobelt H, *et al.*: **UpSet: Visualization of Intersecting Sets.** *IEEE Trans Vis Comput Graph.* 2014; **20**(12): 1983–1992.
PubMed Abstract | Publisher Full Text | Free Full Text

Macalalad AR, Zody MC, Charlebois P, *et al.*: **Highly sensitive and specific detection of rare variants in mixed viral populations from massively parallel sequence data.** *PLoS Comput Biol.* 2012; **8**(3): e1002417.
PubMed Abstract | Publisher Full Text | Free Full Text

McCrone JT, Lauring AS: **Measurements of Intrahost Viral Diversity Are**

Extremely Sensitive to Systematic Errors in Variant Calling. *J Virol.* 2016; **90**(15): 6884–6895.
PubMed Abstract | Publisher Full Text | Free Full Text

Mohammed KS, Githinji G: **Replication Data for: Evaluating the Performance of Tools Used to Call Minority Variants from Whole Genome Short-Read Data.** *Harvard Dataverse, V3.* 2018.
http://www.doi.org/10.7910/DVN/ZIO43M

Moya A, Holmes EC, González-Candelas F: **The population genetics and evolutionary epidemiology of RNA viruses.** *Nat Rev Microbiol.* 2004; **2**(4): 279–288.
PubMed Abstract | Publisher Full Text

Pabinger S, Dander A, Fischer M, *et al.*: **A survey of tools for variant analysis of next-generation genome sequencing data.** *Brief Bioinform.* 2014; **15**(2): 256–278.
PubMed Abstract | Publisher Full Text | Free Full Text

Poon LL, Song T, Rosenfeld R, *et al.*: **Quantifying influenza virus diversity and transmission in humans.** *Nat Genet.* 2016; **48**(2): 195–200.
PubMed Abstract | Publisher Full Text | Free Full Text

Sandmann S, de Graaf AO, Karimi M, *et al.*: **Evaluating Variant Calling Tools for Non-Matched Next-Generation Sequencing Data.** *Sci Rep.* 2017; **7**: 43169.
PubMed Abstract | Publisher Full Text | Free Full Text

Spencer DH, Tyagi M, Vallania F, *et al.*: **Performance of common analysis methods for detecting low-frequency single nucleotide variants in targeted next-generation sequence data.** *J Mol Diagn.* 2014; **16**(1): 75–88.
PubMed Abstract | Publisher Full Text | Free Full Text

Stack JC, Murcia PR, Grenfell BT, *et al.*: **Inferring the inter-host transmission of influenza A virus using patterns of intra-host genetic variation.** *Proc Biol Sci.* 2013; **280**(1750): 20122173.
PubMed Abstract | Publisher Full Text | Free Full Text

Vignuzzi M, Stone JK, Arnold JJ, *et al.*: **Quasispecies diversity determines pathogenesis through cooperative interactions in a viral population.** *Nature.* 2006; **439**(7074): 344–348.
PubMed Abstract | Publisher Full Text | Free Full Text

Watson SJ, Welkers MR, Depledge DP, *et al.*: **Viral population analysis and minority-variant detection using short read next-generation sequencing.** *Philos Trans R Soc Lond B Biol Sci.* 2013; **368**(1614): 20120205.
PubMed Abstract | Publisher Full Text | Free Full Text

Wilm A, Aw PP, Bertrand D, *et al.*: **LoFreq: a sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets.** *Nucleic Acids Res.* 2012; **40**(22): 11189–11201.
PubMed Abstract | Publisher Full Text | Free Full Text

Worby CJ, Lipsitch M, Hanage WP: **Shared Genomic Variants: Identification of Transmission Routes Using Pathogen Deep-Sequence Data.** *Am J Epidemiol.* 2017; **186**(10): 1209–1216.
PubMed Abstract | Publisher Full Text | Free Full Text

Yang X, Charlebois P, Macalalad A, *et al.*: **V-Phaser 2: variant inference for viral populations.** *BMC Genomics.* 2013; **14**: 674.
PubMed Abstract | Publisher Full Text | Free Full Text

Zhang F, Flaherty P: **Variational inference for rare variant detection in deep, heterogeneous next-generation sequencing data.** *BMC Bioinformatics.* 2017; **18**(1): 45.
PubMed Abstract | Publisher Full Text | Free Full Text

Zukurov JP, do Nascimento-Brito S, Volpini AC, *et al.*: **Estimation of genetic diversity in viral populations from next generation sequencing data with extremely deep coverage.** *Algorithms Mol Biol.* 2016; **11**: 2.
PubMed Abstract | Publisher Full Text | Free Full Text

Open Peer Review

Current Referee Status: ✔ ✔

Version 2

Referee Report 13 November 2018

https://doi.org/10.21956/wellcomeopenres.16071.r33893

✔ **Brad A. Chapman**
Department of Biostatistics, Harvard Public School of Health, Boston, MA, USA

Thank you for the detailed response and revision. The changes you've made greatly improve the paper and I especially appreciate the inclusion of run parameters for the tools and additional stratification plots.

These revisions address all the previous suggestions, and my only additional comment from seeing Figure 5 and the run parameters was that most of these tools are producing a high percentage of low frequency noise which will result in poor precision when trying to identify variants in real samples. A useful followup would be explore filtering or ensemble methods to help improve this specificity.

Thank you again for the work on this paper.

Competing Interests: No competing interests were disclosed.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Referee Report 02 October 2018

https://doi.org/10.21956/wellcomeopenres.16071.r33892

✔ **Micha M. Bayer**
Information & Computational Sciences (ICS) group, James Hutton Institute, Dundee, UK

Thank you for addressing the issues raised in my review of Version 1. Everything has been resolved and I have no further comments to add.

Competing Interests: No competing interests were disclosed.

Referee Expertise: Bioinformatics, variomics, genomics, crop plants

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Version 1

Referee Report 13 April 2018

https://doi.org/10.21956/wellcomeopenres.14703.r32418

Micha M. Bayer
Information & Computational Sciences (ICS) group, James Hutton Institute, Dundee, UK

This is a comparative review of variant calling tools with focus on the detection of minority variant calls, specifically with reference to virus data. The authors have provided data on four freely available tools that can be tuned to call minority variants, or are specifically designed for this purpose. The data used here was artificial reads generated with the ART-Illumina tool, which were spiked in with variants at known locations. This was replicated eight times at levels of read coverage spanning four orders of magnitude. The outcome of the experiment was a low level of concordance between tools. None of the tools were able to identify all of the spiked-in variants.

The approach taken here is technically sound, as it is based on a strictly defined truth set, although I am slightly uncomfortable with the lack of real data (but equally I am aware of the methodological problems real data presents in this kind of scenario). The study is useful for the wider community as a) it reiterates the point that using a single tool for a given bioinformatics task comes with the risk of missing information and b) it provides concrete pointers as to which tools perform well in this scenario.

My main reservation is that there is no mention of the importance of parameter settings. Different parameter values can affect variant calling outcomes dramatically, even when just comparing different runs of the same tool. FreeBayes alone has over 70 command line parameters, many of these continuous variables. This makes for an incredibly large parameter space, and using different sets of parameter values can potentially have much more of an impact on the results than the choice of tool. This should at least be mentioned in the Discussion section.

There also doesn't seem to be any data on how the individual tools were parameterised in this study. Were the defaults used in each case? Presumably not, given that we are looking for minority variants (and tools like FreeBayes appear to be developed and parameterised with a human germline use case in mind). This information is critical for reproducibility and must be included (even if it is just a statement saying the defaults were used).

My other comments below reference the online PDF version of the article.

Abstract
- Results section: change "FreeBayes identified majority" to "FreeBayes identified the majority"
- Conclusions section: change "impact" to "impacts"

Introduction
- para 1 line 7: is the "and" redundant?
- para 5 line 7: comma after "callers"

Methods
- para 1 line 1: "we considered eleven published,open-source tools" I counted six excluded tools in Table 1, plus four that have been taken through to evaluation = 10

- para 2 (“Artificial datasets”): The concept of the error profile has not been well explained. This section needs to be expanded. I would like to see some description of what the error profile consists of, how it has been derived from the FASTQC data, and how it has been applied to the reads in practice.
- para 3 line 2:“at varying frequencies” – what were they? I presume we are talking about the proportion of reads at a spiked-in variant site that carried the alternate allele at this site. This information strikes me as absolutely critical for this paper – this is about minority variants after all. Please provide this data.
- para 4 line 8: “written as a comma separated (CSV) files” should be “written as a comma separated (CSV) file”

Results
- para 3 line 1: “FreeBayes identified majority” should be “FreeBayes identified the majority”
- para 3 line 7: “it missed majority” should be “it missed the majority”
- General comment: Vardict should get a mention in this section, as it actually had better recall rates than FreeBayes in 6 out of 8 samples.

Discussion
- para 1 line 5: change “ascertain” to “distinguish”
- para 3 line 1: “Sensitivity for majority” should be “Sensitivity for the majority”

Table 3
- As far as I can tell, the counts for FN + TP should always add up to 166. This is not the case for the majority of the Freebayes runs (except for Freebayes with samples 1 and 3, which are ok).

Supplementary File S1
- To ensure reproducibility, please include the BAMSurgeon command line statement for spiking in the variants.

Supplementary Figure 2
- It is not apparent what the individual figures/pages in the PDF file represent. Please label these appropriately.

Is the work clearly and accurately presented and does it cite the current literature?
Yes

Is the study design appropriate and is the work technically sound?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Partly

If applicable, is the statistical analysis and its interpretation appropriate?
Yes

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?

Yes

Competing Interests: No competing interests were disclosed.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 13 Aug 2018

George Githinji, KEMRI-Wellcome Trust Research Programme, Kenya

We are grateful for taking time to review this work and for providing useful comments that we have considered in this new draft of the manuscript.

- We agree on the challenges of obtaining and using the appropriate actual data when conducting these types of analysis. We have done our best to generate reliable artificial datasets based on an actual RSV reference genome and error profiles from two sequenced samples using the methods described in this manuscript.
- In regard to the importance of parameter settings and lack of data on how individual tools were parameterized, we have explicitly stated the parameter settings in the variant calling section and in the Supplementary File 4. In the discussion section, we highlight the challenge of exploring all possible parameter settings.
- The concept of the error profile has is now well explained in paragraph 1 of the artificial datasets section in the current manuscript.
- The actual frequencies used to spike the variants are now described and provided in Supplementary file 5.
- There was a bug in the script used to calculate the FN and TP counts in FreeBayes runs. This has been corrected and updated to reflect the true FN and TP counts. A revised table and subsequent figures have been provided.

Competing Interests: No competing interests were disclosed.

Referee Report 22 March 2018

https://doi.org/10.21956/wellcomeopenres.14703.r31459

Brad A. Chapman
Department of Biostatistics, Harvard Public School of Health, Boston, MA, USA

The authors validate methods to call low frequency minority variants from viral samples. Detection of minority species in viral population sequencing helps improve treatments by matching drugs to the full spectrum of sub-species, and allows tracking of sub-species within outbreaks.

The paper uses synthetic datasets generated by the ART read simulator (https://www.niehs.nih.gov/research/resources/software/biostatistics/art/) with randomly generated variants. While this won't fully describe the complexity of a real viral mixture, it does provide a solid baseline for assessing callers. The authors used mutation error profiles from good and bad sequencing

samples to model different error rates, and evaluated the impact of differing coverage on sensitivity and specificity.

My main suggestion is to stratify comparisons by variant frequency to assess how callers do with different low frequency events, and I provide more specific comments below.

Methods suggestions

- The stratifications by coverage, error rates and callers are great for identifying the effect of different callers on detection. The one missing component is the allele frequency of the generated variants. The paper does not define the range of frequencies generated, and this will have a large impact on the ability of detection across callers. Looking through the list of generated variants (Githinji_2018_variant.list.tab) the range varies from high frequency (90%+) to very low frequency (<1%). Since the focus is on low frequency detection, I'd suggest stratifying the results into bins: standard (>25%), moderately low frequency (5-25%) and low frequency (<5%) and examine the caller and depth detection within these bins. Pending the results of this, the authors may want to generate additional low frequency variants to help differentiate caller methods.

- A confusing aspect of evaluating the depth metrics is that the number of possible true positives differs between sample depths. In Table 3, the FreeBayes TP + FNs are 166, 165 and 166 for 20x, 50x and 100x. However they change to 61, 55 and 55 for 500x, 1000x and 200x. Other callers seem to be consistent. Why does FreeBayes have different total variant numbers across depth?

Paper suggestions

- Table 1 should get compressed into a single paragraph saying that HaploptypeCaller, Platypus and mpileup target germline calling and not low frequency detection. It doesn't need to be a separate table.

- Table 2 should be replaced into a link to descriptions of true/false positives/negatives and does not need to be a separate figure in the paper.

- The title should reflect that this paper describes viral calling (versus, say low frequency somatic variant detection).

Is the work clearly and accurately presented and does it cite the current literature?
Yes

Is the study design appropriate and is the work technically sound?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Not applicable

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?

Yes

Competing Interests: No competing interests were disclosed.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 13 Aug 2018

George Githinji, KEMRI-Wellcome Trust Research Programme, Kenya

We are very grateful for taking time to review this work and for providing useful comments which we have considered in the new version of the manuscript:

- All the paper suggestions have been considered table 1 and table 2 were removed and replaced with text descriptions.
- We have added a new figure comparing the allele frequency of the spiked and called variants. We highlight the observations in the result and discussion sections. In addition, we provided a stratified result of the variants frequency and discuss how the callers perform for different low frequency events. We also provided an addition set of spiked variants at frequency below 50%.
- The results section in table 3 have been revised to reflect the correct number of TP and FN for FreeBayes. Figure was also revised to reflect the new changes.

Competing Interests: No competing interests were disclosed.