RESEARCH ARTICLE

UPDATED Considerations for clinical read alignment and mutational profiling using next-generation sequencing [v2; ref status: indexed, http://f1000r.es/NMpsFc]

Gavin R Oliver

ALMAC, Craigavon, Co. Armagh, UK

v2 **First published:** 16 Jul 2012, **1**:2 (doi: 10.12688/f1000research.1-2.v1)
Latest published: 20 Sep 2012, **1**:2 (doi: 10.12688/f1000research.1-2.v2)

Abstract

Next-generation sequencing technologies are increasingly being applied in clinical settings, however the data are characterized by a range of platform-specific artifacts making downstream analysis problematic and error-prone. One major application of NGS is in the profiling of clinically relevant mutations whereby sequences are aligned to a reference genome and potential mutations assessed and scored. Accurate sequence alignment is pivotal in reliable assessment of potential mutations however selection of appropriate alignment tools is a non-trivial task complicated by the availability of multiple solutions each with its own performance characteristics. Using targeted analysis of BRCA1 as an example, we have simulated and mutated a test dataset based on Illumina sequencing technology. Our findings reveal key differences in the abilities of a range of common commercial and open source alignment tools to facilitate accurate downstream detection of a range of mutations. These observations will be of importance to anyone using NGS to profile mutations in clinical or basic research.

Article Status Summary

Referee Responses

Referees	1	2	3
v1 published 16 Jul 2012	☑	☑	☒ report
v2 published 20 Sep 2012 UPDATED			☒ report 1

1 **Thomas Friedman**, National Institute on Deafness and Other Communication Disorders (NIDCD), National Institutes of Health USA

2 **Vera Kalscheuer**, Max Planck Institute for Molecular Genetics Germany

3 **Mihaela Pertea**, Johns Hopkins University School of Medicine USA, **Steven Salzberg**, Johns Hopkins University School of Medicine USA

Latest Comments

No Comments Yet

Corresponding author: Gavin R Oliver (goliver@almacgroup.com)

How to cite this article: Oliver GR (2012) **Considerations for clinical read alignment and mutational profiling using next-generation sequencing [v2; ref status: indexed, http://f1000r.es/NMpsFc]** *F1000Research* 2012, **1**:2 (doi: 10.12688/f1000research.1-2.v2)

Grant information: The author(s) declared that no grants were involved in supporting this work.

Competing interests: No competing interests were disclosed.

First published: 16 Jul 2012, **1**:2 (doi: 10.12688/f1000research.1-2.v1)
First indexed: 26 Jul 2012, **1**:2 (doi: 10.12688/f1000research.1-2.v1)

UPDATED **Changes from Version 1**

In response to the reviewers' comments I have made numerous changes and provided clarifications that I believe address many of the criticisms. My work has been limited to some degree by the fact that I have been in the process of switching jobs and moving from the UK to the US. Corporate policy dictated that much of my data could not be taken with me. Nonetheless I feel many of the concerns have been addressed. The exact commands used to run each aligner are now included. These were mistakenly excluded from the first version. Language surrounding FDR, sensitivity etc has been clarified to reflect the fact that they are not direct measures of aligner metrics. Language used in describing run-times has been altered to avoid any ambiguity and mention of the potential effects of larger datasets included. Language used to describe the classification of aligners has been corrected and description of the effect of Indel size on detection clarified. All minor concerns have also been addressed.

See referee reports

Introduction

Since emergence in 2005, next-generation sequencing (NGS) technologies have proven prolific tools in the research setting, permeating a variety of scientific disciplines and demonstrating a range of applications that seems to be limited only by the imagination of the sequencing community. The technology continues to develop at a rapid pace with established instrument manufacturers regularly augmenting their product portfolios and an increasing number of start-up companies promising to disrupt the market. Beyond basic research applications, NGS technologies are now increasingly being applied in the clinical environment, driven partly by their rapid maturation and the arrival to market of smaller, cheaper sequencing platforms.

The potential clinical application of NGS has a broad scope ranging from full human genome profiling[1] to investigation of the microbiome[2] and includes applications such as biomarker discovery, patient diagnosis, prediction of drug response and patient stratification for clinical trials. Such applications often involve the targeted profiling of genes known to be of clinical relevance. These genes harbor diagnostically relevant variants including single nucleotide polymorphisms (SNPs), and small insertions and deletions (INDELs). Individual genes have previously been interrogated in clinical testing using traditional techniques such as Sanger sequencing however NGS technologies have already begun to supplant the previous tools of choice in these areas, offering increased speed and throughput with reduced running costs.

Despite many successes and increasing uptake, the data generated by NGS analyzers is not perfect, with each platform yielding characteristic errors and biases. Furthermore, NGS technologies produce reads that are much shorter than those traditionally produced by Sanger sequencing methods and this can complicate matters further, especially in genomes containing a large proportion of repetitive elements[3]. The effect of these problems is most visible in large scale studies such as genome-wide sequencing where a recent study reported a 1 million variant platform-based discrepancy for a single genome[4]. This fact bestows responsibility on both algorithm and software developers and downstream users to develop deep understanding of the various data types and their idiosyncrasies and to apply this appreciation in their analysis and interpretation, in order to correct or compensate for potential errors. Despite forming an area of active research, data interpretation remains an issue and is no doubt a factor in feeding the inertia of many clinical facilities that are reluctant to adopt the new technologies[5].

Two major computational steps in variant detection from NGS data are read alignment whereby the data are mapped to corresponding locations on a target genome, and mutation calling whereby nucleotides differing from the target genome are assessed and scored on their likelihood of representing a genuine mutation versus an error. While these two stages of analysis may be supplemented with various pre or post processing techniques they represent the most crucial steps and therefore the area of most active software and algorithm development.

Aligners of choice have begun to emerge[6,7] however their strengths are often application specific and different tools are recommended depending on the sequencing platform and individual study goals. Often there is a trade-off between speed and sensitivity at the read alignment stage with speed sometimes prioritized due to the volumes of data produced by NGS technologies and the corresponding time required for analysis. Aligners can be classified as gapped or ungapped based on their ability to produce successful alignments in the presence of small INDELs. Aligners including BWA[8] and Stampy have been shown to produce alignments with a fair degree of success for reads containing a range of INDEL sizes[9] however such abilities will vary based on a range of complicating factors including size and location of the INDEL. As well as generating continuous controversy[10], the *BRCA1* gene presents a particularly interesting set of alignment challenges due to a disproportionately high concentration of INDELs greater than eight nucleotides in length. In fact, 3% of known deleterious mutations in the Breast Cancer Information Core (BIC) database[11] fall into this category and represent a significant over-representation when compared to a healthy genome. Furthermore the *BRCA1* gene contains multiple areas of high shared identity in the form of tandem repeats, posing another difficulty in achieving accurate read mapping. Challenges like this pose particular difficulties in the clinical setting where errors have the potential to translate to misdiagnosis or mistreatment, directly affecting and endangering the lives of patients. No gold-standard clinical alignment tools yet exist and numerous publicized examples of early translational work appear to base their choice of tool on user-friendliness or availability of a graphical-user-interface rather than assessments of performance. We have investigated the performance of a range of popular alignment tools and assessed their ability to facilitate accurate downstream detection of known mutations with a commonly used variant calling pipeline. Several of these tools are already being used as components of diagnostic workflows in the clinical setting. Here we present data generated using *BRCA1* reads created *in-silico* in a simulated, targeted sequencing scenario. Our findings demonstrate the widely varying abilities of common read alignment tools and their impact on downstream variant calling. Furthermore the results suggest a need for careful and thorough evaluation of the tools used in a particular analysis pipeline by simulation and analysis of data of known constitution.

Methods

Aligners

A range of open source and commercial alignment tools were selected for assessment based on their reported ability to facilitate detection of both SNPs and INDELs as well as frequency of citation in both scientific and commercial literature. The aligners included in the comparison were: BWA (0.5.9-r16), Bfast (0.7.0a)[12], Smalt (0.5.8), Stampy (1.0.14), Mosaik (2.1.33), CLC Genomics Workbench's (5.0.1) NGS and Beta aligner, Novocraft's Novoalign (V2.07.18)[13], Omixon's Variant Toolkit (2.1.3), Bowtie 2 (2.0.0-beta5)[14] and Softgenetics Nextgene (2.2) aligner.

Read simulation

Stampy was used to simulate sixty-seven groups of 200,000 90bp paired-end FASTQ Illumina reads from the human *BRCA1* gene (hg19) with an appropriate error profile. Each sequence grouping was mutated *in-silico* with custom scripts used to introduce a combination of 20 SNPs and 13 INDELs from a test set of 2211 (1299 unique) known *BRCA1* variants containing 1340 SNPs, 320 insertions and 551 deletions. The test set was randomly selected from the full collection of *BRCA1* mutations in dbSNP v131[15] and overlapping mutations removed.

Dataset 1: Simulated Illumina BRCA1 reads in FASTQ format

134 Data Files

http://dx.doi.org/10.6084/m9.figshare.92338

Dataset 2: VCF files describing the known mutations in each read file

67 Data Files

http://dx.doi.org/10.6084/m9.figshare.92401

Read alignment

Reads were aligned to hg19 chromosome 17 on a HP DL585 G6 server with 4 six-core AMD Opteron 2.8Ghz processors and 256GB of RAM. Multi-threading with the maximum number of threads supported by the aligner was utilized. The commands used to run the aligners are provided for all aligners with the exception of the CLC and NextGene software which are GUI based and were run with default settings. Each aligner was run in both single-end and paired-end mode with half of the paired-end reads being used to simulate a single-end read dataset. Run-times were recorded based on the wall-clock time taken to align and produce a SAM format output for all 67 sets of FASTQ reads (i.e. 13.4 million reads). Index creation was not included as this is a one-off step for any reference.

Scripts 1: Aligner execution scripts

16 Bash Scripts

http://dx.doi.org/10.6084/m9.figshare.94257

Calling of SNPs and INDELs

Each SAM formatted file was converted to BAM format and processed to ensure downstream compatibility with GATK[16] using a combination of tools from the Picard collection. (SamFormatConverter, AddOrReplaceReadGroups, SortSam and BuildBamIndex respectively). BAM files were then processed in a GATK-based pipeline. The pipeline consisted of local realignment around INDELs (RealignerTargetCreator and IndelRealigner), quality score recalibration (CountCovariates and TableRecalibrator) and finally variant calling (UnifiedGenotyper). The wrapper scripts sam2bam.sh and gatk.sh are provided and can be used to recreate the processing steps from alignment files (SAM format) to variant call (VCF format).

Scripts 2: Alignment processing and variant calling scripts

2 Bash Scripts

http://dx.doi.org/10.6084/m9.figshare.94258

Dataset 3: Aligner-specific VCF files containing the mutations called for each set of reads in single and paired-end modes

200 Data Files

http://dx.doi.org/10.6084/m9.figshare.92404

Results/discussion

Mutation panel

The reads and mutation panel utilized here represent a challenging and multi-functional test set with widely varying INDEL sizes (Table 1) and an extensive range of SNPs providing a useful means of assessing the aligners' effects on variant calling in single and paired-end read modes. The reads were created to contain only known mutations from the human *BRCA1* gene thus facilitating downstream assessment of mutation profiling accuracy whilst remaining comparable to real-world data. Reads were simulated to closely match the error profile of Illumina's sequencing technology enabling a further level of realism to be captured in the simulated test-set. Only homozygous variants at high levels of sequence coverage (70–140x) were included in the test set to focus testing of the alignment tools' abilities rather than the quality of the downstream variant calling methods.

Run-times

Run-times varied widely from seconds to hours Table 2. Novocraft's Novoalign software performed fastest on this dataset and was closely followed by BWA and Bowtie 2 in both single and paired-end mode whilst Bfast's paired-end mode represented the slowest run-time by almost half a day. Nextgene was excluded from this comparison due to the fact it is Windows-based software and it was not possible to assess run-times on the same hardware as for the other aligners. For all aligners studied here, alignment time could arguably be considered acceptable for the application simulated i.e. alignment of targeted sequencing reads for 64 samples. However, it should be noted that a typical exome data set can be around around 7 times the 13.4 million reads used here and full genome datasets are larger still. Thus, some of the aligners tested in this targeted sequencing application would likely prove unsuitable for larger scale applications. What constitutes an acceptable running time will vary and should be considered on a per-application basis in association with other performance characteristics. It should also be noted that the ranking of aligners by speed observed here could vary based on the size of the input dataset.

Table 1. Range of INDEL sizes (in base pairs) in the *BRCA1* mutation panel. A representative range of SNPs and small INDELs corresponding to known mutations were distributed between 67 groups of simulated Illumina reads in order to test the aligners' abilities to facilitate accurate downstream detection of mutations.

Mutation size	# Insertions	# Deletions
1	215	261
2	52	121
3	8	40
4	11	60
5	5	27
6	11	6
7	4	7
8	3	10
9	0	2
11	0	9
12	1	0
13	0	1
14	3	0
15	1	1
16	1	0
17	0	2
18	1	0
19	0	1
24	4	0
29	0	1
62	0	1
64	0	1

Table 2. Aligner run times in wall-clock hours. Each aligner was tested by aligning all 67 groups of simulated Illumina reads in single and paired-end mode. The maximum number of threads utilizable by each aligner were used in testing.

Aligner	# Threads Utilizable (of 24)	PE Time to align	SE Time to align
BWA	24	0.28	0.08
Bfast	24	19.79	7.01
Smalt	8	6.44	8.25
Stampy	1	6.28	3.50
Mosaik	24	2.64	1.56
CLC	24	1.41	0.88
CLCBeta	24	0.67	0.60
Novoalign	24	0.11	0.04
Omixon	24	4.83	1.26
Bowtie2	24	0.30	0.11
Nextgene	24	NA	NA

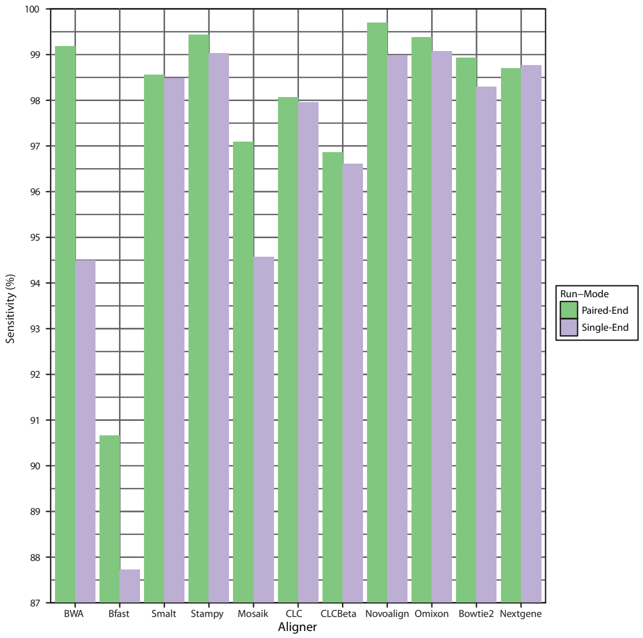


Figure 1. Overall sensitivity of downstream mutation detection. Graphical representation of the total percentage of mutations detected across of 67 groups of simulated Illumina reads following alignment in single-end and paired-end mode with each aligner.

Table 3. Overall detection sensitivities. The total percentage of mutations detected post-alignment across the 67 groups of simulated Illumina reads was recorded for single and paired-end run-modes. Each read group contained 13 INDELs of varying sizes as well as 20 SNPs.

Aligner	Sensitivity PE	Sensitivity SE
BWA	99.23	94.53
Bfast	90.68	87.83
Smalt	98.55	98.51
Stampy	99.46	99.05
Mosaik	97.06	95.07
CLC	98.01	97.92
CLCBeta	96.88	96.74
Novoalign	99.59	99.00
Omixon	99.41	99.14
Bowtie2	98.91	98.19
Nextgene	98.64	98.69

Sensitivity of detection

The greatest overall sensitivities of detection were achieved following alignment with Novoalign in paired-end mode and Omixon's Variant Toolkit in single-end mode (Figure 1, Table 3). Stampy also enabled highly sensitive mutation detection with Bfast performing least favorably. Sensitivities were also assessed based on the category of mutation (Table 4). Some aligners such as Smalt, Bowtie 2 and CLC were clearly seen to facilitate better detection of SNPs than INDELs in general. Nextgene performed similarly for SNPs and deletions but gave way to lower downstream sensitivities of insertion detection. BWA showed obvious decreases in ability to enable accurate detection of INDELs when moving from paired-end to single-end mode. In contrast, Novoalign, Omixon and Stampy performed well regardless of run-mode or mutation type. Overall Novoalign was the best performer in paired-end mode while Omixon achieved the highest downstream detection sensitivities in

Table 4. Downstream detection sensitivities by mutation type. Total percentage of SNPs and INDELs detected across the 67 groups of simulated Illumina reads following alignment with each aligner in single and paired-end mode.

Aligner	% SNPs found PE	% SNPs found SE	% Insertions found PE	% Insertions found SE	% Deletions found PE	% Deletions found SE
BWA	99.48	99.48	99.38	89.38	98.55	85.48
Bfast	92.69	89.85	83.13	78.75	90.20	88.20
Smalt	99.40	99.40	96.56	96.25	97.64	97.64
Stampy	99.48	99.25	99.38	98.13	99.46	99.09
Mosaik	97.01	96.34	96.25	90.00	97.64	94.92
CLC	99.40	99.40	96.56	96.25	95.46	95.28
CLCBeta	97.61	97.61	95.31	95.00	96.01	95.64
Novoalign	99.70	99.48	100.00	98.13	99.09	98.37
Omixon	99.40	98.96	99.38	99.38	99.46	99.46
Bowtie2	99.50	99.50	97.50	96.60	98.40	96.00
Nextgene	99.00	99.00	96.60	96.60	99.10	99.10

single-end mode. In a clinical environment, sensitivity will likely represent the most important metric in evaluating alignment software, however other factors may also be of importance, depending on the test in question. Specificity values are not included here as they were non-discriminatory in this context due to low numbers of false positives relative to the high number of true negatives.

Incorrect identification of mutations

Controlling the rate of false positive results is clinically important in avoiding unnecessary treatment, expense and patient anxiety. The number of incorrectly identified mutations varied widely dependent on aligner and run-mode (Figure 2, Table 5). The highest downstream false-positive rates were obtained after alignment with Bfast, followed by CLC Bio's Beta aligner, Nextgene, Mosaik, Stampy, and Bowtie 2. Novoalign and Smalt had the lowest downstream false-positive rates followed closely by Omixon and CLC. Number of false positives alone is of limited utility in assessing aligner performance, however positive predictive value (PPV) provides a useful metric which combines counts of both true and false positives in a single value (Table 6). PPV was calculated based on the equation:

$$PPV = \frac{True\ Positives}{True\ Positives + False\ Positives}$$

No inferences were made about prevalence in calculating the values. Novoalign achieved the highest downstream PPV in both single and paired-end mode. Smalt, CLC and Omixon's Variant Toolkit also performed strongly on this metric. Notably Stampy performed relatively poorly in contrast to the high downstream detection sensitivity it achieved.

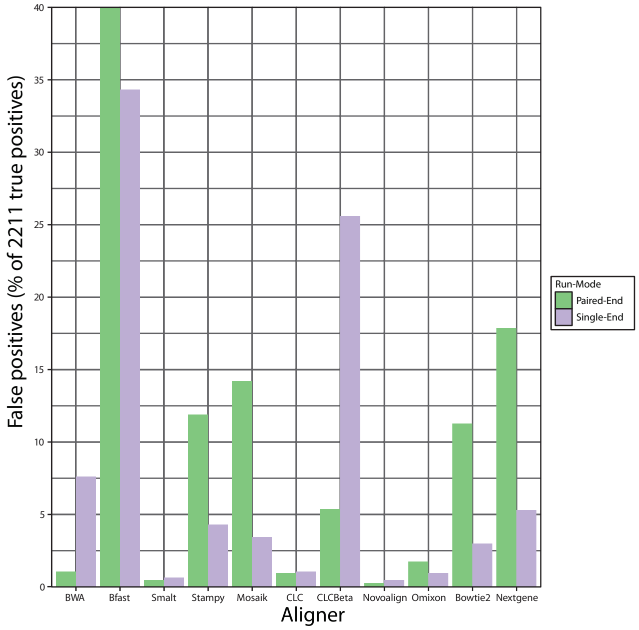


Figure 2. Graphical representation of the total number of downstream false positives expressed as a percentage of the true positive mutations detected following alignment with each each aligner in single and paired-end mode across 67 groups of simulated Illumina reads. Each read group contained 20 SNPs and 13 INDELs.

Paired-end vs. single-end reads

Notably Bfast, Nextgene, Stampy, Mosaik and Bowtie 2 all showed obvious increases in the number of downstream false positives detected following paired-end alignment vs single-end alignment. Conversely, switching from paired-end to single-end reads had varying adverse effects on the downstream detection sensitivities for all aligners except for Nextgene. The worst affected was BWA which retained all downstream SNP calls but lost 13% of INDEL calls. Notably Nextgene and Omixon's Variant Toolkit were the only software to retain all INDEL calls when switching from paired to single-end mode. Omixon's Variant toolkit achieved the highest sensitivity of dowsntream SNP and INDEL detection in single end mode with a 99.14% overall detection sensitivity.

This strong performance in single-end mode is relevant not only from a diagnostic standpoint, but also from a clinical cost-saving perspective as paired-end protocols ultimately incur extra costs per run vs single-end protocols. While paired-end reads generally

Table 5. Raw numbers of false positive and negative mutations detected across 67 groups of simulated Illumina reads following alignment with each aligner in single and paired-end mode. Each read group contained 20 SNPs and 13 INDELs.

Aligner	# False positives PE	# False positives SE	# False negatives PE	# False negatives SE
BWA	23	168	17	121
Bfast	884	759	206	269
Smalt	10	14	32	33
Stampy	263	95	12	21
Mosaik	314	76	65	109
CLC	21	23	44	46
CLCBeta	119	566	69	72
Novoalign	6	10	9	22
Omixon	39	21	13	19
Bowtie2	249	66	24	40
Nextgene	395	117	30	29

Table 6. Positive predictive values (PPVs) calculated for each aligner in single and paired-end mode based upon detection of mutations across 67 groups of simulated Illumina reads each containing 20 SNPs and 13 INDELs. In this case PPV is the total number of true positives divided by the sum of the total number of true positives and true negatives.

Aligner	PPV PE	PPV SE
BWA	98.96	92.56
Bfast	69.40	71.90
Smalt	99.54	99.36
Stampy	89.32	95.84
Mosaik	87.24	96.51
CLC	99.04	98.95
CLCBeta	94.74	79.08
Novoalign	99.73	99.55
Omixon	98.26	99.05
Bowtie2	89.78	97.05
Nextgene	84.67	94.91

represent a saving in terms of cost per megabase, they effectively double sequencing output and this may not be a cost-effective option depending on the logistics of the individual run. Furthermore, researchers who outsource sequencing will often see the available protocol for their relatively small sequencing project dictated by the larger projects they are multiplexed alongside.

Effect of INDEL size on detection

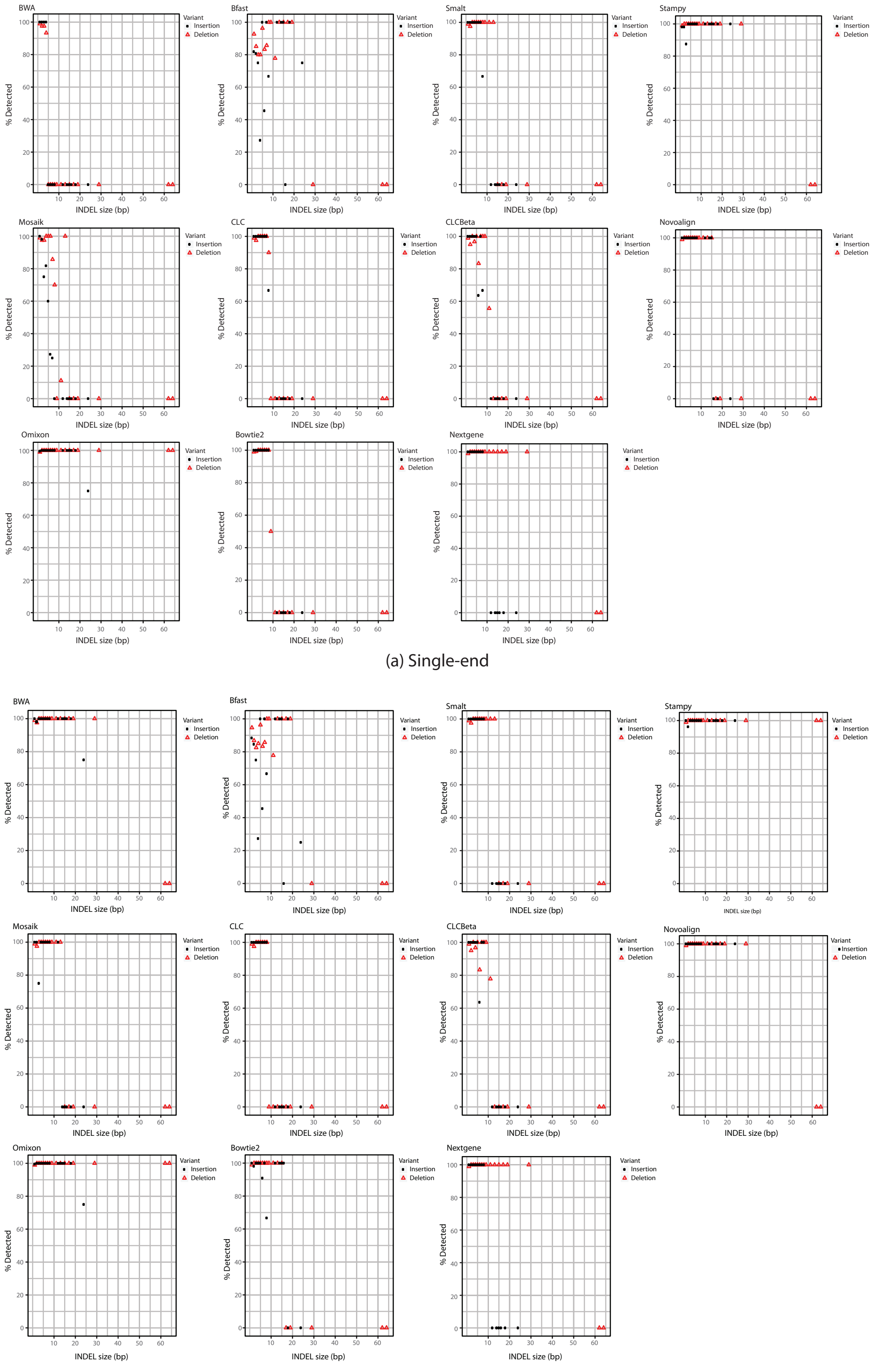
Ability to facilitate downstream calling of INDELs varied by aligner with INDEL sizes influencing detectability in most cases (Figure 3). The effect of size varied by aligner and run-mode with BWA and Novoalign enabling good downstream detection rates for all but the two largest deletions in paired-end mode while others such as Bowtie 2 and the CLC aligners didn't achieve downstream detection far beyond a 10bp INDEL size. Stampy and Omixon eanbled good downstream detectability across the range of mutation sizes in the panel. Run-mode clearly affected detection of INDELs in some instances. For example BWA saw a marked decrease in downstream INDEL calling for single-end mode compared to paired-end while the Omixon results appeared unaffected by switching run-modes. Downstream detection varied between insertions and deletions also. For example, Nextgene enabled higher downstream sensitivity for insertions than deletions. Stampy and Omixon's Variant Toolkit were the only aligners that enabled detection of the two largest deletions in the mutation panel. Collectively these observations highlight a need for those involved with testing and analysis to develop an appreciation of the various mutations that might exist in their target genes and to select their analysis tools appropriately. INDELs in the range represented by the *BRCA1* mutation panel have real-world relevance in genetic disorders and strong aligner performance on larger INDELs appears to be an exception rather than a rule.

Summary and conclusions

Using a simulated, targeted sequencing scenario with Illumina read data, the work presented here highlights several important considerations regarding aligner choice in studies involving profiling of mutations. Furthermore the data presented goes some way to characterizing the performance of a comprehensive selection of commonly used aligners and should represent a useful resource for anyone focused on similar scientific studies.

Whilst the dataset used in this study was engineered to include a challenging range of mutations and efforts were made to simulate the error profile of Illumina sequencing technology, it is nevertheless a simplified representation of real-world data. Experimental artifacts such as PCR stutter have the potential to present further challenges to alignment algorithms and there is no consideration of such issues here. Furthermore, the test dataset used in this study produced a uniform, high coverage of a single target gene and only homozygous variants were simulated. Finally, the use of only a single variant-caller in the study means that some of the errors encountered may not be due to alignment issues. The aim is to follow up the current study by focusing on an expanded gene-set, alternative variant-callers, homo and heterozygous mutations and different sequence formats. Nonetheless, this focused study demonstrates the utility of simulated data in assessing program performance.

With the exception of Bfast, all aligners performed relatively well on the *BRCA1* dataset. Clinical applications necessitate the use of the most highly accurate solutions, however. Only Novoalign, Omixon's Variant Toolkit and Stampy enabled 99% or greater sensitivity in both paired-end and single-end modes. Omixon and Stampy were the only two aligners to facilitate detection of the longest deletions in the dataset however Stampy's performance was let down by a downstream false positive rate which would be considered unacceptably high for many applications. While Novoalign did not enable detection of the largest deletions in the test dataset, it was the most sensitive in paired-end mode and performed fastest on this dataset. Nevertheless, assuming the longer run-times are not an issue, Omixon's superior sensitivity in single-end read mode likely makes it the best option when paired-end protocols are not possible. While the tests here produce some clear winners, they also serve to highlight that program performance can vary widely



(a) Single-end

(b) Paired-end

Figure 3. Effects of INDEL size on success of detection for all aligners in single-end (top panel) and paired-end (bottom panel) mode across 67 groups of simulated Illumina reads, each containing 13 INDELs of varying sizes.

based on the fine-details of a particular run. Even the strongest overall performer can be found lacking in some respects and this means that researchers should be vigilant in their selection of tools for a particular application. In certain instances it may even be necessary to combine two or more approaches to ensure that all relevant aspects of a given dataset are sufficiently characterized and any approach will still require some level of visual inspection and quality control in a clinical setting. The data presented here should facilitate and expedite selection of the correct aligner for a particular task but they do not obviate the requirement for careful consideration nor further testing and analysis on the part of the end-user.

Competing interests

No competing interests were disclosed.

Grant information

The author(s) declared that no grants were involved in supporting this work.

References

1. Lupski JR, Reid JG, Gonzaga-Jauregui C, *et al.*: **Whole-genome sequencing in a patient with Charcot-Marie-Tooth neuropathy.** *N Engl J Med.* 2010; **362**(13): 1181–1191.
   PubMed Abstract | **Publisher Full Text**
2. NIH HMP Working GroupPeterson J, Garges S, Giovanni M, *et al.*: **The NIH Human Microbiome Project.** *Genome Res.* 2009; **19**(12): 2317–2323.
   PubMed Abstract | **Publisher Full Text** | **Free Full Text**
3. Treangen TJ, Salzberg SL: **Repetitive DNA and next-generation sequencing: computational challenges and solutions.** *Nat Rev Genet.* 2011; **13**(1): 36–46.
   PubMed Abstract | **Publisher Full Text** | **Free Full Text**
4. Moore B, Hu H, Singleton M, *et al.*: **Global analysis of disease-related DNA sequence variation in 10 healthy individuals: implications for whole genome-based clinical diagnostics.** *Genet Med.* 2011; **13**(3): 210–217.
   PubMed Abstract | **Publisher Full Text** | **Free Full Text**
5. Dahl A, Mertes F, Timmermann B, *et al.*: **The application of massively parallel sequencing technologies in diagnostics.** *F1000 Biol Rep.* 2010; **2**: 59.
   PubMed Abstract | **Publisher Full Text** | **Free Full Text**
6. Nielsen R, Paul JS, Albrechtsen A, *et al.*: **Genotype and SNP calling from next-generation sequencing data.** *Nat Rev Genet.* 2011; **12**(6): 443–451.
   PubMed Abstract | **Publisher Full Text** | **Free Full Text**
7. Pattnaik S, Vaidyanathan S, Pooja DG, *et al.*: **Customisation of the Exome Data Analysis Pipeline Using a Combinatorial Approach.** *PLoS One.* 2012; **7**(1): e30080.
   PubMed Abstract | **Publisher Full Text** | **Free Full Text**
8. Li H, Durbin R: **Fast and accurate long-read alignment with Burrows-Wheeler transform.** *Bioinformatics.* 2010; **26**(5): 589–595.
   PubMed Abstract | **Publisher Full Text** | **Free Full Text**
9. Lunter G, Goodson M: **Stampy: A statistical algorithm for sensitive and fast mapping of Illumina sequence reads.** *Genome Res.* 2011; **21**(6): 936–939.
   PubMed Abstract | **Publisher Full Text** | **Free Full Text**
10. Salzberg SL, Pertea M: **Do-it-yourself genetic testing.** *Genome Biol.* 2010; **11**(10): 404.
    PubMed Abstract | **Publisher Full Text** | **Free Full Text**
11. Szabo C, Masiello A, Ryan JF, *et al.*: **The breast cancer information core: Database design, structure, and scope.** *Hum Mutat.* 2000; **16**(2): 123–131.
    PubMed Abstract | **Publisher Full Text**
12. Homer N, Merriman B, Nelson SF: **Bfast: an alignment tool for large scale genome resequencing.** *PLoS One.* 2009; **4**(11): e7767.
    PubMed Abstract | **Publisher Full Text** | **Free Full Text**
13. Hercus C: Novoalign v2. 2011; 07.
14. Langmead B, Salzberg SL: **Fast gapped-read alignment with Bowtie 2.** *Nat Methods.* 2012; **9**(4): 357–359.
    PubMed Abstract | **Publisher Full Text** | **Free Full Text**
15. Sherry ST, Ward MH, Kholodov M, *et al.*: **dbSNP: the NCBI database of genetic variation.** *Nucleic Acids Res.* 2001; **29**(1): 309–311.
    PubMed Abstract | **Publisher Full Text** | **Free Full Text**
16. McKenna A, Hanna M, Banks E, *et al.*: **The genome analysis toolkit: a mapreduce framework for analyzing next-generation dna sequencing data.** *Genome Res.* 2010; **20**(9): 1297–1303.
    PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Current Referee Status: ☑ ☑ ☒

Referee Responses for Version 2

☒ **Mihaela Pertea**, Steven Salzberg

McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, MD, USA

Not Approved: 29 October 2012

Referee Report: 29 October 2012

Our 'Not Approved' status still maintains. It seems like he has made some nice improvements but the paper doesn't address our fundamental concern that, despite its claims, it doesn't evaluate aligners, but their capacity to work with the GATK pipeline. All the article really shows is how a particular program, GATK, functions in concert with different aligners. GATK has been fine-tuned to use BWA, as its own developers acknowledge. In our opinion these findings are misleading.

We have read this submission. We believe that we have an appropriate level of expertise to state that we do not consider it to be of an acceptable scientific standard, for reasons outlined above.

Competing Interests: No competing interests were disclosed.

1 Comment

Attila Berces, Omixon, Hungary

Posted: 07 Jan 2013

In this review I make arguments based on some computational evidence in support of Oliver's experimental design and make some observations on the reviews made by Pertea and Salzberg. I declare conflict of interest since I am involved with one of the alignment software reviewed in Oliver's paper. I note that Pertea and Salzberg chose not to declare conflict of interest in a similar position.

Before going into details, I would like to note that both Oliver in this F1000 article and Pertea and Salzberg in their Do-It-Yourself Genetic Testing paper in Genome Biology make important contribution to advance the use of next generation sequencing for BRCA analysis for the benefit of patients. However, they are approaching the same subject from a different perspective. These differences in perspective are reflected in the review of Pertea and Salzberg.

Oliver's experimental design reflects the fact that he carried out this study in a company considering mutational analysis based on next generation sequencing and that in practice such tests use deep targeted sequencing approaches and not whole genome analysis. Some of the arguments made by Pertea and Salzberg reflect experiences with relatively low depth whole genome or whole exome studies, but as I present some evidence here, deep targeted sequencing needs different considerations.

In addition to considering the type of experiment, Oliver considered the mutational profile of the target gene. In some populations one third of inherited breast cancer cases are linked to significant insertions or deletions. The inability of the NGS pipeline to detect long indels would mean that the method fails to screen the third of the target population. Oliver carefully considered variants relevant from this screening perspective. When a patient is diagnosed with deleterious BRCA1/BRCA2 mutations they have to make life-changing decisions that could be as severe as electing for preventive mastectomy or ovariectomy. Consequently, in BRCA screening or any other genetic tests, one cannot afford to have false negative mutations and thus false negatives are the critical evaluation criteria.

In the diagnostic analysis of BRCA, all variants are followed up by visual inspection of the underlying data and if necessary they are confirmed by Sanger sequencing. False positives are relevant to the extent they generate wasted effort of following them up. When the false positive rate is high, the wasted effort can offset the economic benefits of using NGS in the first place.

Oliver's experiment simulated deep, PCR primer-based, targeted sequencing data. The Haloplex and the Ampliseq kits are commercially available for BRCA analysis and are examples of experiments producing such datasets. Oliver considered homozygous mutations since only these are relevant from a genetic screening perspective.

A key argument made by Pertea and Salzberg is that GATK parameterization strongly influences the results and the false variants rate does not reflect the performance of the alignment algorithms. While this argument is correct for whole genome and even whole exome studies considering both homozygous and heterozygous variants, Oliver simulated deep targeted sequencing and examined homozygous mutations only. Restricting the analysis to homozygous mutations removes significant dependence of the results on GATK parameters. GATK parameterization impacts the results of deep, targeted sequencing experiments differently from that of whole genome sequencing. In order to examine the effect of variant calling parameterization as well as the reasons for false negatives, my colleague Tibor Nagy repeated the experiments in Oliver's paper using the Bowtie2/GATK pipeline. We examined the false negatives, as well as the dependence of the false positive rates on the GATK parameterization.

We have visually inspected the read alignment in the region of all 40 false negative variants produced by the single-end Bowtie2/GATK pipeline. The false negatives can be explained by no coverage for 11 mutations, by low mapping quality for 18 mutations, and by alignment error for 11 mutations. GATK parameterization will only affect one of these categories:

(1) The missing variants due to no-coverage will not be found by any parameterization of GATK.
(2) Low mapping quality reads supporting missing variants can be recovered by lowering the corresponding threshold in GATK.
(3) The alignment error is the most complex category and warrants further investigation beyond the scope of this review. However, better false negative rates reflect better read alignment by the aligner.

In contrast to shallow depth whole genome sequencing, where the false negative mutations can be affected by various GATK parameters, mapping quality threshold is the main factor impacting the results in Oliver's paper.

In order to examine the effect of re-parameterization of GATK on the number of variants, we used the read filter (-rf) option of GATK and reassigned mapping quality with the ReassignMappingQuality -DMQ 60 (unsupported) option. As we expected, we could reduce the number of false negatives from 40 to 22 and recovered the variants missing due to low mapping quality. However, this recovery comes at the cost of increasing the false positives from 66 to 304. This observation is in line with that of Pertea and Salzberg's that "in our experience we can easily increase the number of SNPs by a factor of 5-fold simply by varying its parameters, REGARDLESS of the alignments provided at the front end".

We also note here that most of the improvement in the false negative rates going from single-end to paired-end analysis are related to improved mapping quality since read pairing improves the specificity of read mapping. Consequently, for a paired-end dataset re-parameterization of GATK can mostly impact the false positive rates without much change in false negatives.

This evidence presented here is to the contrary of the main concern of Pertea and Salzberg that "our fundamental concern that, despite its claims, it doesn't evaluate aligners, but their capacity to work with the GATK pipeline." In fact this evidence shows that the false negatives are mostly affected by the ability of the aligners to produce (1) coverage around the mutation, (2) high mapping quality, and (3) correct alignment.

While it is true that false positive rate can be significantly impacted by GATK parameters, the main concern of genetic screening is to reduce false negative rates and to keep false positives within a reasonably low level. Based on the evidence presented here, false negative rates can only be improved by GATK re-parameterization at the cost of excessive increase in false positive rates compared to the default settings used in Oliver's paper. Mapping quality threshold is the main influencer of the single-end sequencing results but it has a standard definition independent of the mapping software. For this reason, it does not affect different aligners differently. Heterozygous mutation rates would be significantly impacted by GATK parameters, but it was not a subject of Oliver's study. Consequently, re-parameterization of GATK for every aligner would make the results less and not more comparable.

Pertea and Salzberg argue that „All the article really shows is how a particular program, GATK, functions in concert with different aligners. GATK has been fine-tuned to use BWA, as its own developers acknowledge." The actual findings of Oliver's paper are contrary to this argument since BWA is the second worst performer in false negative rate in single-end mode. If fine-tuning the GATK parameters to the particular aligner impacted the results more than the quality of read alignment then BWA/GATK would outperform other pipelines.

In my opinion Oliver's paper can benefit from the following improvements and revisions:
(1) Better explaining the implicit assumptions behind the design of the experiment
(2) Emphasizing more the importance of false negative variants from a diagnostic perspective
(3) Emphasizing that only the false negative rates are comparable across pipelines and that the actual values would depend on the mapping quality threshold
(4) Showing that the false positive rate is sensitive to the parameterization of the GATK and should only be compared in order of magnitude across pipelines
(5) Removing ppv values since those can be misleading

With these improvements this article fills a gap. In contrast to other comparative studies on mapping and alignment where the emphasis is made on re-mapping accuracy of the reads, this study gives a meaningful metric for those who want to use NGS for BRCA screening. This paper advances our understanding of NGS analysis and serves patients who are tested for BRCA mutations.

In the interest of those labs defying Myriad's BRCA patent and apply NGS for the benefit of patients, I ask Pertea and Salzberg to reconsider their rejection of the paper and give clear instructions to the author what reasonable changes he has to make in order for the paper to be accepted to advance science for the benefit of patients

Competing Interests: No competing interests were disclosed.

Referee Responses for Version 1

☒ **Mihaela Pertea**, **Steven Salzberg**
McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, MD, USA

Not Approved: 27 July 2012

Referee Report: 27 July 2012
In this paper the author sets out to investigate the performance of several alignment tools and to assess their ability to accurately detect known mutations when used in a variant calling pipeline. This is an important issue to address before designing a particular analysis pipeline for variant detection. However, this paper makes multiple very strong claims about the superiority of various alignment algorithms based on highly flawed computational experiments. Overall the results are at best misleading, and many of the conclusions are simply wrong.

Our concerns are related to the following issues:

1. Concerns about the experimental design:

The experiment claims to measure the accuracy, and in particular the sensitivity and FDR rate, for many sequence aligners. Unfortunately, it simply doesn't measure anything of the sort. Instead, it measures the sensitivity and FDR of the GATK SNP pipeline, a complex series of programs with many, many parameters, with different aligners fed into the very first step of GATK. GATK is exquisitely sensitive to these parameters; in our experience we can easily increase the number of SNPs by a factor of 5-fold simply by varying its parameters, REGARDLESS of the alignments provided at the front end. Unless the author optimizes GATK for each aligner – which he explicitly did not do – these results are simply invalid. Thus the whole experiment is deeply flawed.

It is not sufficient, in a benchmarking test like this one, to use only default running parameters (as the author says he did), and to make no effort at careful evaluation of what would be the best parameters to use for each aligner in that specific experiment. If the author wishes to compare aligners as part of a complex pipeline (GATK), he needs to do much more work than the simple push button runs he did here.

The whole point of simulated data ought to be that one can check each read and see if it was aligned to the correct place. This should be easy to do as all the reads are simulated and therefore their location is known a priori. If (and only if) the author compared the alignments to the true alignment, then he could report valid findings about the sensitivity at finding SNPs, indels, etc. He did not do this, which is somewhat astonishing. As it stands, the main results – including Tables 3 and 4 and Figure 1 – are simply wrong.

Also, in order to make his results reproducible, the author should provide the alignment results for all programs, as well as the exact command lines used for each aligner. Just specifying that he ran the aligners with parameters "as close as possible" to defaults is not enough.

2. Running time evaluations:

Another major conclusion of the paper concerns run times, which the author reports in a separate section (3.2). An obvious flaw here is that running the aligners on such small datasets (each only 200,000 reads) cannot properly differentiate the relative running times of the different programs, especially the faster ones. Exome sequencing, a very common experiment today, generates roughly 100 million reads per experiment – 500 times larger than each sample data used here. Whole-genome data sets are much larger. To provide any realistic run time findings, the author needs to load at least an exome-sized data set and run it. He doesn't need to use simulated reads – many exomes are publicly available. Since he is only measuring run time, he doesn't need to worry about the sensitivity of these alignments, just speed.

If the author wants to report findings about run-time, he needs to scrap this experiment and run a more realistic data set. If 100 million reads, not large by today's standards, swamps the ability of any aligner to handle it, then he can report that.

Other comments in the alignment section are not justified. For example, claiming that "most alignment times recorded here might be considered manageable for most purposes" seems to be little more than the author's unsupported opinion, based on a relatively tiny number of reads.

3. Other significant concerns:

a). The author used the Stampy package to simulate the reads from the BRCA1 region. What was the reason that this particular read simulator was used, and not another one that is independent from all aligners involved? E.g., the Mason simulator is considered to be relatively realistic. The Stampy simulator might give an unfair advantage to the Stampy aligner.

b). Why did the author align the simulated reads only to chromosome 17? If this is supposed to simulate a targeted sequencing experiment, why not just align to the BRCA1 region, which is far, far smaller than the entire chromosome? A much more realistic design would be to align to the whole human genome, which is normally done for real data where contamination from other parts of the genome is common. The author should also specify how he obtained the index required by the different aligners, and how long it took to create such an index (from the running times of the programs presented in the paper I assume this time was not included).

c). The way the programs were run is completely unclear, since no command line options are provided. Besides a step required to create an index (see above), some of the aligners require two steps to be run (e.g. BWA requires both an 'aln' and a 'samse/sampe' commands to be run; Stampy can be run in a hybrid version with a BWA option first). Were both of these steps included in the running times presented?

Most of these programs have many options that can increase their sensitivity at the cost (sometimes small, sometimes not) of increased run time.

d). The author makes a technical error in classifying aligners into two categories, “based on either hash tables or suffix trees.” The Burrows-Wheeler Transform (the basis of Bowtie, BWA, and SOAP2) is simply not a suffix tree. Further, it is not only simplistic but incorrect to state that hash-based programs are generally more sensitive, while the ones based on suffix trees are faster. That is wrong in multiple ways; there are many examples of hash-based approaches that are fast but not sensitive, and suffix-tree approaches don’t have to be faster. These features (speed/sensitivity) depend much more on the numerous implementation details, of which the author appears to be unaware.

e). The two wrapper scripts (sam2bam.sh and gatk.sh) that the author mentions that he made available do not seem to be present.

f). Each of the 67 data sets presented in the paper include 20 SNPs and 13 indels. Why use 67 data sets? And why have exactly the same number of SNPs and indels in each one? What criteria were used to include these particular numbers of SNPs and indels? Since each data set is representative for only one variant of the BRCA1 gene, how likely it is that in real data these 20 SNPs and 13 indels will appear at the same time in the gene? This is an unrealistic data set that has a bizarrely skewed bias.

g). The author states – when referring to Figure 3 – that the size of the indels influences their detection rates. He specifically says that the “size of the effect varied by aligner with BWA and Novoalign showing good detection rates for all but the largest deletions.” This statement is simply not correct: BWA cannot find large deletions (by design). Neither can Bowtie. However, GATK can find larger deletions in some cases, even if the input alignments don’t detect them. There are also entirely separate programs (e.g., Pindel) designed to find larger indels, and researchers looking for large indels know about these programs (and use them). This whole discussion again reflects the fundamental flaw in the experimental design: the author is measuring GATK’s performance, not the performance of the aligners. In addition, the author’s interpretation of Figure 3 seems biased, and is not supported by the data in the figure itself.

4. Minor concerns:

a). PPV is defined differently in the main body of the paper and in Table 6′s caption.
b). The author needs to include citations or at least web addresses for all the aligners presented in the paper.
c). I assume GLG in Table 5 is in fact CLC.
d). Where did the author collect the known mutations for the BRCA1 gene from? He needs to provide citations.

We have read this submission. We believe that we have an appropriate level of expertise to state that we do not consider it to be of an acceptable scientific standard, for reasons outlined above.

Competing Interests: No competing interests were disclosed.

Vera Kalscheuer
Department of Human Molecular Genetics, Max Planck Institute for Molecular Genetics, Berlin, Germany

Approved: 26 July 2012

Referee Report: 26 July 2012

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.

Competing Interests: No competing interests were disclosed.

☑ Thomas Friedman

Laboratory of Molecular Genetics, National Institute on Deafness and Other Communication Disorders (NIDCD), National Institutes of Health, Rockville, MD, USA

Approved: 25 July 2012

Referee Report: 25 July 2012

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.

Competing Interests: No competing interests were disclosed.