Bioinformatics Advance Access published November 29, 2014

Bioinformatics, 2014, 1–8
doi: 10.1093/bioinformatics/btu730
Advance Access Publication Date: 7 November 2014
Original Paper

Genome analysis

Ulysses: accurate detection of low-frequency structural variations in large insert-size sequencing libraries

Alexandre Gillet-Markowska[1,2], Hugues Richard[1,2], Gilles Fischer[1,2,]* and Ingrid Lafontaine[1,2]*

[1]Sorbonne Universités, UPMC University Paris 06, UMR 7238, Biologie Computationnelle et Quantitative and [2]CNRS, UMR7238, Laboratory of Computational and Quantitative Biology, F-75005 Paris, France

*To whom correspondence should be addressed.

Associate Editor: Alfonso Valencia

Received on June 17, 2014; revised on October 24, 2014; accepted on October 27, 2014

Abstract

Motivation: The detection of structural variations (SVs) in short-range Paired-End (PE) libraries remains challenging because SV breakpoints can involve large dispersed repeated sequences, or carry inherent complexity, hardly resolvable with classical PE sequencing data. In contrast, large insert-size sequencing libraries (Mate-Pair libraries) provide higher physical coverage of the genome and give access to repeat-containing regions. They can thus theoretically overcome previous limitations as they are becoming routinely accessible. Nevertheless, broad insert size distributions and high rates of chimerical sequences are usually associated to this type of libraries, which makes the accurate annotation of SV challenging.

Results: Here, we present Ulysses, a tool that achieves drastically higher detection accuracy than existing tools, both on simulated and real mate-pair sequencing datasets from the 1000 Human Genome project. Ulysses achieves high specificity over the complete spectrum of variants by assessing, in a principled manner, the statistical significance of each possible variant (duplications, deletions, translocations, insertions and inversions) against an explicit model for the generation of experimental noise. This statistical model proves particularly useful for the detection of low frequency variants. SV detection performed on a large insert Mate-Pair library from a breast cancer sample revealed a high level of somatic duplications in the tumor and, to a lesser extent, in the blood sample as well. Altogether, these results show that Ulysses is a valuable tool for the characterization of somatic mosaicism in human tissues and in cancer genomes.

Availability and implementation: Ulysses is available at http://www.lcqb.upmc.fr/ulysses.

Contact: ingrid.lafontaine@upmc.fr or gilles.fischer@upmc.fr

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Our current understanding of the structural and functional impact of SV onto the biology of genomes has largely benefited from the development of the second generation of DNA sequencing technologies. The computational detection of SV has mainly relied on the development of four methodological strategies, the 'read-depth' method (Campbell *et al.*, 2008; Alkan *et al.*, 2009; Chiang *et al.*, 2009; Yoon *et al.*, 2009; Mills *et al.*, 2011), the 'split-read' method (Lam *et al.*, 2010; Zhang *et al.*, 2011; Jiang *et al.*, 2012), the *de novo* genome assembly (Wang *et al.*, 2011) and the 'Paired-End' method [PEM (Chen *et al.*, 2009; Korbel *et al.*, 2009; Lee *et al.*, 2009; Hormozdiari *et al.*, 2010; Quinlan *et al.*, 2010; Zeitouni *et al.*, 2010; Qi and Zhao, 2011; Marschall *et al.*, 2012; Sindi *et al.*, 2012; Hart *et al.*, 2013)]. Some detection tools have increased SV detection specificity and breakpoint resolution by combining several of these detection strategies (Ye *et al.*, 2009; Medvedev *et al.*, 2010; Abyzov and Gerstein, 2011; Handsaker *et al.*, 2011; Rausch *et al.*, 2012; Yang *et al.*, 2013).

Computational analyses using these approaches demonstrated that large SV are major contributors to the genomic polymorphism between individuals (Conrad and Hurles, 2007; Korbel *et al.*, 2007; Kidd *et al.*, 2008; Mills *et al.*, 2011). Polymorphic SV were shown to contribute to both common diseases and rare genomic disorders and to alter normal gene function during cancer development (Fanciulli *et al.*, 2007; Hollox *et al.*, 2008; Stephens *et al.*, 2009; Pinto *et al.*, 2010; Girirajan *et al.*, 2011). New approaches have also started to reveal the quantitative importance of somatic SV in healthy tissues such as neuron or blood cells (Singer *et al.*, 2010; Laurie *et al.*, 2012; McConnell *et al.*, 2013; Voet *et al.*, 2013).

However, the true level of somatic mosaicism probably remains underestimated, owing to the limitations inherent in classical short-range PE libraries. More SV are now theoretically accessible thanks to the recent development of long-range Mate Pair (MP) libraries in which the two reads can be separated by several kilobases. MP libraries present major advantages over classical PE libraries because large inserts can span over large repeated regions often involved in SV formation and because MP libraries provide, for the same number of reads, a much higher physical coverage of the genome. Higher physical coverage triggers the possibility of uncovering SV that are present at low frequency in mosaic genomes. However, MP libraries involve a ligation step during the library construction which generates a large amount of chimerical Read Pairs (RPs), making those library prone to higher rates of false-positive SV. In addition, MP libraries suffer from wide insert size (IS) distributions, which bring additional noise to the detection of deletion and insertion events. These limitations explain why currently available SV detection tools that were developed for short-range PE libraries perform badly on MP data.

Here, we report a new PEM-based software, called Ulysses, specifically designed to detect SV in MP datasets. Ulysses comprises a SV scoring module, which improves SV detection accuracy in MP libraries. Our algorithm can annotate the full spectrum of SV, including deletions (DEL), segmental duplications (DUP), inversions (INV), small insertions (sINS, with a size smaller than the library IS), large insertions (INS), reciprocal translocations (RTs) and non-reciprocal translocations (NRT). Benchmarks on real MP sequencing datasets from the 1000 Human Genome project, on MP simulated datasets as well as on a breast cancer tumor MP library showed that Ulysses outperforms three commonly used detection tools [Breakdancer (Chen *et al.*, 2009), GASVpro (Sindi *et al.*, 2012) and Delly (Rausch *et al.*, 2012)] for all types of SV and notably for low frequency structural variants in MP libraries. In addition, Ulysses is on par with or outperforms the three other tools on PE datasets, making it a highly versatile detection tool.

2 Methods

2.1 Overview of Ulysses

Ulysses is a PEM-based algorithm, which comprises two independent parts: library parsing (Steps 1–2) and SV detection (Steps 3–5, Fig. 1A). The algorithm automatically tunes parameters for SV detection from the set of statistical properties derived from the library parsing. Then, Ulysses builds simple undirected graphs describing groups of discordant RP that consistently support the existence of the same structural variation (SV). The problem of using cliques to predict SV has already been exactly solved and largely implemented in the past few years (Lee *et al.*, 2008; Hormozdiari *et al.*, 2009; Sindi *et al.*, 2009). Cliques are defined in Ulysses in a way closer to (Rausch *et al.*, 2012). However, because a parameter ensures that all IS within a clique are in a comparable size range (ISc_n, see below), Ulysses adds new constraints on cliques. Note that our clustering rules only reflect our implementation rather than an exact solution to the problem of cliques identification. Finally, Ulysses assesses the statistical significance of each candidate variant in a principled manner using for each type of SV an explicit model for the generation of chimerical RP (Fig. 1A). The five main steps are detailed below and further details can be found in the Supplementary Material.

2.1.1 Step 1—Statistics of the library and detection parameters

Starting from a library alignment file (BAM format), Ulysses derives summary statistics [read-pair orientation, empirical IS distribution (f_l), IS median (μ) and median absolute deviation (σ)] from 1 million of randomly sampled RP. The median and median absolute deviation estimates were preferred over mean and standard deviation as they are more robust to outliers. These values are used to set SV detection parameters described below (d_n, ISc_n, $p_{d_n/\ell_k}(\tau)$ and p_{IS}, Fig. 1B).

Two descriptors, d and ISc, are defined to assess whether two RP are consistent (Fig. 1B). Given two RP of size IS_1 and IS_2, spanning the genomic intervals $[l_1, r_1]$ and $[l_2, r_2]$, respectively, we consider:

– their maximal interdistance: $\text{d} = \max(|l_1 - l_2|, |r_1 - r_2|)$,
– their IS difference: $\text{ISc} = |\text{IS}_1 - \text{IS}_2|$.

Two RP are consistent if they satisfy $\text{d} \leq \mu + n\sigma$ and $\text{ISc} \leq n\sigma$ (note that ISc only applies to intra-chromosomal RP). For further reference, we will name those thresholds d_n and ISc_n (see Section 2.5). In addition, two probabilities, $p_{d_n/\ell_k}(\tau)$ and p_{IS}, are defined to assess the statistical significance of groups of consistent RP (Fig. 1B):

– $p_{d_n/\ell_k}(\tau)$ the probability that τ RP have consistent positions in a chromosome of length ℓ_k,
– $p_{\text{IS}}(\tau)$ the probability that τ RP have consistent IS.

Details about the computation of both probabilities are given in Supplementary Material.

2.1.2 Step 2—Selection of discordant RP

To define SV, Ulysses relies on the identification of discordant RP (with mapping quality ≥20), i.e. RP that map incongruously onto the reference genome (Fig. 1C). RP are considered as discordant when they fulfill at least one of the three following criteria:

– Incongruous RP orientation: any RP not in a [−,+] orientation (throughout the text, read orientations are given for MP and must be reversed for PE libraries).
– IS deviating from the expected range: any RP with an IS outside the range $[\mu - n\sigma, \mu + n\sigma]$.
– Incongruous read location: any RP with the two reads on two different chromosomes.

At the end of the library parsing part (Steps 1 and 2), a set of alignment files containing all identified discordant RP is produced (Fig. 1A). The following SV detection part (Steps 3–5) is independent and can be run separately.

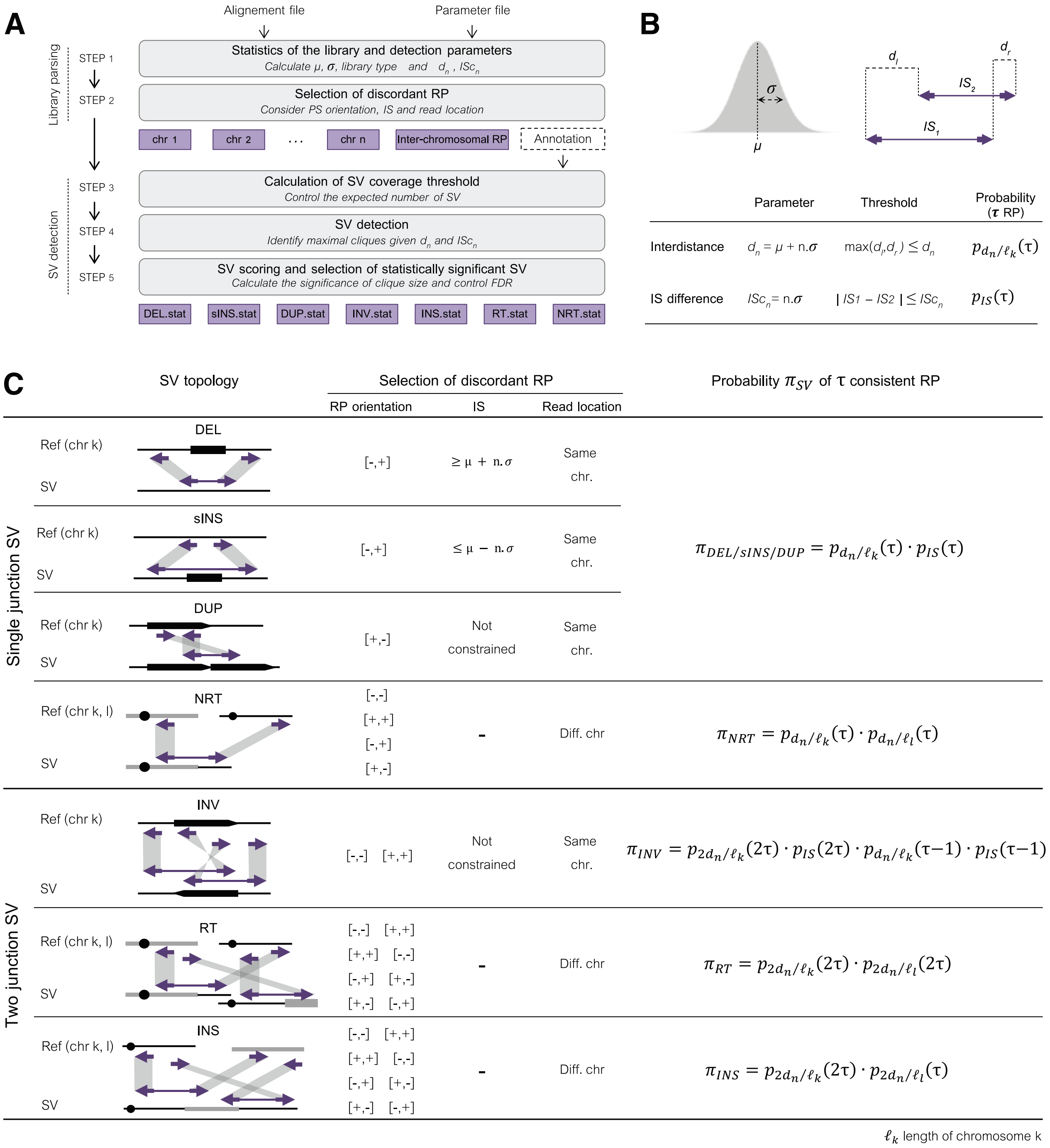

Fig. 1. Ulysses design. (**A**) Flowchart of the Ulysses algorithm indicating all five processing steps (grey) and output files (light purple). The program is composed of two independent parts, the library parsing and the SV detection, comprising two and three steps, respectively. The five steps are detailed in Section 3 (Overview of Ulysses). (**B**) Consistency parameters. Distribution of IS (top left) and schematic representation of two overlapping RP (purple arrows, top right) allow defining the interdistance and the IS difference parameters, thresholds and probabilities of having τ discordant RP (see Supplementary Material). ISc_n is a threshold only applicable to intra-chromosomal RP. (**C**) SV detection characteristics. For each type of SV, the description of discordant RP, including their topology (left panel), properties (middle panel) and probabilities (right panel) is detailed. Mapping read orientations are intended for MP libraries [−,+] and should be reversed for PE libraries. The π_{SV} formula gives the probability for τ discordant RP to be consistent (see the paragraph 'Overview of Ulysses'). *k* and *l* are two different chromosomes of sizes ℓ_k and ℓ_l. Note that π_{SV} can be directly expressed as a product between $p_{d_n/\ell_k}(\tau)$ and $p_{IS}(\tau)$ because the same value of *n* is used for both d_n and ISc_n and therefore there are only two degrees of freedom to fully define the three parameters $|l_1 - l_2|$, $|r_1 - r_2|$ and $|IS_1 - IS_2|$

2.1.3 Step 3—Calculation of SV coverage minimal threshold

The minimal number τ of consistent RP that is required to support each type of SV is set in order to limit the expected number of candidate SV (N_{SV}). This is especially justified when libraries have a wide distribution of ISs and/or a high number of chimerical reads that will generate a high number of false positives (FPs) (as it is often the case with MP libraries).

For each type of SV, if π_{SV} is the probability that x discordant RP are consistent, the number of expected candidate SV supported by x RP is $N_{SV} = \binom{D_{RP}}{x} \cdot \pi_{SV}$ with D_{RP} being the total number of discordant RP of this type. The probability π_{SV} depends on the topology of the SV and can be explicitly formalized using the probabilities $p_{d_n/\ell_k}(\tau)$ and $p_{\mathrm{IS}}(\tau)$ defined above (Fig. 1C). In practice, $\mathrm{N_{SV}}$ is limited to be at most 10 000. The value of τ (default $\tau = 2$) is automatically increased while $\mathrm{N_{SV}}$ is above this limit.

2.1.4 Step 4—SV detection

Discordant RP are categorized into one of the seven SV types reported in Figure 1C. RP are then sorted by chromosomes and coordinates, and used to build a simple undirected graph. Each discordant RP represents one vertex of the graph and one edge is drawn between each pair of consistent RP. Ulysses defines a SV as a maximal clique in the graph where all consistent RP are connected to each others. Note that the same vertex can belong to two or more different maximal cliques (meaning that a given RP can be used to define different SV).

DEL, sINS and DUP are SV that produce a single new DNA junction in the rearranged chromosome when compared with the reference genome. For those single-junction SV, we directly use the corresponding maximal cliques to identify candidates SV. Each INV, INS or RT produces two new DNA junctions when compared with the reference genome. These two-junction SV are consequently detected by two maximal cliques, which, in Ulysses, need to be interconnected (by RP orientation and relative coordinates, Fig. 1C and Supplementary Material). This requirement increases the specificity of the detection in comparison with methods that consider only one junction, even for the detection of two-junction SV. All combinations of compatible pairs of cliques are conserved, thus allowing the same RP to be re-used in several different pairs. In addition, when the position of the centromeres is provided, the relative orientation of the two cliques is checked and an RT event will be reported only when both of the rearranged molecules contain a single centromere. Otherwise, the corresponding SV is classified as an INS. For both INS and RT, the number of RP must be equally distributed between the two cliques (ratio > 0.1). Otherwise, the corresponding SV is classified as a NRT.

2.1.5 Step 5—SV scoring and selection of statistically significant SV

Ulysses evaluates whether the number m of RP that supports each candidate variation is significant. This step is essential to filter out false-positive predictions from the artefacts in the library. The significance level of each candidate is then corrected by controlling the false-discovery rate (FDR).

Deletions and small insertions: The statistical significance of each candidate deletion is estimated by calculating the probability b_m to randomly sample at least m RP identifying this deletion, given the local physical coverage C. Given the smallest RP describing the deletion with an IS s, each RP drawn from the IS distribution has a probability $\mathrm{prop_{IS}}$ to be consistent with the smallest RP ($\mathrm{prop_{IS}} = \sum_{l\geq s} f_l$, see Supplementary Material). The probability b_m will result from the binomial sampling of the RP:

$$b_m = \sum_{i\geq m} \binom{\mathrm{C}}{i} \mathrm{prop}_{\mathrm{IS}}^{i} \cdot (1 - \mathrm{prop}_{\mathrm{IS}})^{\mathrm{C}-i}.$$

In practice, C is estimated over the region corresponding to the DEL position (considering at most the 10 first kilobases of the deleted region). The significance of small insertions can be evaluated in the same manner, by flipping around the IS distribution.

Segmental duplications, non-reciprocal translocations and two-junction SV: For those SV, the statistical significance is computed with a binomial distribution as being the probability ν_m that at least m RP among all combinations of discordant RP of each type are consistent, given the probability π_{SV} (see Fig. 1A):

$$\nu_m = \sum_{i\geq m} \binom{C_m}{i} \pi_{\mathrm{SV}}^{i} \cdot (1 - \pi_{\mathrm{SV}})^{C_m - i},$$

where $C_m = \binom{D_{\mathrm{RP}}}{m}$ is the number of ways of having m RP among a total of D_{RP} discordant ones on the SV considered (see Step 3 and Fig. 1C).

ν_m is calculated for each chromosome independently (or each pair of chromosomes for inter-chromosomal rearrangements).

After SV scoring, we set up a P-value cut-off by controlling the FDR (default value 0.01). Q-values are estimated using a bootstrap method (Storey *et al.*, 2004). During our tests with experimental and simulated data, this approach greatly improved the specificity of the detection (see Section 3).

2.2 Sequencing datasets

The details and the accession numbers for all real sequencing datasets used in this study (NA12878 and breast cancer BT71) are provided in Supplementary Methods.

For simulated sequencing datasets, a MP ($\mu = 2000$ bp, $\sigma = 1487$ bp, normal distribution, read-length = 50 bp) and a PE ($\mu = 233$ bp, $\sigma = 10$ bp, normal distribution, read-length = 50 bp) Illumina-like datasets were generated with wgsim 0.3.1–*r*13 (Li *et al.*, 2009) with default parameters using a 162 Mb human genomic region as a reference (GRCh37/hg19 chromosomes 20, 21 and 22) at three different levels of coverage (10×, 30× and 60×). The MP dataset was generated with variable proportions of random chimerical RP (0.1%, 1%, 2.5% and 5%) in order to sample different levels of experimental artefacts that can derive from the circularization step during MP library construction. Chimerical RP were generated by randomly sampling the RP and by shuffling the mates. As a consequence, chimerical RP can have any orientations and be either intra or inter-chromosomal. The simulated reads were remapped on the reference genome using BWA aln 0.7.3*a*–*r*367 with default parameters (Li *et al.*, 2009). Artificial SV were then added to both PE and MP simulated library datasets after the remapping step in order to avoid mapping artefacts that could bias SV detection. Artificial SV were designed with variable numbers of discordant RP (with 4 and 8 RP in the 10× dataset; with 4, 8, 16 and 32 RP in the 30× dataset and with 4, 8, 16, 32 and 64 RP in the 60× dataset). These numbers correspond to a relative coverage varying from 0.06 (4 RP in the 60× library dataset: 4/(60 + 4) = 0.06) up to 0.52 (64 RP in 60× dataset: 64/(60 + 64) = 0.52). Each set of artificial SV comprises 50 SV of each type (DUP, DEL, INV, INS, RT and NRT) for each relative coverage, generating a total of 600, 1200 and 1500 SV for 10×, 30× and 60× datasets, respectively. DUP, DEL and INV were generated with random sizes, varying from 1 to 50 kb. The detection of one SV was considered as true positive (TP) when at least one of the RP that defined the simulated SV was recovered. All simulated datasets are available at http://www.lcqb.upmc.fr/ulysses/simulations.

2.3 Benchmark with other detection tools

All benchmarks analyses were performed using the AMADEA Biopack platform developed by ISoft (http://www.isoft.fr/bio/biopack_en.htm). Breakdancer v1.2.6 (Chen *et al.*, 2009), GASVpro-HQ v1.2 (Sindi *et al.*, 2012) (referred to as GASVpro) and Delly v0.5.6 (Rausch *et al.*, 2012) were configured to detect SV defined by at least two consistent RP (which is the lowest limit to define a clique). All other parameters remained set by default. The sensitivity (Sn) of a detection tool is the proportion of true SV that are detected ([TPs]/[TPs + false negatives]) and its precision (or positive predictive value, PPV) is the proportion of true SV among all detected SV ([TPs]/[TPs + FPs]). The F1-score is an estimator of the trade-off between Sn and PPV ($2 \times \text{Sn} \times \text{PPV}/(\text{Sn} + \text{PPV})$). Supplementary Tables S1, S2 and S3 report Sn, PPV, F1-scores and running times of Ulysses, Delly, Breakdancer and GASVpro on all simulated datasets.

Note that GASVpro cannot perform SV detection for MP datasets (both simulated and experimental). The execution of GASVpro on these data was stopped after more than 166 h of computation on a single CPU after splitting the sequence files by chromosomes (core i7 870, 32GB RAM). As a consequence, GASVpro does not appear in any of the MP analyses.

2.4 Ulysses detection parameter *n*

The performance of Ulysses was tested as a function of the n value, which controls the main detection parameters for all types of SV (the interdistance d_n and IS difference ISc_n, see Section 3). For both MP and PE simulated datasets, F1-scores remain highly stable for n values ranging from 3 to 10 (Supplementary Fig. S1). For real sequencing dataset (NA12878 MP 30×), a n value set to 6 provides a good compromise between Sn and PPV (Supplementary Fig. S2). This value of $n = 6$, set by default in the parameter file, is therefore well suited for most applications and does not require any manual adjustment by the user.

3 Results

3.1 Performance of Ulysses on MP sequencing datasets

The performance of Ulysses was compared with three other widely used SV detection tools: Breakdancer, GASVpro and Delly. Given that BreakDancer and Delly do not discriminate between INS, RT and NRT, these three types of SV were merged into a single class of events called inter-chromosomal (INTER) for analysis. The performance of each method was estimated with PPV, representing the total number of TPs divided by the total number of detected SV.

3.1.1 Performance on real sequencing data from the 1000 Human Genome Project

The three tools were benchmarked on a real MP sequencing dataset (30× coverage) coming from a single individual (NA12878, see Section 2). We focused on DEL detection because a robust GS containing 2209 DEL was available (see Section 2). Results are presented only for Ulysses, Delly and BreakDancer (SV detection could not finish in reasonable time with GASVpro on MP datasets, even with sequences split by chromosome, see Section 2).

The results on Figure 2A are presented as ROC curve showing the number of TPs DEL as a function of the total number of predicted DEL (TP + FP). The main interest of our approach lies in the scoring method that allows to filter out a large number of FP cliques that directly result from chimerical RP. The scoring module filters out 99.88% of the 166 057 (167 335 − 1278) FP detected. It also filters out 86.31% of the 1278 TP, keeping as statistically significant 175 TP after filtering. Note that the final sensitivity of Ulysses after the scoring step (175 TP) is very close to that of BreakDancer (197 TP) and higher than that of Delly (48 TP). As a result, the scoring method strongly increases Ulysses PPV from 0.76% to 45% (a 59-fold increase), which explains the massive gain in specificity of Ulysses over the other tools (Fig. 2A, Supplementary Table S1).

We have checked that the higher detection accuracy of Ulysses did not result from a biased distribution of DEL sizes in the GS that could have favoured Ulysses over the other tools. The median deletion size of DEL from the GS is 4.9 kb, whereas the median deletion sizes of DEL detected by Ulysses, Delly and Breakdancer are of 13.2, 21.5 and 5.5 kb, respectively. Thus, Ulysses and Delly do not detect small DEL from the GS. BreakDancer is able to predict such small DEL at the cost of precision, with the concomitant detection of about 21 000 FPs.

In order to further characterize Ulysses higher detection accuracy, we plotted the relative precision (proportion of TP DEL or PPV) and the FDR (proportion of FP DEL) of the three tools as a function of DEL physical coverage (Fig. 2B). This plot shows that Ulysses precision remains relatively constant (between 0.24 and 0.57) over the whole range of coverages whereas BreakDancer precision increases with increasing coverage (from 0.0024 to 0.54). Delly achieves correct precision only for intermediate coverage values. Compared with the other tools, Ulysses precision is particularly higher for the lowest physical coverage values (between 1 and 20 RP). These results show that Ulysses is particularly efficient for the detection of low frequency SV.

3.1.2 Detection of somatic mosaicism in tumour sample

We used 8 kb MP data from a luminal A breast tumor to analyse the somatic mosaicism present in cancer cells (data taken from Inaki et al. (2014)). In this paper, the authors suggest that tandem duplications appear to be early events in tumour evolution, especially in the genesis of amplicons. We performed the differential detection of tandem DUP in the blood and the tumour samples with Ulysses and Delly (Supplementary Fig. S4). Nearly all DUP detected in the blood sample with allele frequencies higher than 0.2 were also found in the tumour, as expected for germline SV. Note that Uysses detected 49 germline DUP whereas Delly only found 5 such DUP. We found nearly no evidence of high allele frequency (AF >0.2) somatic DUP in the blood sample whereas in the tumor sample, we detected a significant number of somatic DUP that probably occur early during tumour development as suggested by their AF higher than 0.2 (28 with Ulysses versus 3 with Delly, Supplementary Fig. S4A and B, respectively). The highest level of somatic mosaicism was found by Ulysses as low frequency DUP (AF <0.2) in the tumour sample. Ulysses detected 4551 low frequency DUP whereas Delly found only 256 such DUP (Supplementary Fig. 4A and B, respectively). In this analysis, somatic mosaicism was also detected in the blood sample, as already reported (Jacobs *et al.*, 2012; Laurie *et al.*, 2012), but to a lesser extent than in the tumour sample (Supplementary Fig. S4).

3.1.3 Performance on simulated SV

A MP sequencing dataset with wide IS distributions ($\mu = 2$ kb, $\sigma = 1487$ bp) was simulated at different sequencing coverage values (10×, 30× and 60×, see Section 2), using three human chromosomes as reference. Fifty simulated SV of each type (DUP, DEL, INV, INS, RT and NRT) were added to the libraries at relative physical coverages varying from 0.06 up to 0.52. Intra-chromosomal SVs (DUP, DEL and INV) were generated with random sizes, varying from 1 kb to 50 kb. Note that simulated SV represent the TP SV in this dataset. In addition, variable proportions of random chimerical RP (from 0.1% to 5% of the reads) were added to the library in order to simulate the typical experimental noise that derives from the circularization step during MP library construction.

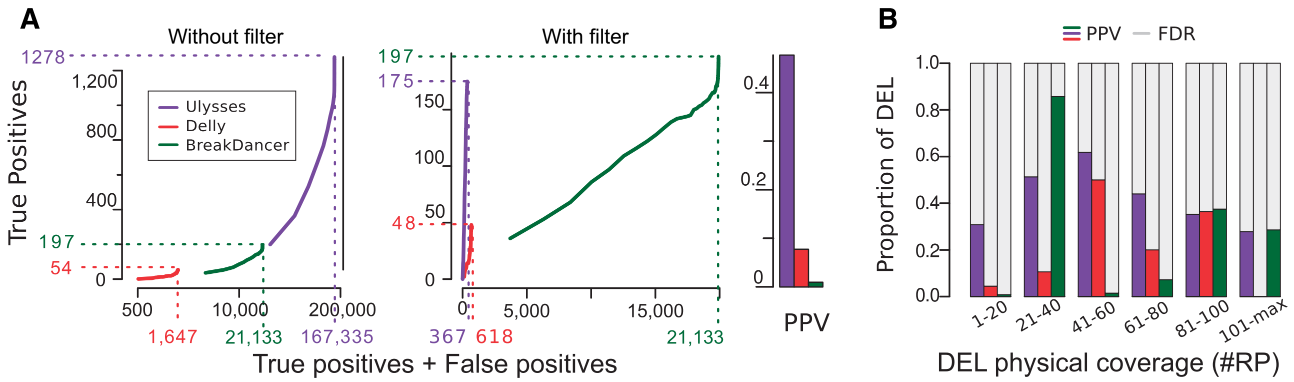

Fig. 2. Deletions in the NA12878 MP 30X dataset. (**A**) Left and middle panels: ROC-like curves representing the number of TP DEL, without and with statistical filter (see Gold Standard in Section 2) as a function of the total number of predictions (TPs + FPs) using the relative SV coverage as a cumulative varying threshold. GASVpro could not be run for the MP library (see text). Right panel: detection accuracy with filter represented as PPV. (**B**) Relative precision (proportion of true-positive DEL) and FDR (proportion of false-positive DEL) of the three tools as a function of DEL physical coverage in MP30X NA12878 dataset. Blueviolet, red and dark green bars represent the precision (PPV) of Ulysses, Delly and Breakdancer, respectively. Grey bars represent the proportion of false-positive DEL (FDR, 1-PPV)

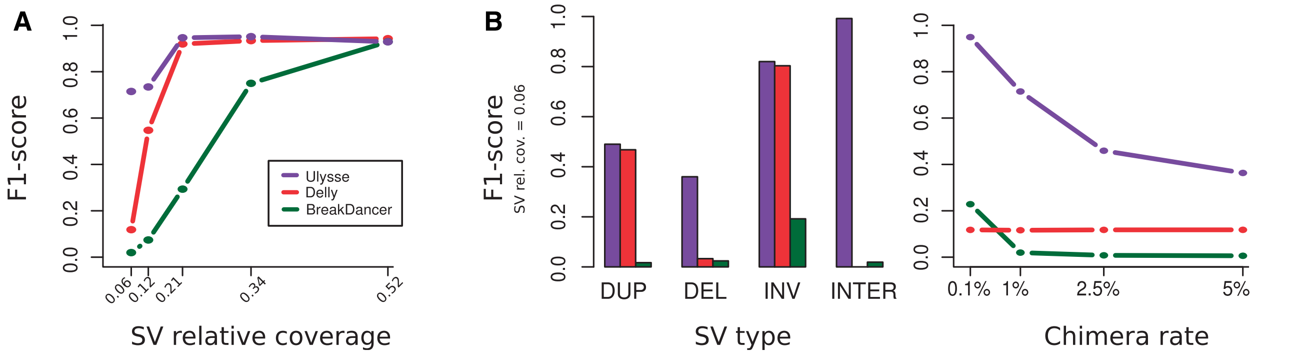

Fig. 3. Results on a 60X MP dataset with RP distribution $\mu = 2$ kb, $\sigma = 1487$ bp and 1% of chimerical RP. (**A**) F1-score as a function of SV relative coverage. Note that GASVpro is absent from the MP simulations because the execution of the program on this dataset exceeded 166 h of computation (see Supplementary Table S2). (**B**) Left panel: F1-score for the different types of SV with a relative coverage of 0.06. The INTER class of SV gathers results from INS, RT and NRT. Right panel: F1-score value are represented as a function of the proportion of chimerical RP present in the dataset (SV relative coverage = 0.06)

No major difference was found across the range of sequencing coverage values. Therefore, the results, presented on Figure 3, only focus on the 60× coverage dataset (see Supplementary Figs. S5–S9 and Supplementary Tables S2 for 10× and 30× results). We first assessed the performance of the three algorithms across variable SV physical coverage values (Fig. 3A). For SV present at high relative coverage (0.52, corresponding to a heterozygous SV in a diploid cell), all three algorithms perform globally well, with similar accuracy as estimated by the F1-scores. However, as soon as the SV relative coverage decreases, the performance of Delly and BreakDancer rapidly deteriorates (Fig. 3A). At the lowest SV relative coverage (0.06), Ulysses retains a detection accuracy of 71% whereas Delly and BreakDancer drop down to 11% and 2%, respectively. We found that detection accuracy of Delly and BreakDancer deteriorates because of high over-prediction of FP SV (see ROC curves in Supplementary Figs. S5–S9). These results show that Ulysses is the only algorithm able to detect low frequency SV (i.e. physical coverage between 0.06 and 0.34) with high accuracy.

Next, we compared the three tools for each type of SV, across all physical coverage values. Again, for the highest relative coverage (0.52), all tools achieve comparable accuracies on all types of SV (Supplementary Fig. S10). For all SV types, the performance of Ulysses compared with the other tools gradually improves with decreasing relative coverage. For the lowest relative coverage value, Ulysses outperforms both Delly and BreakDancer for all types of SV (Fig. 3B left panel).

Because MP libraries often comprise high proportions of chimerical RP, we also tested whether the conclusions drawn here for a dataset containing 1% of chimerical RP also apply to datasets with other levels of chimeras (between 0.1% and 5%). Again, for the lowest SV coverage value (0.06), Ulysses shows the highest F1-scores over the entire range of chimera (Fig. 3B right panel). For relative coverage values from 0.12 to 0.34, the difference between the tools gradually reduces and totally vanishes for the highest coverage (0.52, Supplementary Fig. S10).

3.2 Performance on PE libraries

We also tested Ulysses performance on classical PE libraries, with smaller ISs, and compared it with Delly, BreakDancer and GASVpro. For real sequencing dataset (NA12878), Ulysses and BreakDancer achieve the best detection accuracies at low coverage (5×). For a high coverage 200× dataset, Ulysses clearly outperforms all tools (Supplementary Fig. S11A). Noticeably, GASVpro, which also includes a SV scoring and filtering module, performs very poorly on low coverage data but reached the second best PPV on the high coverage data. For simulated dataset, RP were generated with characteristics similar to those described above for MP data (in terms of sequencing coverage, SV type and physical coverage, see Section 2). Both Ulysses and BreakDancer behave well at all SV relative coverages and across all types of SV whereas Delly and GASVpro show lower F1-scores for low frequency DEL (Supplementary Fig. S11B).

4 Discussion

Ulysses uses a PEM approach that relies on the identification of groups of discordant RP to detect the full spectrum of SV. This strategy, also implemented in the three other detection tools tested here, is highly sensitive but usually lacks specificity when used alone on MP data with wide IS distribution and high proportion of chimerical RP. To overcome these limitations, we developed in Ulysses a scoring module that statistically assesses the genuineness of all candidates SV, given an explicit model for the generation of chimerical RP. To deal with wide IS distribution, Ulysses evaluates IS consistency between RP and filters out groups of RP with inconsistent IS. To deal with high proportions of chimerical RP (up to 5%), Ulysses uses statistics based on the relative coverage of candidate SV. These two parameters automatically adjust to the characteristics of the library such that no manual tuning is required to achieve detection with good accuracy across all types of SV. As a result, Ulysses is the only tool that performs equally well on MP and PE data. On a real MP sequencing dataset from the 1000 Human Genome Project (NA12878, 30×), Ulysses achieves a better sensitivity than the other tools and by several orders of magnitude the best precision. Ulysses also achieves the highest detection accuracy on PE datasets, showing that globally, Ulysses outperforms the other tools no matter the type of sequencing library.

Ulysses scoring module brings a major benefit by enabling accurate detection of low coverage variants. Low coverage SV can correspond to rearrangements occurring in genomic regions which are difficult to sequence and where the local coverage could dramatically drop (some regions were shown to be consistently prone to low coverage). Alternatively, low coverage SV could also result from rearrangements only present in a small subset of the cell population. This is of particular interest when analyzing samples that contain polymorphic somatic mutations such as cancer samples. The analysis of somatic mosaicism in a breast tumour sample revealed that Ulysses achieves an efficient detection of both germline and somatic DUP. We also showed that Ulysses is able to detect somatic mosaicism in a blood sample. Furthermore, recent insights onto somatic mosaicism showed that subclonal cell heterogeneity is not restricted to cancer cells and could be common between cells from a single tissue sample. Given such an unsuspected level of somatic genome plasticity, the availability of a SV detection tool like Ulysses, with high accuracy for rare SV is of primary importance.

Acknowledgements

The authors thank our colleagues from LCQB for fruitful discussions and Jean-Philippe Meyniel from ISoft for invaluable tips and advices for pipeline developments in AMADEA.

Funding

This work was supported by the Agence Nationale pour la Recherche grant 2010 BLAN1606 and by an ATIP grant from the Centre National de la Recherche Scientifique (CNRS). H.R. was partly supported by a Japan Society for the Promotion of Science (JSPS) fellowship [PE11014].

Conflict of interest: none declared.

References

Abyzov,A. and Gerstein,M. (2011) AGE: defining breakpoints of genomic structural variants at single-nucleotide resolution, through optimal alignments with gap excision. *Bioinformatics (Oxford, England)*, **27**, 595–603.

Alkan,C. *et al.* (2009) Personalized copy number and segmental duplication maps using next-generation sequencing. *Nat. Genet.*, **41**, 1061–1067.

Campbell,P.J. *et al.* (2008) Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. *Nat. Genet.*, **40**, 722–729.

Chen,K. *et al.* (2009) Breakdancer: an algorithm for high-resolution mapping of genomic structural variation. *Nat. Methods*, **6**, 677–681.

Chiang,D.Y. *et al.* (2009) High-resolution mapping of copy-number alterations with massively parallel sequencing. *Nat. Methods* **6**, 99–103.

Conrad,D.F. and Hurles,M.E. (2007) The population genetics of structural variation. *Nat. Genet.*, **39**(Suppl. 7), S30–S36.

Fanciulli,M. *et al.* (2007) Fcgr3b copy number variation is associated with susceptibility to systemic, but not organ-specific, autoimmunity. *Nat. Genet.*, **39**, 721–723.

Girirajan,S. *et al.* (2011) Relative burden of large CNVs on a range of neurodevelopmental phenotypes. *PLoS Genet.*, **7**, e1002334

Handsaker,R.E.*et al.* (2011) Discovery and genotyping of genome structural polymorphism by sequencing on a population scale. *Nat. Genet.*, **43**, 269–276.

Hart,S.N. *et al.* (2013) SoftSearch: integration of multiple sequence features to identify breakpoints of structural variations. *PLoS One*, **8**, e83356.

Hollox,E.J. *et al.* (2008) Psoriasis is associated with increased beta-defensin genomic copy number. *Nat. Genet.*, **40**, 23–25.

Hormozdiari,F. *et al.* (2009) Combinatorial algorithms for structural variation detection in high-throughput sequenced genomes. *Genome Res.*, **19**, 1270–1278.

Hormozdiari,F. *et al.* (2010) Next-generation VariationHunter: combinatorial algorithms for transposon insertion discovery. *Bioinformatics (Oxford, England)*, **26**, i350–i357.

Inaki,K. *et al.* (2014) Systems consequences of amplicon formation in human breast cancer. *Genome Res.*, **11**, 1–13.

Jacobs,K.B. *et al.* (2012) Detectable clonal mosaicism and its relationship to aging and cancer. *Nat. Genet.*, **44**, 651–658.

Jiang,Y. *et al.* (2012) PRISM: pair-read informed split-read mapping for base-pair level detection of insertion, deletion and structural variants. *Bioinformatics (Oxford, England)*, **28**, 2576–2583.

Kidd,J.M. *et al.* (2008) Mapping and sequencing of structural variation from eight human genomes. *Nature*, **453**, 56–64.

Korbel,J.O. *et al.* (2007) Paired-end mapping reveals extensive structural variation in the human genome. *Science*, **318**, 420–426.

Korbel,J.O. *et al.* (2009) PEMer: a computational framework with simulation-based error models for inferring genomic structural variants from massive paired-end sequencing data. *Genome Biol.*, **10**, R23.

Lam,H.Y.K. *et al.* (2010) Nucleotide-resolution analysis of structural variants using BreakSeq and a breakpoint library. *Nat. Biotechnol.*, **28**, 47–55.

Laurie,C.C. *et al.* (2012) Detectable clonal mosaicism from birth to old age and its relationship to cancer. *Nat. Genet.*, **44**, 642–650.

Lee,S. *et al.* (2008) A robust framework for detecting structural variations in a genome. *Bioinformatics (Oxford, England)*, **24**, i59–i67.

Lee,S. *et al.* (2009) MoDIL: detecting small indels from clone-end sequencing with mixtures of distributions. *Nat. Publ. Group*, **6**, 473–474.

Li,H. *et al.* (2009) The sequence alignment/map format and samtools. *Bioinformatics*, **25**, 2078–2079.

Marschall,T. *et al.* (2012) Clever: clique-enumerating variant finder. *Bioinformatics*, **28**, 2875–2882.

McConnell,M.J. *et al.* (2013) Mosaic copy number variation in human neurons. *Science*, **342**, 632–637.

Medvedev,P. *et al.* (2010) Detecting copy number variation with mated short reads. *Genome Res.*, **20**, 1613–1622.

Mills,R.E. *et al.* (2011) Mapping copy number variation by population-scale genome sequencing. *Nature*, **470**, 59–65.

Pinto,D. *et al.* (2010) Functional impact of global rare copy number variation in autism spectrum disorders. *Nature*, **466**, 368–372.

Qi,J. and Zhao,F. (2011) inGAP-sv: a novel scheme to identify and visualize structural variation from paired end mapping data. *Nucleic Acids Res.*, **39**(Web Server issue), W567–W575.

Quinlan,A.R. *et al.* (2010) Genome-wide mapping and assembly of structural variant breakpoints in the mouse genome. *Genome Res.*, **20**, 623–635.

Rausch,T. *et al.* (2012) Delly: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics*, **28**, I333–I339.

Sindi,S. *et al.* (2009) A geometric approach for classification and comparison of structural variants. *Bioinformatics (Oxford, England)*, **25**, i222–i230.

Sindi,S.S. *et al.* (2012) An integrative probabilistic model for identification of structural variation in sequencing data. Genome Biol., **13**, R22.

Singer,T. *et al.* (2010) Line-1 retrotransposons: mediators of somatic variation in neuronal genomes? Trends Neurosci., **33**, 345–354.

Stephens,P.J. *et al.* (2009) Complex landscapes of somatic rearrangement in human breast cancer genomes. *Nature*, **462**, 1005–1010.

Storey,J.D. *et al.* (2004) Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *J. R. Stat. Soc. Ser. B Statist. Methodol.*, **66**, 187–205.

Voet,T. *et al.* (2013) Single-cell paired-end genome sequencing reveals structural variation per cell cycle. *Nucleic Acids Res.*, **41**, 6119–6138.

Wang,J. *et al.* (2011) CREST maps somatic structural variation in cancer genomes with base-pair resolution. *Nat. Methods*, **8**, 652–654.

Yang,L.X. *et al.* (2013) Diverse mechanisms of somatic structural variations in human cancer genomes. *Cell*, **153**, 919–929.

Ye,K. *et al.* (2009) Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics (Oxford, England)*, **25**, 2865–2871.

Yoon,S. *et al.* (2009) Sensitive and accurate detection of copy number variants using read depth of coverage. *Genome Res.*, **19**, 1586–1592.

Zeitouni,B. *et al.* (2010) SVDetect: a tool to identify genomic structural variations from paired-end and mate-pair sequencing data. *Bioinformatics (Oxford, England)*, **26**, 1895–1896.

Zhang,Z.D. *et al.* (2011) Identification of genomic indels and structural variations using split reads. *BMC Genomics*, **12**, 375.