Project No. *283496*

transPLANT

Trans-national Infrastructure for Plant Genomic Science

Instrument: **Combination of Collaborative Project and Coordination and Support Action**

Thematic Priority: FP7-INFRASTRUCTURES-2011-2

D12.1
Development and test of sophisticated statistical methods to model variation in large plant genomes

Due date of deliverable: 31/08/13
Actual submission date: 6/09/13

Start date of project: 1.9.2011 Duration: 48 months

Organisation name of lead contractor for this deliverable: INRA

Project co-funded by the European Commission within the Seventh Framework Programme (2011-2014)		
Dissemination Level		
PU	Public	x
PP	Restricted to other programme participants (including the Commission Services)	
RE	Restricted to a group specified by the consortium (including the Commission Services)	
CO	Confidential, only for members of the consortium (including the Commission Services)	

Contributor

INRA (5), BIOGEM (7)

Introduction

Deliverable reference number:12.1

The aim of this deliverable is to assess the ability of different tools to detect SNPs and indels, particularly large indels, using NGS data with a reference genome assembly.

Several tools are available to detect structural variations (SVs). However they all have some specificity according to the type of variation they can detect (see Table 1 for an overview).

tool	method	deletion	insertion	inversion	translocation	duplication	CNV	comment
NovelSeq	Assembly	no	Yes	no	no	no	no	not mapped and singly mapped reads assembly
SOAPIndel	Assembly	Yes	Yes	no	no	no	no	Finds inserted sequence
inGAP-SV	combination	yes	Yes	yes	yes	yes	yes	includes visualization of results with graphical interface
SVMerge	combination	yes	Yes	yes	yes	yes	yes	combines different tools
CnD	RD	no	No	no	no	no	yes	specific to Mouse
CNVnator	RD	no	No	no	no	no	yes	maps ambiguous reads randomly
ReadDepth	RD	no	No	no	no	no	yes	GC content and mappability corrections
BreakDancer	RPM	yes	Yes	yes	yes	no	no	provides a score can deal with Mate-Pair data
HYDRA	RPM	yes	Yes	yes	yes	no	no	documentation on mapping step available
PEMer	RPM	yes	Yes	yes	yes	yes	no	can deal with 454 data
SVDetect	RPM	yes	Yes	yes	yes	yes	no	both clustering and sliding window strategy can deal with Mate-Pair data
SECluster	SMR	no	Yes	no	no	no	no	included in SVMerge
ClipCROP	SR	yes	Yes	yes	yes	yes	no	finds exact breakpoint using soft-clipped reads
Pindel	SR	yes	Yes	yes	yes	yes	no	finds exact breakpoint using singly mapped reads anchors or soft-clipped reads

Table 1: Selection of tools and methods used to call SVs from NGS data. RD = Read Depth ; RPM = Read Pair Map ; SMR = Singly Mapped Read ; SR = Split Read

Our study focuses on long indels detection, anchored on a reference sequence, in the aim to have tools able to move from species description with a single reference genome toward pan-genome full description. We tested two different strategies, one based on Mate-Pair (MP) and the other on Pair-End (PE) reads.

The MP strategy is based on observed length inconsistencies between the two MP reads positions when mapped on the reference genome. Any discrepancy suggests an insertion (if shorter than expected) or a deletion (if longer) on the sample studied. MP reads would allow the detection of large indels, whereas PE could not as indels should be larger than the PE insert size. We tested this approach with real-world data from Maize.

The PE strategy requires the mapping and the assembly of the reads to detect indels larger than the PE insert size. This approach needs higher read coverage for the assembly, but would allow the detection of break points and the recovery of inserted sequences. For this approach, we used Grapevine real-world and simulated dataset.

Methods

1 The MP strategy

We sequenced 2 lanes of MP library from a given maize genotype, on a HiSeq2000 Illumina sequencing machine, which resulted in approximately 18x of read coverage (for more details see http://supportres.illumina.com/documents/myillumina/0a36163e-5fc0-4ae0-a944-a0ee51aa0eb2/matepair_v2_2-5kb_sampleprep_guide_15008135_a.pdf).

We used novoalign to map reads against B73 genome with default settings. We ran the different SV callers on the mapped result, filtered for reads mapped with a MAPQ greater or equal to 30, in order to have high quality results anchored on low copy regions. Results less than 500bp from a gap on B73 assembly as well as those in

centromeres, were filtered out in order to avoid false positives. SVDetect[1] results were filtered using the number of supporting pairs: 5 minimum. SECluster[2] results were also filtered according to number of supporting single mapped reads: 5. All results were filtered to keep only SVs longer than 100bp and shorter than 5Mb.

2 The PE strategy

Two main strategies can be used for such detection. They both require the mapping of sequenced reads onto the reference, prior to detection. We used BWA[3].

A first method is based on variant callers aimed at retrieving SNPs and/or indels, using the mapping information such as mismatches/gaps, split-reads or PE insert sizes variations. These tools are mostly used to detect SNPs, small (1-3) and medium indels (4-30). We have tested MAPHiTS[4] (a pipeline using Samtools) , SOAPindel[5] and Pindel[6].

The second method is based on reference-guided assembly, where all the alignments of the re-sequenced reads against the reference are used to derive consensus draft of contigs. Secondly, the unmapped reads are *de novo* assembled, and anchored on the obtained contigs using PE information, in order to assemble the newly sequenced genome. This approach has been proven to generate a more accurate assembly than *de novo* approach, using less computing time. Finally, the resulting assembly is aligned onto the reference to predict structural variations (SV) from the unaligned sequences, either on the reference or the guided-assembly. We expected that this method could retrieve larger indels than others. We have tested Velvet-Columbus[7] for guided assembly, followed by Nucmer[8] for genomes alignment. For large indels detection, the contigs were scaffolded using PE information with SSpace[9] prior to genomes alignment.

In order to accurately assess the variant calling results, we implemented quite an unusual approach: in our benchmarking dataset, we simulated SNPs and indels in *Vitis vinifera* reference genome assembly (Chromosome 17) and used original PE reads from Illumina re-sequencing of the same cultivar. Thus we developed a sequence variation simulator to create SNPs and indels in the *V. vinifera* reference sequence,. Deletions and insertions into reference genome simulate insertions and deletions in the re-sequenced genome reads, respectively. This tool will be soon available, as a part of the REPET package (http://urgi.versailles.inra.fr/Tools/REPET). This approach allowed us to accurately calculate the sensitivity and False Discovery Rates of the benchmarked tools with real world reads. In addition, to assess the impacts of increased re-sequencing coverage, we also simulated high-coverage PE sequencing, based on the native *V. vinifera* genome sequence (chromosome 17).

Three simulated chromosome 17 were generated relying on 3 different classes of indels sizes. Small (1 to 3 nt),

[1] Bruno Zeitouni; Valentina Boeva; Isabelle Janoueix-Lerosey; Sophie Loeillet; Patricia Legoix-ne; Alain Nicolas; Olivier Delattre; Emmanuel Barillot (2010) SVDetect: a tool to identify genomic structural variations from paired-end and mate-pair sequencing data. Bioinformatics 2010; 26: 1895-1896

[2] Wong K, Keane TM, Stalker J, Adams DJ (2010) SVMerge: Enhanced structural variant and breakpoint detection by integration of multiple detection methods and local assembly, Genome Biology, 11:R128

[3] Li H. and Durbin R. (2010) Fast and accurate long-read alignment with Burrows-Wheeler Transform. Bioinformatics, Epub. [PMID: 20080505]

[4] http://urgi.versailles.inra.fr/Tools/MAPHiTS

[5] SOAPindel: Efficient identification of indels from short paired reads, Genome Research 2012

[6] Ye K, Schulz MH, Long Q, Apweiler R, Ning Z. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. Bioinformatics. 2009 Nov 1;25(21):2865-71. Epub 2009 Jun 26

[7] Velvet: algorithms for de novo short read assembly using de Bruijn graphs. D.R. Zerbino and E. Birney. Genome Research 18:821-829

[8] Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL: Versatile and open software for comparing large genomes, Genome Biol. 2004;5(2):R12. Epub 2004 Jan 30

[9] Boetzer M, Henkel CV, Jansen HJ, Butler D and Pirovano W. 2010. Scaffolding pre-assembled contigs using SSPACE. Bioinformatics.

medium (4 to 30 nt) and large (> 2Kb, average size=4Kb). The large indels were based on transposable elements sizes (2 to 5kb), previously annotated on this grape chromosome. Locations for SNPs, insertions/deletions were randomly chosen, not overlapping between each other. The small and medium indels datasets both contains 100 insertions, using random sequences, and 100 deletions. The large indels dataset contains 409 insertions, using random sequences, and 409 deletions. In addition, about 2% of SNPs randomly located where added in the three datasets.

We used 2 real world (76nt PE reads at 6x and 17x coverage) and 2 simulated Illumina datasets (76nt pair-end reads, respectively 17x and 60x coverage) from *V. vinifera* (chromosome 17). The simulated reads datasets were generated using the wgsim tool (SamTools).

We benchmarked the tools using the following metrics:

- Sensitivity: it assesses the proportion of true-positive SVs (SNPs, INDEL) that have been retrieved. High value is expected for good sensitivity (up to 100%)
- False Discovery Rate (FDR): it estimates the percentage of false positive SVs retrieved by a given tool. We used thereafter the value "1-FDR" to give a good idea of the specificity. High value expected for good specificity (up to 100%)

Results

3 The MP strategy

89% of reads could be aligned to B73, but only 41% as proper MP (according to insert-size and orientation of the reads). Indeed the MP library protocol generates a certain amount of PE data (14% in our case) and chimeric reads as well.

Although, some of these tools are capable of calling different types of SVs (including translocations for example), we are only interested here in indels. Approximately 5% of results from each type (insertion/deletion) and from each tested tool were randomly selected for validation. This validation was done using sequence capture. The experiment was designed so that it was able to capture flanking regions of each SV. Both genotypes (under test and B73) were captured and sequenced using this same design. Captured reads were then aligned to B73 genome:

1- to confirm the presence of the SV with tested genotype reads at the exact same location,
2- to confirm the absence of SV with B73 reads.

	SVDetect		Pindel		SECluster
	deletions	insertions	deletions	insertions	insertions
raw SV calls	97601	16677	4738	1207	46441
gaps and centromere filtered	45268	14377	3551	1149	44380
selected SVs	13432	4097	3551	1149	13835
Random selection for validation	756	20	193	40	904
validation rate	**46.85%**	**0.00%**	**83.23%**	**48.65%**	**8.98%**

Table 2: Validation results for SVDetect, Pindel, and SEcluster. Validation rates correspond to true positives percentage (Specificity)

As reported in Table 2, Pindel gives better results than SVDetect for calling deletions (83% validated compared to 47% for SVDetect), even though SVDetect finds 3 times more deletions. However, if Pindel seems to be the

best choice when looking for deletions with low false positive rate and exact breakpoint definition, SVDetect could be an interesting complement, when exhaustivity is important and exact indel breakpoint not necessary.

However, none of the tools have more than 50% of validated insertions results. Here again, the best choice seems to be Pindel with 49% of validation, while SVDetect has 0% and SECluster 9%. These tools are thus not very good at finding insertions, even with MP data.

4 The PE strategy

First, we compared the performance of *de novo* assembly from PE reads using the Velvet assembler, with guided assembly using its Velvet-colombus module (Table 3). We compared the assemblies obtained with the reference genome (note that PE reads comes from the same grape clone accession). We confirm here that guided assembly generates a better assembly than *de novo* approach, using less computing time.

Coverage	*De novo* – Control chr17		Guided – Control chr17	
	N50	Reference Genome Fraction Covered	N50	Reference Genome Fraction Covered
6x	1813	76.41%	1972	86.24%
17x	1750	79.75%	6611	92.62%

Table 3: De novo *vs* guided assembly comparison

Different parameters combinations have been tested, only best ones are presented here. The native *V. vinifera* chromosome17 was also used as a negative control for variants calling tests (data not shown).

4.1 SNPs

Even if our goal is to detect indels, we took the opportunity to also test for SNP detection as the tools and the data used will allows such a benchmark. Our test shows (Figure 1) that retrieving SNPs is not a problem, even with low coverage re-sequencing data (i.e. 6x). MAPHiTS is very sensitive and quite specific, while Nucmer is less sensitive but often more specific. This indicates a higher confidence on SNPs supported by an assembly. Note that there is also the possibility to cross results between tools to find more reliable SNPs. As expected, higher re-sequencing coverage improves both sensitivity and specificity, especially for MAPHiTS.

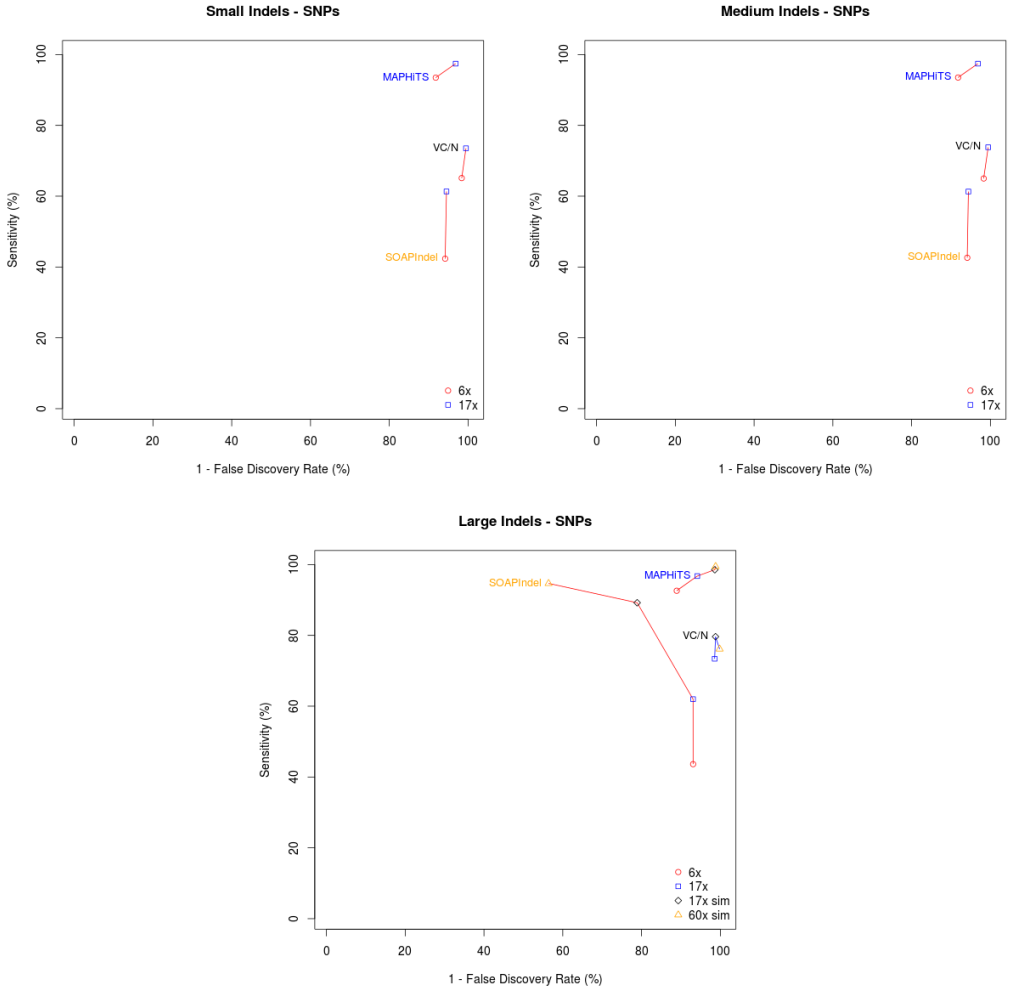

Figure 1: SNP detection in the 3 simulated chromosomes: "Small Indels", "Medium Indels", and "Large Indels"

4.2 Small and medium indels:

Indels smaller than read size can be detected by methods based on reads mapping, as they use reads alignment gaps and read clipping. Methods based on PE insert size variation (here 423 bp including reads size), such as those tested in the MP strategy, cannot detect small and medium indels as their impact on insert size is not significant.

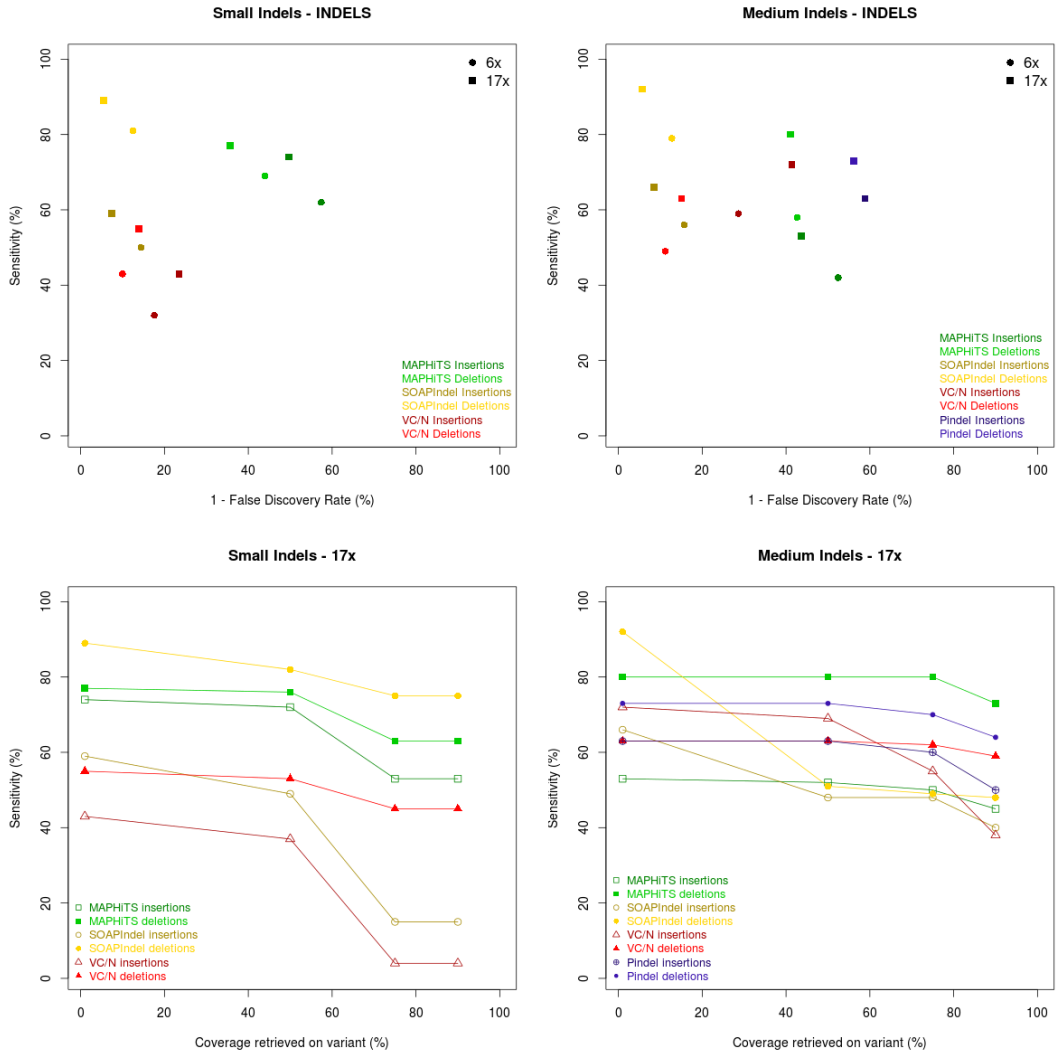

Figure 2: Small and Medium indels detection in the "Small Indels" and "Medium Indels" simulated chromosomes. Coverage retrieved on variant corresponds to the fraction of the variant retrieved by the SV caller, sensitivity is cumulative.

Our test shows (Figure 2) a high false positive rate for both small and medium indels. Consequently, further filtering and validation would be needed for this approach to be really useful. Moreover, even when indels are detected they do not overlap the whole expected/inserted SV. MAPHiTS is more specific in these cases as it detects more accurately variants boundaries. Pindel has a similar profil, with slightly better FDR but lesser sensitivity. Note that deletions are better detected than insertions.

4.3 Large indels:

As expected, methods based on read mapping like MAPHiTS are not able to retrieve indels larger than reads length (Figure 3). Simulated indels lengths seem to be too large to be detected by insert size variation based methods like SOAPindel. On the contrary, Pindel manages to fetch large deletions with good sensitivity and specificity. To do so, when only the first member of a pair is anchored, it tries to uniquely map the second one within a user-specified distance. This procedure allows to span the deletion and to detect its boundaries almost perfectly, but can't be applied on insertions. Reference-guided assembly approach (VC/N) has much better results on both large insertions and deletions, as expected. Large gap in contigs without mate on the reference is considered as signature of a large insertion in the re-sequenced genome, whereas the reverse observation signs a large deletion. Specificity is good, but sensitivity is still quite low, even with scaffolding and parameters fine-tuning.

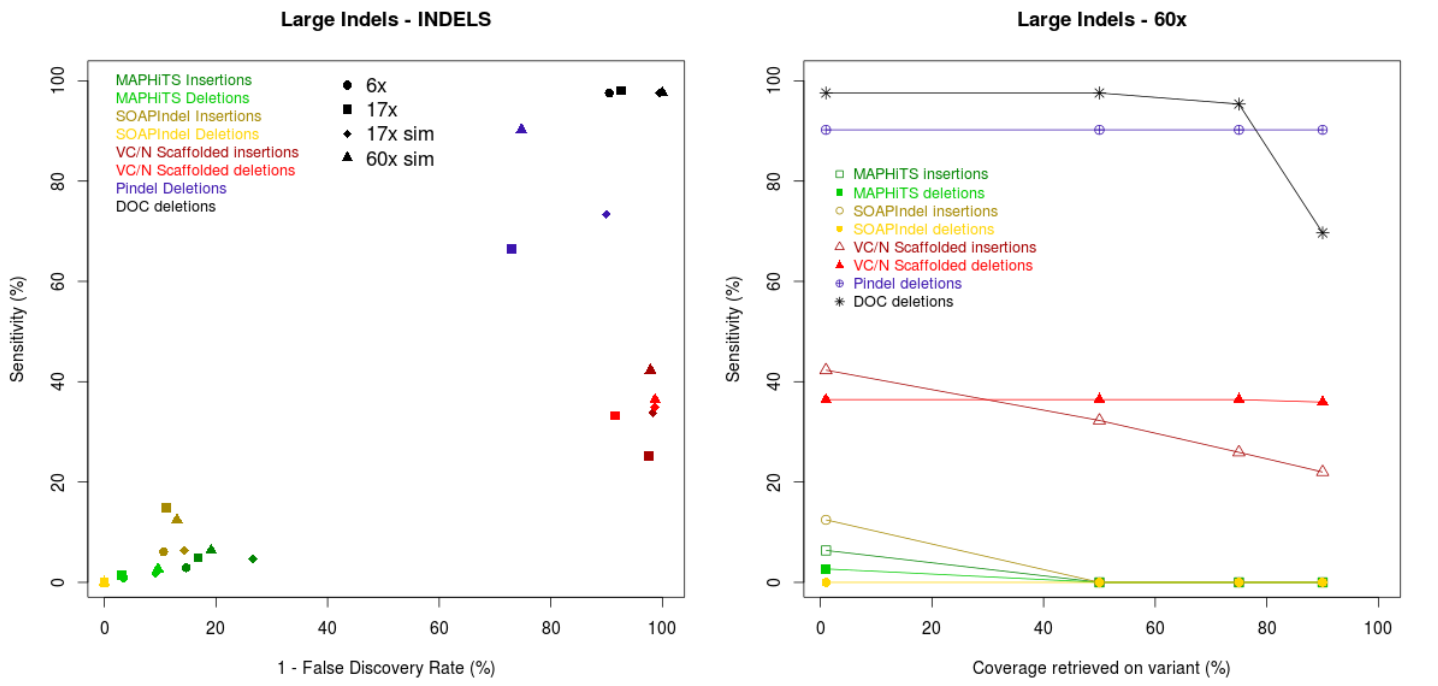

Figure 3: Large Indels detection in the "Large Indels" simulated chromosome

We also implemented a Deph Of Coverage (DOC) approach to detect deletions based on the fall of read coverage expected when a deletion occurs in the sample studied. To reduce false positive rate due to local falls of coverages, only deletions longer than 500bp were kept. This approach gives very high sensitivity and specificity in addition to good enough boundaries retrieval, but only allows the detection of deletions. Nevertheless, it appears to be a promising approach to complement the reference-guided assembly method, especially as the boundaries detection algorithm has not been refined.

Conclusions

In order to find deletions with exact breakpoints and low false positive rate from MP reads, we recommend to use Pindel. SVDetect could also be used, but it needs an experimental validation in order to eliminate the false positives and define exact breakpoints. For insertions, Pindel could also be used as long as they don't exceed few dozen nucleotides, but experimental validation is required. However, it retrieves only a small number of them. Finally, MP data does not seem to help getting long insertions, at least when using SVDetect. However, MP reads could help to scaffold reference-guided assembly.

The PE strategy appears here to be more efficient than MP, and also cheaper. However, it still requires at least a 6x coverage for SNP detection and 17x for other indels. Considering the algorithm used by tools such as Pindel and MAPHiTS, long reads are key to improve the maximum indel size able to be retrieved by these tools. The insert size of MP and PE may also impact greatly the size of the indels that can be accurately detected. Accurate sizing of PE and MP may be important to get solid results.

Large deletions can be detected with a reasonable efficiency with Pindel and the VC/N strategy. Our DOC tool appears to be very efficient for this type of SV. Large insertions are still an issue for all the tools, even if best results are obtained with VC/N. The user needs to refine variant-callers parameters according to his datasets, as default ones are rarely the best. K-mer size for guided assembly has to be determined each time, as it depends on coverage, read size, species, … Biological validation is recommended if FDR > 50% (small and medium size indels). Small FDR would be preferred on high sensitivity when there is cost issue to the validation experiment.

The optimal strategy is probably to combine several tools for detecting different types of SVs, MAPHiTS for SNPs and small indels, Pindel for medium ones, DOC for large deletions and SV/N for large insertions. In addition, when the highest possible specificity is required, predicted SVs cross-validation between tools would be a very efficient strategy.