Increasing diagnostic yield of fast whole genome diagnostics for new borns

‘Improved variant calling of medium sized InDels’

Janine Reurink
Student Biomedical Sciences
s2011670
April 2nd 2015

Introduction

Recently, the 5 genes per minute project (5GPM) was launched in the UMCG. This project focusses on finding the genetic cause of disease in severely ill children, which were hospitalized in the neonatal intensive care (NICU). These new borns couldn’t be diagnosed with any other disease using the standard diagnostic methods. In order to find the probable genetic cause of their disease, next generation sequencing (NGS) is used. The results of this are analysed in order to find anything that could cause the disease.

Among the polymorphisms that can be detected are single nucleotide polymorphisms (SNPs) and short insertion or deletion (InDels) that might be responsible for the disease. However, any deviation larger than a few base pairs (bp) and smaller than what can be found using arrays cannot be called optimally. This includes the so called medium sized InDels, which range from 30 bp to the read length (100 bp in this case).[1]

The aim of this project is to find the most optimal tool (or tools) for detection of medium sized InDels in samples of the 5GPM project.

Tools

The tools that were included in the testing are listed below. Table 1 gives an overview of the methods that are used by the tools to detect InDels.

Gustaf

Gustaf (Generic mUlti-SpliT Alignment Finder) is a SV detection tool that claims to detect structural variants from 30 to 100 bp and larger than 500. It uses a generic multi-split alignment to detect the breakpoints of such a SV. Unfortunately, Gustaf requires input files in GFF format, which can be generated from a BAM file using the local aligner Stellar.[2]

Delly

Delly is a tool that uses short insert paired-ends, long-range mate-pairs and split-read alignments to detect structural variants. In this project, only insertions and deletions are tested.[3]

ABRA

ABRA is an assembly-based realigner that uses de novo assembly and realignment to detect insertions, deletions, SNPs, MNPs and complex events. Unfortunately, ABRA produces output in BAM-format. The call variant tool Freebayes had to be implemented to get output in VCF format.[4]

MATE-clever

MATE-clever stands for Mendelian-inheritance-AtTEntive CLique-Enumerating Variant finder and claims to find InDels longer than 30 bp. The tool focusses on finding InDels in family trio’s and quartets, but has the possibility to test single samples as well.

MATE-clever combines the tools Clever (for detection of InDels shorter than 100 bp) and Laser (for alignment of split reads).[5]

Scalpel

Scalpel claims to detect InDels by combining mapping and assembly.[6]

CoNIFER

Conifer uses the number the number of sequenced reads to calculate a so called RPKM-value. When the RPKM reaches a certain threshold, duplication and deletion breakpoints can be detected. Unfortunately Conifer requires at least 8 samples to prevent systemic bias.[7]

Table 1: Display of the methods the tools claim to use for detection of InDels.

Tool \ Type of detection	***Coverage***	***Split reads***	***Insert size***	***Realignment/ De novo***
GUSTAF		*x*		
DELLY		*x*	*x*	
ABRA				*x*
MATE-clever		*x*	*x*	
Scalpel		*?*		*x*
CoNIFER	*x*			

Methods
The tools that were used for testing were selected from literature. They were reviewed according to several criteria:

- Documentation
Documentation about how to install and use the tool should be present and up to date.

- *Feasibility and workability of the tool*
  The tool should be easy in installation and use.
- *Detects medium sized INDELS in single sample NGS data*
  The tools should be able to detect medium size InDels in single samples, ranging from 30 bp to 100 bp.
- *Common input and output*
  The tools should accept BAM and/or FastQ files as input and produce VCF-files as output.
- *Written in a common language*
  The tool should be written in a common language, to ease handling and solving possible errors.
- *Speed of analysis*
  Although not often mentioned in literature. And of course dependant of the size of the DNA that is tested and the number of threads used for computing. Time of analysis should preferably be within a day.
- *Sensitivity*
  The number or true positives, false positives and false negatives detected by the tool. The amount of true negatives can be infinitive, so specificity cannot be calculated.

The selected tools were installed on the cluster and tested on both in silico test data of chromosome 20 and real NGS data. The in silico test set was generated containing several SNPs, insertions, deletions and other CNVs of various sizes. When comparing the InDels present in the test set with the calls the tools made, the number of true positives, false positives and false negatives can be calculated. From there, the sensitivity can be calculated for each tool as a degree of validity.

Results were stratified on their size in base pairs. Some of the tools claim to detect InDels until the read length (100 bp), while others do detect larger deletions. This can result in lack of results in a specific range in size for some tools. This might lead to calling a missed result as false negative, while it’s actually outside the detection range of that tool. By stratifying the result in two groups, from 30 to 100 bp and bigger than 100 bp, such distorted results could hopefully be prevented.
The results were compared with the InDel present in the test set using Bedtools intersect. A mutual overlap of respectively 50, 75 and 90% between the true results and the detected results was required for giving a true positive call. This can give some insight in the precision of the calls.
The real data was tested with the tools that gave reasonable results for the test data. The size distribution of the results was plotted to see if the results looked as could be expected. Then, the results were compared with each other using Bedtools intersect with a required 50% mutual overlap. Although the tools had a different range of detection size, this could gain a little insight in the probability of the results.

Results

Performance of tools

Gustaf

Analysis with Stellar on the test set couldn't be performed on multiple threads. After 170 hours only 1% of the reads were analysed. Hence, it was decide to stop the alignment and omit the tool for further analysis.

Delly

Delly performed well for detection of deletions. Analysis of the whole exome was performed in 9 minutes on one thread. However, detection of insertions did not succeed well. The tool seemed to require more than 200 gb memory, after which it crashed due to exceeding the memory limit. Therefore, detection of insertions with Delly couldn't be continued.

If not testing on whole genome data, Delly needs an exclude file instead of a regular regions file. This file was made by running Bedtools complement on the regions bed-file to the reference genome file. The command used to test Delly on sample 1404 can be found in appendix 1.

ABRA

It took some effort to get ABRA running, including sorting and indexing of the BAM file and variant calling with Freebayes. The resulting VCF file contains SVs in 4 categories, insertions, deletions, MNPs and complex events.

When creating the regions BED file for ABRA, it appeared ABRA cannot handle more than 3 columns in the file. Awk was used to extract the first three columns. Subsequently, the regions file for freebayes had to be created using the freebayes script fasta_generate_regions.py.

ABRA contains the options to enable structural variant searching and to assemble unaligned reads. This was enabled to see if this could improve the results. Unfortunately, it did not work for the test set. The options did work on the real data, but did not seem to influence the amount of calls that were made. Therefore, the results of ABRA shown below do not include these two options in its command. Analysis of the whole exome was performed in less than 3 hours on 8 threads. The actual command for testing of sample 1404 can be found in appendix 2.

MATE-clever/ Clever

Unfortunately, MATE-clever gave an error when trying it on the test set. Clever did work on this test set, and the results produced by Clever were used for validation. In the first place, MATE-clever seemed to generate results when testing on the real data. Unfortunately, it appeared that MATE-clever had caused an error and these results were actually generated by Clever. The MATE-clever error could not be solved immediately and MATE-clever couldn't be tested at all. All samples were tested with Clever instead. This gave results in the right format (VCF) but did not give a genotype for a call.

The command used for testing Clever (and MATE-clever as well) on sample 1404 can be found in appendix 3. Analysis of the whole exome was performed in 22 minutes on 8 threads.

Scalpel

Scalpel was tested on the test set, but the output VCF-file was either empty or barely contained information about the found structural variant. Since other tools did give promising results when results for Scalpel were generated, testing with Scalpel was aborted.

Conifer

Conifer requires at least eight samples to perform solid detection. These were not available, so testing with Conifer couldn't be continued.

Testing was continued with the three tools that gave promising results: ABRA, Delly & Clever.

Results on in silico test data

Since the three remaining tools had a different range of detection size, results were stratified in two groups, from 30 to 100 and from 100 bp. Furthermore, distribution of the found InDels by the tools was computed. The same was done for the in silico generated test set as shown in figure 1. Seven InDels of 30 to 100 bp (3 insertions and 4 deletions) and 17 InDels longer than 100 bp (7 insertions & 10 deletions) were present in the test set. These InDels were compared with the InDels found by the tools.

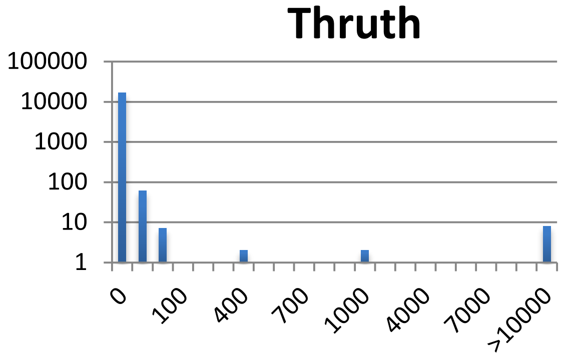

Figure 1: Distribution (per 10 bp) of InDels in the in silico test set. The most right column indicates all InDels bigger than 10000 bp.

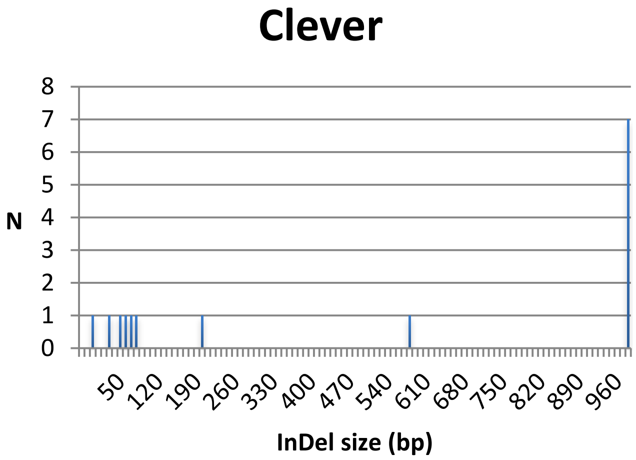

Figure 2: Distribution (per 10 bp) of InDels found by Clever. most right column indicates all InDels bigger than 1000 bp.

Clever

Since MATE-clever didn't run on the test data, analysis was only performed with Clever. In total 15 InDels were detected, ranging from 24 to almost 40000 bp. Most InDels that were found were deletions. Results are shown in figure 2.

ABRA

Figure 3 shows the results of analysis with ABRA. As shown, ABRA distinguishes between insertion, deletions, MNPs and complex events. Complex events can be any combination of SNPs, MNPs and InDels. Between 30 and 100 bp, 8 complex events were detected. One deletion bigger than 100 bp was detected (not shown).

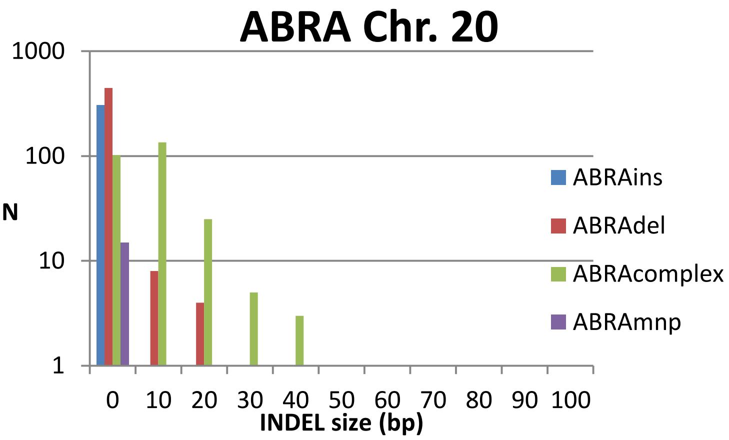

Figure 3: Distribution (per 10 bp) of InDels found by ABRA.

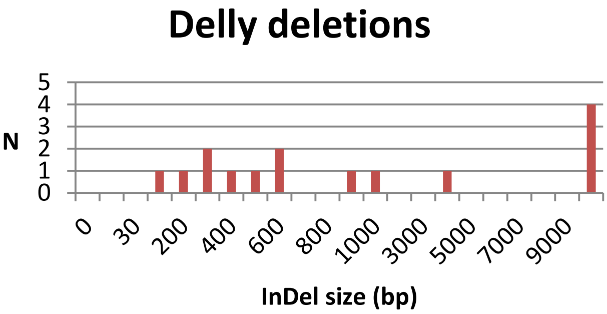

Figure 4: Distribution of deletions found by Delly. The most right column indicates all deletions bigger than 10.000 bp.

Delly

Figure 4 shows the results of the InDels detected by Delly. 15 deletions bigger than 100 bp were found.

The results of the comparison of the InDels detected by these three tools and the true InDels present in the test set are shown in figure 5 and figure 6. The number of found true positives is plotted against the found false positives. The requirements for calling a detected InDel true positive were a minimal mutual overlap with the true InDels in the test set of respectively 50(least strict), 75 and 90%(most strict). Delly and Clever also give a quality score for the call (either pass or low quality) and the precision of the breakpoints (precise or imprecise). When using only calls that give a pass and precise score, the amount of false positives decrease for Delly. Clever unfortunately did not give any calls with that score.

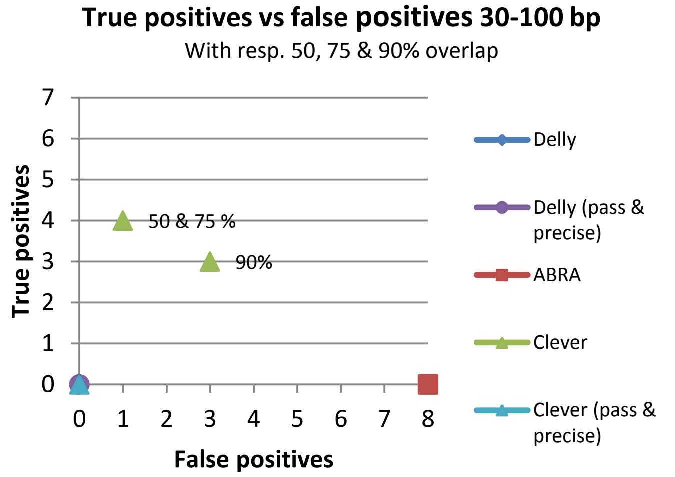

Figure 5: The found true positives InDels of 30 to 100 bp plotted against the number of detected false positives. In an ideal situation, a tool should have found 7 true positives (the number present in the test set) and 0 false positives. The required mutual overlap between the real InDel in the test set and the InDel detected by a tool was set at 50, 75 and 90%. This can lead to up to 3 data points for each tool, but the results can overlap as well.

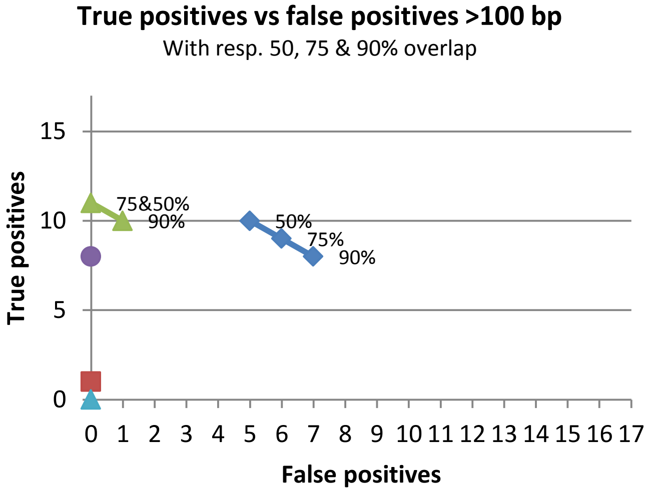

Figure 5: The true positives of >100 bp plotted against the false positives. 17 InDels were present in the test set. The required mutual overlap between the real InDel in the test set and the InDel detected by a tool was set at 50, 75 and 90%. This can lead to up to 3 data points for each tool, but the results can overlap as well.

Results on real data

The three tools were tested on real samples as well. Those two samples (1404 & 1402) were real unsolved cases from the 5GPM project. Although the results of this testing couldn't be validated, it might still be interesting to see what these tools detect. The distribution of the InDels detected in sample 1404 by the tools is shown in figure 7, 8 & 9. Results for sample 1402 are in appendix 4. The results of 1402 look similar to those of 1404, except for Clever. That tool gave little of result for sample 1402. This might be caused by lack of insert size selection during processing the sample in the lab.

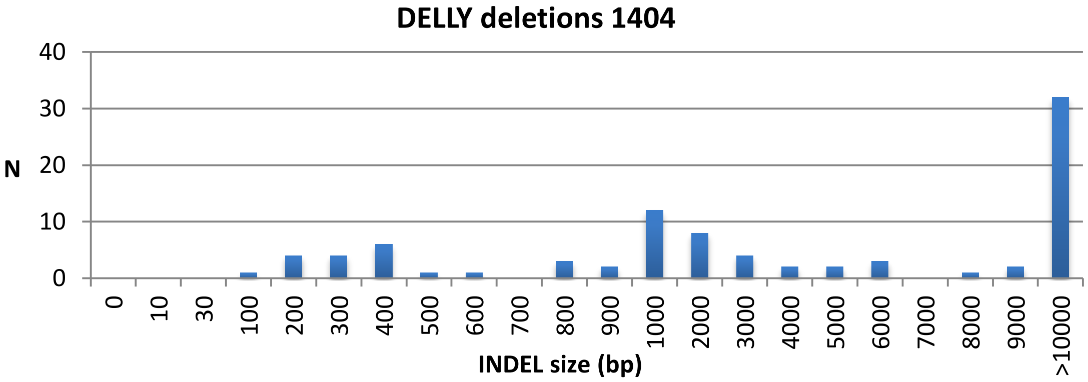

Figure 6: Distribution of the deletions detected by Delly.

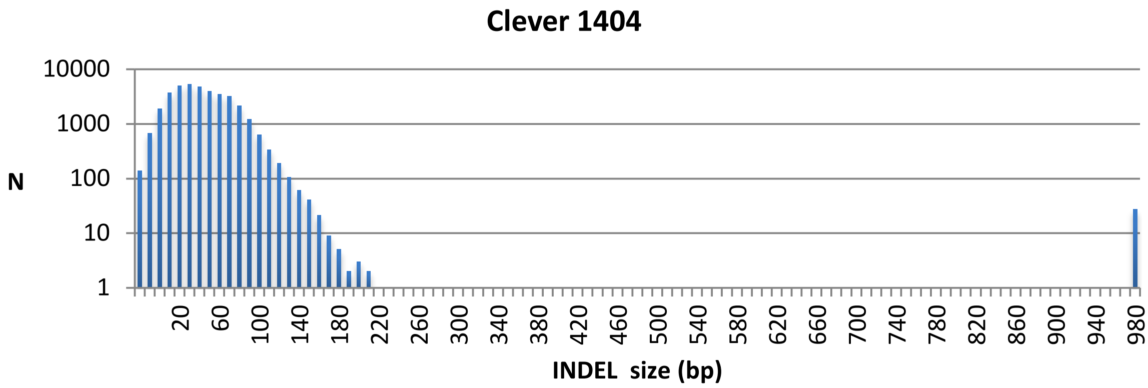

Figure 8: Distribution of the InDels detected by Clever. The graph is in log-scale, which causes that single calls are not visible, although they were present in the region between 220 and 990 bp. The most right column shows all InDels bigger than 1000 bp.

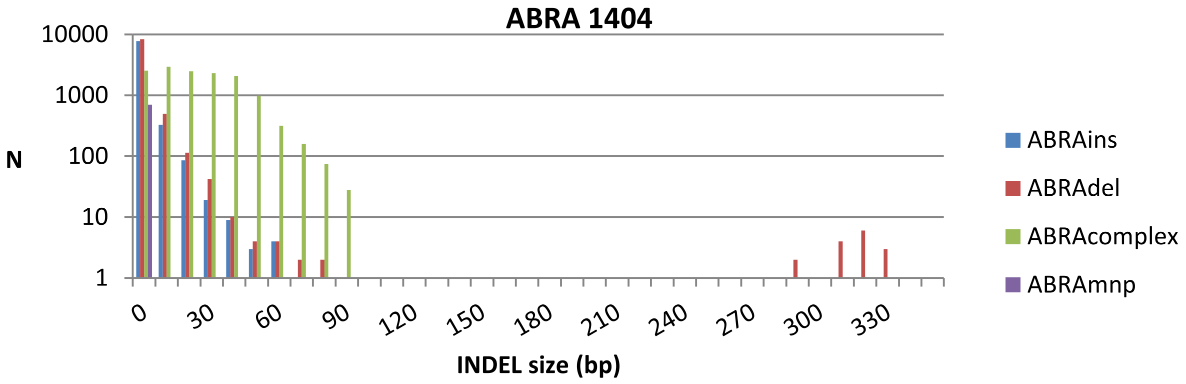

Figure 7: Distribution(per 10bp) of the InDels called by ABRA. ABRA distinguishes between insertions(ins), deletions(del), complex events and MNPs. The graph is in log-scale, which causes that single calls are not visible, although there were some present between 100 and 300 bp and bigger than 350 bp.

Although these results couldn't be validated, they were compared with each other. This might gain some insight in the overlap in the InDels that were detected by each tool. Comparison was performed with Bedtools intersect with a minimal required mutual overlap of 50%. The number of overlapping InDels is shown in figure 10. Delly gives less calls then ABRA and Clever. This can be explained by the fact that Delly only calls deletions larger than 100 bp. Those are less abundant than the medium sized InDels that the other tools find.

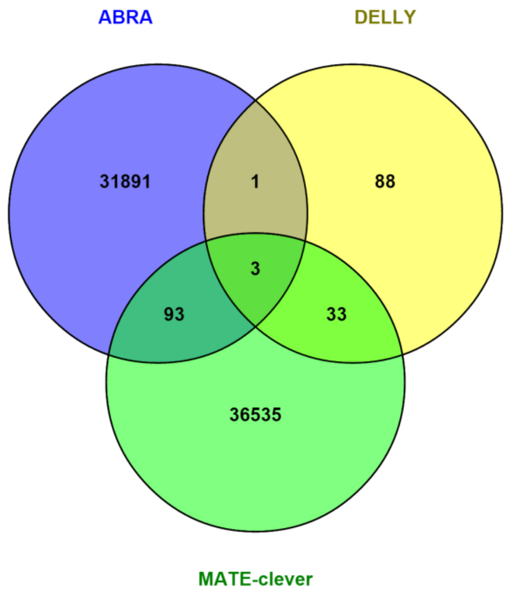

Figure 10: The overlap in detected InDels of the three tools.

Discussion and conclusion

All three tested tools show that they are able to give sensible InDel calls. However, there is still space for improvement and some remarks can be made on the results of the tools.

For ABRA, complex events should be omitted from the results, because they can be any combination of SNPs, MNPs & InDels. In the test set, several complex events bigger than 30 bp were detected by ABRA. When further considered, it turned out that these complex events were actually a number of SNPs located closely to each other and not an actual InDel. These events might be useful for diagnostic purposes, but they can create a bias if it's not known how to handle those complex events. Possibly such complex events can be reanalysed.

MATE-clever needs more research as well. As explained before, MATE-clever does not work well yet. An error appeared when trying to run MATE-clever after Clever on several samples. Since clever is an InDel caller on its own as well, a result in VCF format was generated anyway. These calls did not contain genotype results, while MATE-clever claims to give those. When MATE-clever could also give calls including genotype, this can improve usability of the results.

Fortunately, Delly gives such genotype calls. The tool can distinguish between heterozygous and homozygous deletions. In the test set it detected a true homozygous deletion of almost 40 kb.

ABRA also gives genotype calls of either 0/1 for a heterozygous event and 1/1 for a homozygous event. This information can improve the usefulness of a call. Heterozygous InDels are not necessarily harmful when the other allele of the gene is normally present. Information of the zygosity can therefore be crucial to determine the probability of a certain InDel to be cause of the disease.

In general the overall results look promising, but there is definitely more validation required. The three considered tools seem to all have a different range of detection. Especially Delly starts to make calls from 100 bp. ABRA calls up to the read length of 100 bp, except from deletions which can be detected even when they are larger than 100 bp. Clever can detect InDels up to larger than 1000 bp, although it mainly calls deletions, but barely insertions. This indicates that those tools mainly detect insertions. Perhaps other tools can be used for detection of insertions.

All tools meet the required criteria that were mentioned in the methods section, namely: good documentation, feasibility, ability to call medium sized InDels in single samples, the required input and output format, the language it's written in, speed of analysis and the sensitivity.

However, the speed of analysis and sensitivity vary for the tools. The sensitivity is visualized in figure 5 and 6 and seems to be best for Clever and Delly. For ABRA, sensitivity seems to be lower. This might be due to the fact that only complex events were called between 30 and 100 bp, and no insertions or deletions. The speed of analysis ranges between 9 minutes for Delly deletions until almost 3 hours for ABRA. Clever takes 22 minutes.

Combining the results of tools would increase overall yield of InDel detection. Especially Delly and Clever should be used, because they show to give quite sensitive results and together cover a broad range of detection. However, more validation and possibly optimization should be performed.

References

1. http://www.molgenis.org/wiki/Courses/ComputationalMolecularBiologyResearch/P1.

2. Trappe, K., Emde, A. K., Ehrlich, H. C. & Reinert, K. Gustaf: Detecting and correctly classifying SVs in the NGS twilight zone. *Bioinformatics* 30, 3484-3490 (2014).

3. Rausch, T. *et al*. DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* 28, i333-i339 (2012).

4. Mose, L. E., Wilkerson, M. D., Hayes, D. N., Perou, C. M. & Parker, J. S. ABRA: improved coding indel detection via assembly-based realignment. *Bioinformatics* 30, 2813-2815 (2014).

5. Marschall, T., Hajirasouliha, I. & Schonhuth, A. MATE-CLEVER: Mendelian-inheritance-aware discovery and genotyping of midsize and long indels. *Bioinformatics* 29, 3143-3150 (2013).

6. Narzisi, G. *et al*. Accurate de novo and transmitted indel detection in exome-capture data using microassembly. *Nat. Methods* 11, 1033-1036 (2014).

7. http://conifer.sourceforge.net/.