Published online 25 June 2012 *Nucleic Acids Research, 2012, Vol. 40, No. 18* ***e145***
doi:10.1093/nar/gks606

SVM^2: an improved paired-end-based tool for the detection of small genomic structural variations using high-throughput single-genome resequencing data

Matteo Chiara[1,*], Graziano Pesole[2,3,4] and David S. Horner[1,*]

[1]Department of Biomolecular Sciences and Biotechnology, University of Milan, Milan 20133, [2]Institute of Biomembranes and Bioenergetics, National Research Council, [3]Department of Biosciences, Biotechnology and Pharmacological Sciences, University of Bari and [4]Center of Excellence in Genomics (CEGBA), Bari 70125, Italy

Received December 22, 2011; Revised May 29, 2012; Accepted May 30, 2012

ABSTRACT

Several bioinformatics methods have been proposed for the detection and characterization of genomic structural variation (SV) from ultra high-throughput genome resequencing data. Recent surveys show that comprehensive detection of SV events of different types between an individual resequenced genome and a reference sequence is best achieved through the combination of methods based on different principles (split mapping, re-assembly, read depth, insert size, etc.). The improvement of individual predictors is thus an important objective. In this study, we propose a new method that combines deviations from expected library insert sizes and additional information from local patterns of read mapping and uses supervised learning to predict the position and nature of structural variants. We show that our approach provides greatly increased sensitivity with respect to other tools based on paired end read mapping at no cost in specificity, and it makes reliable predictions of very short insertions and deletions in repetitive and low-complexity genomic contexts that can confound tools based on split mapping of reads.

INTRODUCTION

The characterization of intra-specific genomic diversity has enormous implications for biomedical sciences and for biology in general and is one of the principal objectives of contemporary genomics. Recently, ultra high-throughput next-generation sequencing [NGS (1)] technologies have greatly facilitated ambitious genome resequencing projects and associated studies focused on human health [e.g. http://www.1000genomes.org/ (2)] and on generating a wider understanding of genome evolution (3,4).

One of the most interesting general conclusions to emerge from such studies is that, contrary to long-held assumptions, structural variations (SVs)—genomic re-arrangements, including insertions, deletions, copy number variations and inversions—typically explain a very significant proportion of normal intra-specific genetic variation (5–9). Although the widespread association of SV with hereditary diseases and cancer (10–17) justifies their study, the use of SVs as molecular markers in non-human systems, for genome-wide association studies, genetic mapping and marker-assisted breeding approaches is also increasing.

Bioinformatics tools to detect SV with high-throughput resequencing data tend to be specialized to accommodate specific types of data or rely on different expected patterns of mapping of reads from a resequenced (donor) genome on a reference sequence in the vicinity of SVs. For example, in the context of the 1000 genomes project (2), mixed samples of genomic DNA from multiple individuals have often been sequenced together as part of an effort to generate a comprehensive catalog of variants and haplotypes in human populations. Dedicated and highly sophisticated tools that use probabilistic methods to identify variations that are not present in all sequenced individuals have been developed and shown to be highly effective (18,19).

Tools developed to detect SVs from high-coverage individual genome resequencing may be categorized as alignment based or statistics based. Approaches dependent on

*To whom correspondence should be addressed. Tel: +39 02503 14884; Fax: +39 50315044; Email: david.horner@unimi.it
Correspondence may also be addressed to Matteo Chiara. Tel: +39 02503 14923; Fax: +39 50315044; Email: matteo.chiara@unimi.it

the alignment of reads to a reference sequence may include partial *de novo* assembly of reads (20) or may rely on split mapping of short reads (21). Although such methods should be capable of precisely identifying break points, difficulties in *de novo* sequence assembly, the presence of sequencing errors, limits on the maximum detectable size for insertions and the impact of repetitive genome sequences mean that they are incapable of identifying all SV events [reviewed in (22,23)].

Statistics-based methods include read density-based approaches that exploit the same principle as DNA hybridization arrays. These tools are particularly efficient in detecting copy number variation but cannot easily identify the introduction of novel sequences (24,25). Paired end (PE) read-based approaches are particularly suited for identifying insertion and deletions. Such methods aim to identify genomic loci where donor reads map at inconsistent distances. A number of tools based on this principle have been developed and either detect genomic loci exhibiting statistically significant clustering of PE reads with anomalous mapping distances (26,27) or compare local distributions of mapping distances to an expected distribution in an attempt to identify regions harboring SVs (26,28). The first approach is more suited to the identification of long deletions, whereas the second tends to be more computationally intensive but generally applicable. One obvious disadvantage of statistics-based methods is that they do not identify precise break points.

Most older approaches to detect SV from resequencing data are based, essentially, on one of the aforementioned metrics. However, several recent publications attempt to incorporate multiple types of information. For example, Breakpointer (29) uses specific coverage measurements to identify regions potentially harboring SV and split-mapping data to provide additional support and to fine map break points, whereas GASVPro (30) directly incorporates both read depth and mapping distance information to increase the specificity of prediction of large deletions and inversions. Indeed, different types of genomic rearrangements, even those involving only a few base pairs of DNA, are expected to generate complex and particular signatures in mapping patterns of PE reads. Additionally, each sequencing reaction and reference genome has a series of characteristics which can, in principal, impact on methods used to identify SV. For example, each library has a characteristic insert-size distribution, and each sequencing reaction shows a particular profile and frequency of sequencing errors. Furthermore, individual reference genome sequences show a particular distribution of repetitive sequences. All these factors are relevant to the selection/parametrization of appropriate statistical tests for the identification of SVs from insert-size perturbations.

Support vector machines (SVMs) are an ensemble of statistics/computational techniques that have been widely used in biological classification problems including the recognition of micro-RNA precursors, the discrimination of coding from non-coding sequences, the classification of differential gene expression profiles from microarray data, the recognition of protein secondary structure and the identification of candidate drug targets. SVM uses a series of training data points, each known to belong to one of two (or more) classes of origin and described by a number of quantitative features, and, having transformed them into a higher dimensionality than allowed by the number of associated features and through the use of a kernel function, identifies the hyperplane that maximizes their separation by class in a multidimensional space. Once the optimal discriminating function has been established, it is used to classify unknown instances [for an introductory review see (31)]. Several software libraries implementing SVM are freely available, and the method can be adapted to function in multiple category classification problems.

In this study, we show that the incorporation of different characteristics of mapping data derived from PE resequencing reads can improve the sensitivity of detection of relatively small indels (1–30 bp) that constitute the majority of intra-specific SV events (32). We use SVMs to incorporate these diverse mapping characteristics to address the indel finding/classification problem. The SV mapping using SVMs (SVM^2) software presented herein calculates and integrates a combination of features based on statistics and resequencing coverage measures for windows around a given genomic coordinate. The method does not make *a priori* assumptions regarding the insert-size distribution of a particular library or on the optimal *P* value cutoff to be used in any of the statistical tests that it uses, rather, it is trained using a given resequencing data set and reference genome sequence. In this work, SVM^2 was trained to discriminate genomic loci flanking four classes of events (deletions, insertions shorter than the library insert size, insertions longer than the library insert size and hypervariable regions) from normal genomic regions, although in principal there is no restriction to the number of classes/sizes of events that could be recognized.

SVM^2 attains a similar specificity and a far superior sensitivity than state of the art PE-based methods using the same data and seems to be more robust than conventional split mapping to the confounding effects of some genomic contexts.

Recent surveys confirm that comprehensive detection of SV events of different types between donor and reference sequences is best achieved through the combination, with rigorous filters, of predictions made by methods based on different principles (2). In this light, the improvement of individual predictors is of course desirable. Indeed, resequencing is becoming ever more accessible and economical, and in some experimental contexts, notably the development of molecular markers for crop and animal positional cloning and marker-assisted breeding programs, workers are likely to prefer to use one or two methods to maximize the detection of small to medium insertions and deletions (1–30 bp) without the requirement of implementing and optimizing particularly complex bioinformatics pipelines. We provide evidence that combining our method with split mapping could provide a reasonable starting point for the identification of small- to medium-sized SV events.

MATERIALS AND METHODS

SVM features

For each chromosome, we store the read mapping data in a sorted (ascending order, by mapping coordinates) doubly linked list. Every node (N) in the list contains the following information:

(1) Genomic coordinates (start end) or inner distance (ID).
(2) Coverage by paired and unpaired reads on each strand (within the coordinates).
(3) Observed insert-size distribution of PE 'covering' the node on each strand.

Consecutive positions with identical coverage statistics are merged into single nodes, and positions with no coverage are not incorporated into nodes. Positions and lengths of uncovered regions can be trivially calculated from the difference between the coordinates of two consecutive nodes.

For a given node N (which includes a genomic position X), we call M (genomic position Y) the node exactly 1 insert-size downstream in covered bases. The objective is to identify a site in the reference genome that is beyond the SV event and corresponds to the expected position of mapping of the partners of reads mapping to X. It is acknowledged that bases covered by only redundant mapping reads will lead to errors in the calculation of M as will insertions in the donor genome.

We define the following genomic windows (in ID) X-read length to X, X-10 to X, X to X + 10 and X to X + read length (and equivalent intervals for Y). The windows of 1 read length correspond to expected positions of peaks of BP reads as X moves within one insert size of an SV event, whereas the windows of 10 bases were chosen arbitrarily with the objective of accommodating errors in the estimation of position Y and to aid the precise delineation of the sharp peaks of broken pairs of reads (BPs) expected to flank junctions of deletions in the donor genome.

For each of these windows, we calculate (for each genomic strand) the mean total coverage per base and normalize these values to the total coverage of X or Y, respectively.

For each of these windows, we also record (for each genomic strand) the mean proportion of reads mapping to each of these windows that are BPs.

We define an additional window: X-read length to Y + read length (in ID).

We record the length of this long window in genomic bases, the longest interval of consecutive uncovered bases contained within it and the total number of uncovered bases in the interval.

For each visited node in the long window, we perform the following statistical tests: a *Z* test to compare the observed 'upstream' read length distribution to the global insert-size distribution, another *Z* test to compare the observed 'downstream' read length distribution to the global distribution, a Student *t*-test (Welch) and a Kolmogorov-Smirnov (KS) test to compare the 'downstream' to the 'upstream' distribution to each other. We record the proportion of genomic positions in the window supporting a perturbation of mapping distance according to a particular test with a *P* value within the following ranges: $\leq 10^{-5}$, 10^{-5} to 10^{-4},10^{-4} to 10^{-3} and 10^{-3} to 10^{-2}.

Finally, for each node in the long window, we compute the BPs to total number of reads ratio for each strand and record the proportion of positions on each strand in the window with ratios within the following ranges 0.15–0.25, 0.25–0.50, 0.50–0.75 and >0.75.

The aforementioned statistics are recorded in an ordered vector and used as a feature set for SVM analysis.

Cluster formation and SV calling

Sites classified as non-normal and of the same type by the SVM are merged into clusters when located less than five bases apart on the genome. Clusters with a number of 'non-normal' positions that exceeds an 'indicator cutoff parameter' are promoted to the status of indicators and a comparison of mapping distances for all paired reads mapping to the cluster and pointing toward the putative SV event with the global mean mapping distance is used to estimate where a complementary strand cluster/indicator is expected to fall (two mean insert sizes plus or minus the estimated size of the SV event in the case of deletions and insertions, respectively). If a cluster or an indicator with the same type of predicted event is identified in the expected interval, an event of that type is called at the base falling half way between the outer coordinates of the two supporting clusters. If a cluster or indicator of contradictory type is identified in the expected region, an indeterminate indel (IndIndel) is called. When intervals between two called events overlap by more than 80% of their lengths, the predictions are merged.

Size estimation and detection of heterozygosity

The expected position of the event (or break points) is evidently half way between the two clusters. Once a position has been predicted, a more accurate estimate of the size of the event is obtained by identifying all pairs of reads mapping across the predicted break point and comparing their mean insert size to the mean global insert size.

To discriminate between homozygous and heterozygous events, we use an EM algorithm and a log-likelihood test similar to that implemented in the software Modil (27). In brief, for any genomic locus where an indel has been predicted, we model the mapping distances of reads covering the predicted event data a single distribution (homozygous) or (heterozygous) a pair of distributions, one of which is constrained to the global insert-size distribution and compute the respective likelihoods. At least 30% of reads covering the position must be assigned to each distribution. A log-likelihood test with 1 degree of freedom is used to verify whether the two distribution models are significantly more likely (*P* value $\leq 10^{-3}$).

Coarse filters for the identification of regions potentially containing SVs

The genomic sequence (read map) is traversed in a 5′–3′ direction on each strand. Only positions with total coverage above a 'minimum coverage' parameter are considered. To avoid unnecessary calculations, the SVM is invoked only by sites that satisfy at least one of two 'coarse filter' criteria: if the ratio of BP reads to mapped pair reads overlapping the position is in the highest 'BP proportion parameter' percentage of genomic sites or if the mean insert size falls outside of 'map distance deviation' standard deviations (SDs) of the mean of the global insert-size distribution.

Training and parameter optimization

Randomly selected genomic regions of at least 15 kb in length, within which no bases would invoke SVM analysis and where all bases show coverage to expected coverage ratios of between 0.5 and 4 are selected as templates for SVM training and parameter estimation.

To produce the training set for the SVM, we use 100 selected regions and randomly introduce a single insertion or deletion of length 1–2500 bp to each. Real sequence reads are then remapped to the *in silico*-mutated genome. The process is repeated to give a total of 1500 indel events. To simulate the effect of hypervariable regions, random windows of length 35–500 were selected and subjected to random mutation at different substitution rates (10–25%). A total of 1000 simulations were performed for each combination of size and substitution rate. Each position on the positive strand upstream by less than an insert size from *in silico* break points (or polymorphic hot spots) and every position on the negative strand downstream by less than an insert size are labeled with the relevant type of event (deletion, small insertion, large insertion and hypervariable). For each set of remapped reads, the initial coarse filters are re-applied and features calculated around positions, which would invoke the SVM. Appropriately labeled feature vectors are used in conjunction with the *libsvm* facilities to train a multi-class SVM and obtain the SVM model file. The polynomial kernel was used in all experiments.

Several parameters required for the analysis must be specified at runtime. The minimum coverage parameter determines the minimum total read coverage of a site for consideration. The 'BP proportion parameter' and the 'map distance deviation' parameters govern the invocation of the SVM, whereas the cluster promotion parameter is required in the definition of indicators in the post processing step. These values can be determined by the user or optimized after SVM training using provided scripts. These tools perform simple parameter sweeps and attempt to the select parameter values that minimize the number of overlapping predictions and false-positive predictions with the optimized SVM model and a subset of the simulated events that were not used in SVM training.

Data pre-processing and mapping of reads

To evaluate the proposed method, we downloaded 3.5 billion reads (1.75 billion pairs of reads) from the NCBI short read archive: ftp://ftp-private.ncbi.nlm.nih.gov/sra. All reads were 36 bases in length, and the libraries contained theoretical insert sizes of ~208 bases. Similar to Hormozdiari *et al.* (28), we removed any read (and its mate) where the average phred quality was below 20 and pairs of reads where one read contained more than 2 Ns. This leads to the elimination of 650 million pairs. We aligned the reads to the human genome hg18 reference assembly using SOAP2 (33), allowing only unique mapping reads/pairs with up to 2 mismatches/read.

This generated 1 billion uniquely mapping pairs and 40 million uniquely mapping unpaired reads.

Predictions from other tools, data download and comparison criteria

We ran BreakDancer on our mapping data using the parameters reported in the original article. PinDel predictions from the same data set were downloaded from http://www.ebi.ac.uk/~kye/pindel/ and Variation Hunter predictions from http://compbio.cs.sfu.ca/strvar.htm.

Repeat and gene annotations were downloaded from the UCSC genome browser (genome-mysql.cse.ucsc.edu).

To compare different validation and prediction set, we used the latest version of the intersectBED program from the BEDtools (34) suite and custom Perl scripts. We used simple overlap ($\geq$1 bp) between different sets as main criterion of validation/equivalency. As the significant intervals predicted by PE-based tools tended to be longer (avg 290 bp), respect to the predictions by Pindel, we extended Pindel predictions by 60 bp upstream and downstream.

RESULTS

Rationale and description of the approach

In the vicinity of indels between a donor and a reference genome, three types of perturbations in the 'normal' PE mapping pattern are expected—in different degrees—depending on the type of event (deletion in donor genome, insertion smaller than library insert size and insertion larger than library insert size).

First, PE reads spanning the indel will show a perturbation from the expected mapping distance (increased distance for a deletion, decreased for an insertion in the donor genome provided that the insertion event is smaller than the library insert size. Insertion events larger than the library insert will lead to an absence of PE reads spanning the junction on the donor genome). These phenomena are expected to be observed within one library insert size 5′ of junctions of rearrangements.

Second, given sufficient sequence coverage and presuming correct and comprehensive mapping of reads, a peak of BP mappings is expected to be observed from one library insert size 5′ of rearrangement junctions, toward the junctions. In the case of deletions in the donor genome, this peak will be narrow (one read length) as only reads mapping on the rearrangement junction will fail to map, whereas in the case of an insertion in the donor genome, this peak will extend the length of the insertion toward the rearrangement junction.

Finally, and as a corollary to the previous observation, the rearrangement junctions (and the region deleted in the case of deletions in the donor genome) will show an absence of coverage by any reads (PE or BP). A schematic illustration of these expected patterns is provided in Supplementary Figure S1.

Existing tools to exploit PE mapping perform a single statistical test, comparing the local insert-size distribution with that for all mapped PE reads. The assumption underlying our approach is that avoiding the use of stringent statistical cutoff values by using a series of *ad hoc* descriptors of read mapping patterns, supervised learning and searching for concordance between neighboring genomic sites, it might be possible to improve sensitivity of SV detection without loss of specificity.

In this method, for any genomic position, we first attempt to identify the expected mapping position for the partners of PE reads covering that position. We then define a series of genomic windows centered on these positions (see 'Materials and Methods' section and Supplementary Figure S2). For these windows, statistics regarding the aforementioned phenomena is recorded, and a multi class SVM classifier is used to assign the site to one of several different categories ('normal', flanking a deletion, flanking a small insertion, flanking a long insertion and flanking a hypervariable region).

In practice, as the starting position approaches a SV event, the disposition of different types of perturbations along the different windows changes, meaning that a single characteristic pattern of feature value biases cannot be associated linearly with a single type of event. However, the advantage of SVM over hierarchical methods such as decision tree is that it is not necessarily 'looking' for a single combination of feature values to make a classification, rather, it should recognize different patterns that were associated with a class in training.

It is of course expected that multiple sites flanking a single SV event (upstream on each strand) will be recognized by the SVM classifier as indicating a similar type of event, and this expectation is exploited in a post-processing step that detects relevant clusters of indicative sites on each strand of the genomic sequence and calls insertion and deletion events between complementary clusters, where such cluster conflicts in their assignment of the nature of their event, we assign an indeterminate indel (IndIndel). Finally, dimensions of called events are estimated by comparing map distances of PE reads spanning the predicted event to the global mean insert size, and a likelihood-based method is applied to identify heterozygous SVs. As for other mapping distance methods, an inherent weakness of our approach is its relative inability to detect insertion events larger than the PE library insert size. Indeed, although it uses BP data and might be expected to detect some such events in regions of high sequence coverage, it is unable to estimate the insertion size.

Although the implementation of our approach is efficient and rapid, it is not necessary to apply the SVM to all positions in the reference genome. Our method uses initial filters to identify a subset of genomic positions where either the ratio of mapped unpaired reads to paired reads or the mean insert size of reads on one strand are potentially anomalous with respect the global situation. Given that most SVs are very short (too short to actually perturb the insert-size distribution), the net effect is that the SVM is mostly invoked as a consequence of the presence of BP reads.

The SVM itself is trained using the experimental data and genome sequence under study, with simulated insertion and deletion events. Several parameters relevant to the analytical pipeline are also optimized automatically during the training of the system for a particular combination of data set and genome. The method is implemented in the software SVM^2—a package written in C++ with accompanying Perl scripts and uses the freely available Libsvm package (35). SVM^2 is rapid and requires only limited RAM memory after the initial read mapping phase.

Simulation

To estimate the specificity and sensitivity of our method, we artificially implanted 9000 random insertions and deletions of different sizes (1–600 bp—note that beyond the library insert size, detection of insertion events is not influenced by dimensions of the insertion) into human chromosome 17 (hg18 assembly) and generated artificial reads (theoretical coverage 30X, error rate 1%) from the mutated genome (mate pairs: insert size, 208; standard deviation, 13 and theoretical coverage, 35X) using the dwgsim program (36).

Results presented in Table 1 show good overall recall rates (88% and 91% for insertions and deletions, respectively) and generally low false-positive rates. The column 'recall' indicates percentage of simulated events that were correctly classified as insertion or deletion (subdivided by the actual length of the event simulated), whereas the 'recall as any' column shows the total proportion of simulated events that were identified as either insertions or deletions. It is clear that both false-positive predictions and misclassification of the nature of events constitute significant issues only with predictions of very short events (less than 10 bases). An exception to this trend is provided by insertions longer than the insert size, which are recovered with a slightly lower recall rate. This is unsurprising given that detection of such events relies exclusively on the presence of BPs.

Heterozygosity

The identification of heterozygous SV events is an inherently difficult problem for non-alignment-based methods. Heterozygosity reduces apparent perturbation in insert sizes and lowers ratios of unpaired to paired reads. Low read depth also raises the probability of unequal sampling of haplotypes, further complicating the issue. However, we anticipated that with sufficient depth of coverage our method should be able to recognize longer indel events.

Accordingly, we simulated a set of SVs of different size (1–40 bp, 250 events/size/category) with a theoretical 40X depth of coverage, and for each SV, we generated the heterozygous and the homozygous version. The results are summarized in Table 2, for each set of SV, we computed the recall rate for the homozygous and the heterozygous event and the fraction of heterozygous SVs that was correctly classified as heterozygous (see 'Materials and Methods' section). The results confirm that although our method is particularly accurate in detecting homozygous SV of any size, it lacks sensitivity both in the detection and

Table 1. Simulation

Size	Recall[a] (%)	Recall as any[b] (%)	FP rate (%)
Deletions			
1–5	69	83	8.5
6–10	82	89	6.3
11–20	91	92	2
21–40	94	94	0
41–60	97	97	0
61–100	95	95	0
101–200	97	97	0
>200	97	97	0
Insertions			
1–5	70	86	9
6–10	82	88	6
11–20	94	94	2
21–40	92	92	0.5
41–60	93	93	0
61–100	91	91	0
101–200	89	89	0
>200	86	86.00	0

[a] Correctly classified as insertion or deletion.
[b] Correctly identified locus, includes indindel and hypervariable predictions.

Table 2. Simulations of heterozygous events

Size[a]	Recall rate[b] (%)	Correctly classified[c] (%)	Recall rate if homozygous[d] (%)
Deletions			
1	10	0	83
3	10	0	87
5	13	15	94
10	40	20	98
15	53	29	99
20	63	45	99
30	85	87.5	99
40	87.5	93.5	99
Insertions			
1	10	0	80
3	10	0	86
5	17	3	94
10	28	14	99
15	48	32	99
20	57	47	99
30	81	89	99
40	88	96	99

[a] Size of the event.
[b] Recall rate for the heterozygous case.
[c] Proportion of recalled indels classified as heterozygous.
[d] Recall rates for equivalent (same locus) homozygous indels.

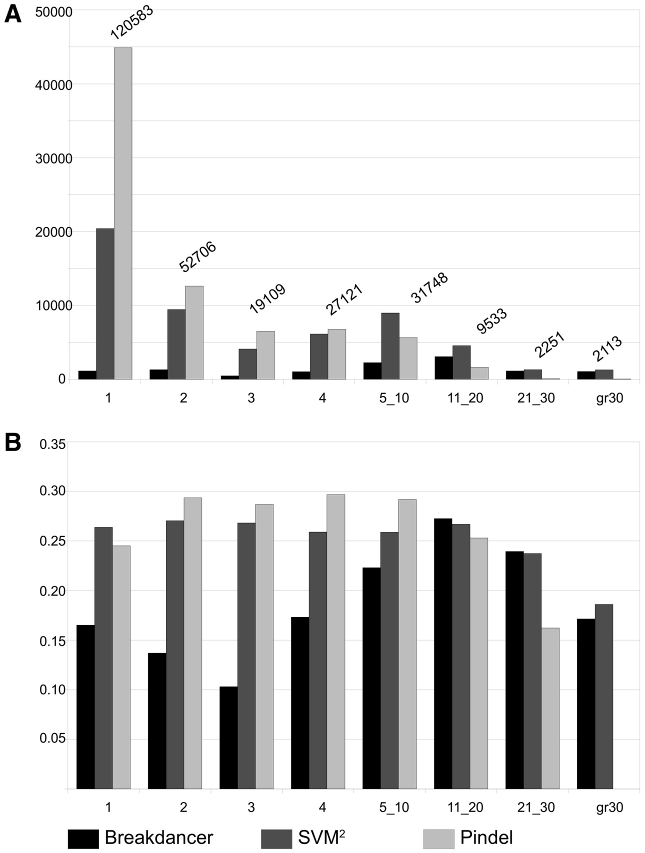

Figure 1. Sensitivity and specificity of different methods with the Kidd *et al.* data set. (**A**) Number of indels from the Kidd *et al.* data set (binned by size of event in bp) recalled by each method. (**B**) Proportion of predicted indels (binned by predicted sizes) that are validated by an indel in the Kidd *et al.* validation set. Size bins: size ≤ 1, size ≤ 2, size ≤ 3, size ≤ 4, 5 ≤ size ≤ 10, 10 < size ≤ 20, 20 < size ≤ 30 and size > 30.

correct classification of heterozygous SVs less than 20 bases in length. Additional simulations showed that, as expected, proportions of heterozygous alleles sampled in resequencing impacts on detection and classification (Supplementary Table S1).

Comparison with other tools

To compare the performance of our method to other tools using real PE resequencing data, we have taken advantage of publicly available PE resequencing data from an anonymous human donor (37) generated with the Illumina technology. The peculiarity of this data set is that a large and consistent set of SV was previously detected and validated using low-coverage (0.3X) longer insert (Sanger 40 + kb fosmids) from the same individual (38), thus it has been widely used as a benchmark to compare different SV detection tools. Indeed, the Kidd *et al.* data were recently subjected to a second analysis (39), and in this study, we consider the union of both sets of predictions as a validated indel set (265 264 events).

We compared the performance of our tool with that of BreakDancer (26) (a widely used PE-based method) that, in previous studies of the same data set, exhibited the highest sensitivity and specificity among PE-based tools in detecting relatively small indels (indicatively greater than 10 bp) and PinDel (21), a popular split-mapping approach.

The sensitivity (the proportion of indels in the validation set that was recovered by each method, as a function of the validated size of the indel) of each method is shown in Figure 1A (and Supplementary Table S2). Under this criterion, SVM^2 outperforms BreakDancer in all size categories, overall recalling 4.5 times as many events. As expected, the split-mapping method (PinDel) is more sensitive in the detection of very small indels (up to 5 bp), although SVM^2 recalls a larger proportion of events over this threshold.

The number of predictions and apparent specificity by predicted event size (proportion of predicted indels of coinciding with any indel in the validation set as a function of the predicted size of the indel) for each method are shown in Figure 1B (and Supplementary Table S3). It should be noted that the genome coverage of the Kidd *et al.* data, 0.3X, represents the maximum theoretical specificity in this benchmark. All the evaluated methods demonstrate similar overall performance. PinDel in particular shows a marginally better specificity with respect to the smallest events (<10 bp), whereas the size/specificity profile of SVM^2 and BreakDancer are relatively uniform at approximately 26–27% 'validation' for each size bin. Both SVM^2 and BreakDancer suffer an apparent loss in specificity with regard to predicted events greater than 30 bp or more. This last observation is likely a stochastic effect due to the fact that larger rearrangements constitute a very small minority of SVs. To partially ameliorate the low genome coverage of the validation set, we compared predictions with all events in dbSNP 130, which contains more than 4.2 million known rearrangements derived mostly from Sanger sequencing data (39). The 81.5%, 80.6% and 80.4% of the predictions made by BreakDancer, PinDel and SVM^2, respectively, correspond to known human SV events. The specificity by size profile strongly resembles that observed with the Kidd *et al.* SVs (Supplementary Figure S3A and Supplementary Table S4). Cross referencing the predictions from the various methods with the collection of human genomic SVs provided by the 1000 genomes project, derived from NGS data (2) (1.32 million events), showed that 61% of BreakDancer predictions, 69% of SVM^2 predictions and 80.7% of PinDel predictions were coincident with events present in that database; 54% of the Kidd *et al.*/Sanger-based validation set events were present in the 1000 genomes database (Supplementary Figure S3B and Supplementary Table S4).

Although the identification of very large SVs is not a primary objective of our method, we also compared the capacity of several methods to identify 98 long deletions (10 kb or more) called in the original work of Kidd *et al.* In this particular task, Variation Hunter, a method developed specifically for the identification of large SVs (28) recovered 65 events, whereas BreakDancer and SVM^2 recalled 55 and 51 events, respectively. All tested methods demonstrated a lower apparent specificity than GASVPro (30), underlining the capacity of that method in its specialized function—namely the detection of large deletions.

The Venn diagram in Figure 2 shows the overlap of validated calls made by SVM^2, BreakDancer and PinDel. The union of all methods identified 108 158 of the 265 264 events recovered from the Sanger data (41%); 24 842 (23%) are found by PinDel and SVM^2, and 9122 (8.5%) are identified by BreakDancer and SVM^2. Only 1730 (1.5%) are found by BreakDancer only, whereas 49 972 (46%) are unique to PinDel and 20 974 (19%) are unique to SVM^2; 87% of validated BreakDancer predictions are also made by SVM^2. Taken together, these observations confirm that the incorporation of additional mapping information in SVM^2 allows a great increase in sensitivity over methods that use only mapping distance information. Furthermore, it is evident that a notable proportion of events are recovered by SVM^2 but not other methods. When compared with the sensitivity profile by event size (Figure 1A), it is evident that SVM^2 identifies a significant number of small events not detected by PinDel.

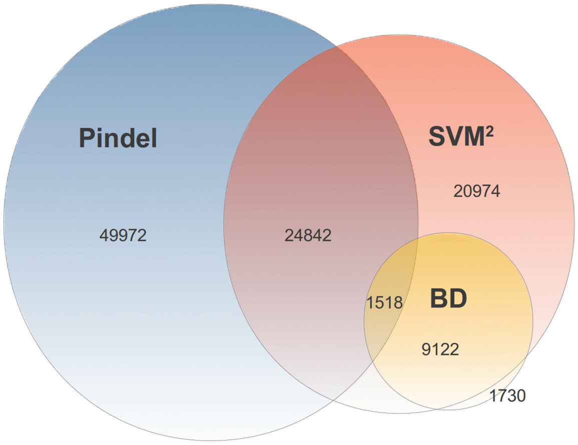

Figure 2. Venn diagram showing intersection between validated (by Kidd *et al.*) predictions by each method.

Accuracy of classification and genomic context of predictions

The analysis of simulated data suggested that, for small events, SVM^2 may lack precision in classification. Furthermore, genomic clustering of SV in variation 'hot spots' may additionally complicate classification of real events. As for the simulations, SVM^2 showed a tendency to misclassify only small events ($\leq$5 bp). Table 3 summarizes the classification patterns for such events, whereas Supplementary Figure S4 illustrates profiles of SV size prediction accuracy for SVM^2. SVM^2 shows a tendency misclassify small ($\leq$5 bp) deletions rather than insertions. Consistent with the difficulty of classifying small events, our hypervariable and indindel predictions almost exclusively contain small indels at a similar validation rate to other categories.

Next, we asked whether for a series of size range bins, the sensitivity by genomic context showed obvious differences between methods. Figure 3 confirms that for the smallest events ($\leq$5 bp), PinDel outperformed the other methods in most genomic contexts. However, the sensitivity of SVM^2 in SINEs and low-complexity regions was comparable with that of PinDel, whereas in simple repeats, SVM^2 outperformed PinDel. As expected, given the small number of predictions by BreakDancer in this size range, the sensitivity was low. For events of between 6 and 10 bp in size, SVM^2 was the most sensitive method, dramatically outperforming BreakDancer in all genomic contexts. PinDel was almost as sensitive as SVM^2 in DNA transposons and non-repetitive DNA. As event size increases, PinDel shows decreasing sensitivity particularly

Table 3. Classification accuracy of short indels predicted by SVM^2

SVM^{2a}	Total^b	Kidd^c			
		Insertions	Deletions	Validation rate^d	Misclassification rate^e
Insertions	50 688	11 068	2288	13 356 (26.3%)	17.1
Deletions	46 102	3991	8111	12 102 (26.2%)	32
IndIndels	9118	1308	1049	2357 (25.8%)	
Hypervariable	8503	1268	982	2250 (26.4%)	

[a] Class predicted by SVM^2.
[b] Number of predictions by SVM^2 by category.
[c] Class of the validating event in the data set.
[d] Validation rate for each category of SVM^2 predictions (applies for small only).
[e] Misclassification rate for validated insertions and deletions.

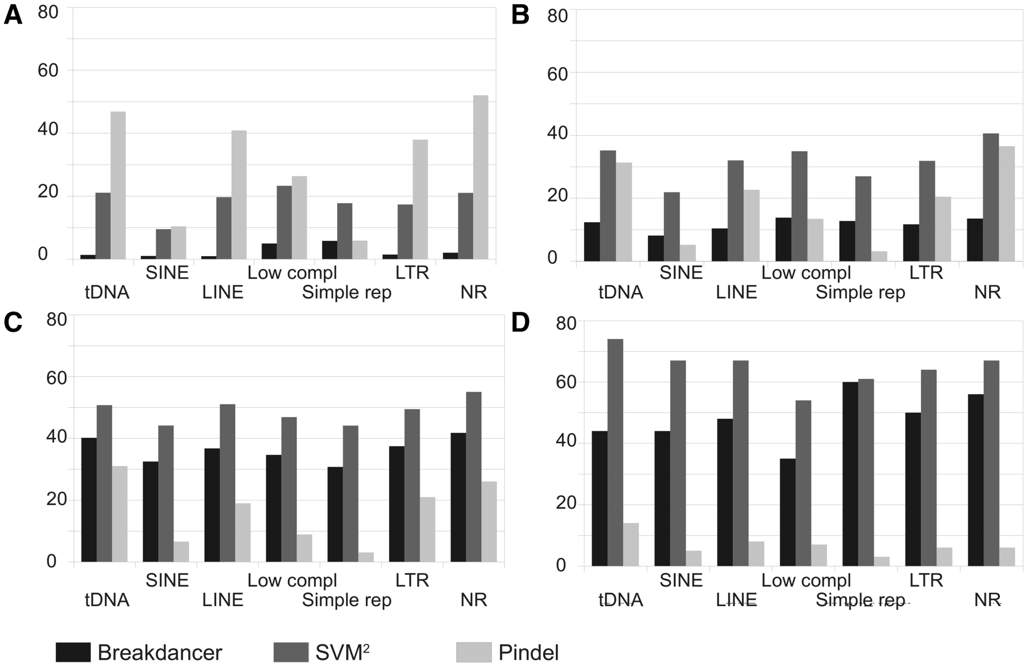

Figure 3. Sensitivity by size and genomic context. Fraction of events in the Kidd *et al.* data set, in different genomic contexts (tDNA = DNA transposon, LTR = long terminal repeats, NR = non-repetitive), recalled at different size ranges [size ≤ 5 (**A**), 5 < size ≤ 10 (**B**), 10 < size ≤ 20 (**C**), size > 20 (**D**)] by different methods.

in low-complexity regions and simple repeats (an inevitable property of split-mapping methods). Even for larger (>20 bp) events, which BreakDancer was designed to detect, SVM^2 is more sensitive in all genomic contexts. It is notable that, overall, SVM^2 and BreakDancer seem to show much less dependence on genomic context than PinDel.

We were intrigued by the difference of apparent specificities between methods previously observed when using the 1000 genomes SV catalog (but not when using dbSNP or the Kidd *et al.* data) as a validation set and by the relatively large proportion of the small (<10 bp) events found by SVM^2 but not PinDel that fall in low-complexity and simple repeat regions (10 037/19 274, 52%). We reasoned that these observations might be linked by the fact that the 1000 genomes catalog used split mapping to identify small event, and showed that a notable proportion (>97%) of the part of the genome deemed 'inaccessible' by their low coverage data, fell in regions annotated as 'high copy repeats or segmental duplications' (2). Accordingly, we investigated the genomic distribution of predictions validated by Sanger sequencing but not by the 1000 genomes catalog by event size and method. We observed that a relatively small proportion of the small events (<10 bp) validated by the Kidd *et al.* data and predicted by PinDel but not supported by the 1000 genomes data set fall in low-complexity regions and simple repeats (1483/16 081, 9.25%), whereas the equivalent numbers for SVM^2 were (5991/18 450, 32%) suggesting that SVM^2, or similar methods, might effectively complement existing tools and pipelines in the detection of very short SVs, particularly in repetitive and low sequence complexity areas of the genome.

Finally, we compared the frequency of predictions by SVM^2 in genic regions with the rest of the genome, reasoning that SV events should occur at lower frequency in the former; 1.2% and 0.27% of predictions fell in genic and CDS regions, respectively (using refseq genes). We estimated the significance of the difference between expected and observed frequencies using the Poisson distribution. The departure from the null model that predictions are distributed randomly along the genome was $<10^{-20}$ for both categories.

DISCUSSION

With simulated data, both the sensitivity and specificity attained by our method were exceptionally high, although it should be emphasized that other methods have generated similarly impressive results in similar benchmarks but show, in particular, lower sensitivity with real data (26,27). This is unsurprising as the effects of repetitive sequences and inherent biases in sequence coverage tend to be minimized in simulations. However, for the study of heterozygous events, simulation for now provides the only realistic possibility, due to the lack of large scale validated heterozygosity catalogs associated with individual genomes. SVM^2 showed relatively poor accuracy in the detection and classification of very short heterozygous SV. All mapping-distance-based methods are expected to suffer from this limitation as distance perturbations are diluted at heterozygous loci. In addition, our current approach uses measures of coverage, and in the case of heterozygous deletions, a reduction rather than an absence of reads in the deleted region would be expected. Conversely, reduced perturbations of BP mapping patterns are expected upstream of heterozygous insertions. These limitations might be partially addressed by some of the potential developments in the strategy that are envisaged (see later). However, in simulation at least, we note a satisfactory performance by SVM^2 in the identification and classification of larger heterozygous events.

In this work, SVM^2 was trained to recognize hypervariable regions as distinct from SV events. In practice, few predictions of this type were made. Indeed, an examination of these predictions suggested that they showed a similar specificity in detection of SVs as the other categories of prediction—although all validated predictions in this category corresponded to events of four bases or less. This is likely a function of the read mapping strategy used. Allowing up to 2 mismatches in 35 base reads tends to allow correct mapping of the majority of reads in intra-specific comparisons, and in any case, perturbations of read mapping caused by hypervariable genomic regions are expected to be extremely subtle.

The Bentley *et al.*/Kidd *et al.* data represent one of the few cases where extensive Sanger resequencing and SV calling have been performed on an individual for which PE NGS data are also available, providing an 'independent' validation set. For this reason, the data set has been widely used in other studies (21,26,27) and allows immediate comparisons between methods. These considerations notwithstanding, the data set has several relevant limitations that complicate interpretation of results and merit discussion. First, the coverage by Sanger sequencing is rather limited (theoretical coverage 0.3X), suggesting that, even if we make the—optimistic—assumption that all reads were mapped correctly and uniquely, at most less than a third of the SV events between this individual and the hg18 reference could be detected. Second, the low coverage implies that the accurate annotation of heterozygous events should be, at best, extremely limited. Finally, the original study of Kidd *et al.* only attempted to identify events of less than 100 bp in length, and although a second evaluation of these data (39) was more comprehensive, the detection of large insertions is limited by the properties of split-mapping methods. It has been suggested that the majority of intra-specific SVs are small (32), and although this generalization is almost certainly correct, our knowledge of the frequency of medium to large events remains rather limited. Our method made few predictions of insertions larger than the insert size of the library. However, this is an inherently difficult category of events to detect by any current approach and, with the available data, it is difficult to perform statistical analysis of sensitivity and specificity of tools with respect to detection of such events.

Taken together, these observations render the objective assessment of the overall specificity of methods, with respect to both homozygous and heterozygous SV, extremely difficult. Additionally, the probability that a proportion of the Kidd *et al.* and Mills *et al.* predictions are heterozygous complicates estimates of sensitivity with respect to homozygous events. In this context, we believe that although limited in precision, apparent sensitivity and specificity are the best available metrics for comparison of the performance of different methods. By all metrics and validation sets used, SVM^2 outperformed BreakDancer in terms of sensitivity over a range of SV event sizes, attaining at least the same apparent specificity. This is perhaps not surprising given that additional mapping information, not used by BreakDancer, is used by SVM^2. Perhaps more relevant is the observation that SVM^2 identified a large number of small SVs that were not detected by a contemporary split-mapping method.

One alternative to the use of individual genome Sanger resequencing as a biological validation set would be to estimate specificity by comparing genome wide predictions to collections of validated population level SVs [dbSNP (40), 1000 genomes project (2)] making the assumption that coincidence of predictions with an annotated SV implied the presence of the same SV in the donor genome. However, a recent study demonstrated a relatively low overlap between the two aforementioned databases, implying that a significant fraction of human SVs remain undetected (39). It is also worth noting that the 1000 genomes set of SV events was generated from NGS data. Given that our objective was to explore the potential of this very type of data to uncover additional, previously undetected events, we consider that the use of 'independent' data from the individual genome under study as our principal validation set to be a justified strategy. Nevertheless, comparisons of apparent specificity of different methods when 'validated' by Sanger or NGS-based data sets showed interesting patterns, particularly with respect to the genomic context of indel events.

The 'elephant in the room' of all methods to determine locations of SV from resequencing data, be they based on split mapping or on statistical approaches, is the abundance of repeated sequences in complex genomes. Sequence reads (from any technology) that fall within perfectly repeated regions cannot be unambiguously mapped. PE approaches (dependent on library insert size and repeat length) can ameliorate this problem to some extent, as can probabilistic mapping strategies (28), but the fundamental problem remains. For example, SVs within recent segmental duplications present an almost insurmountable problem for all approaches apart from read-depth methods—and even these will not be able to specify the location of the event. For now, the most promising way to address the problem of repeats may be the maximization of read length and the use of different insert-size libraries. The use of larger insert-size libraries will aid the detection of larger SV events by insert-size-based methods (and contribute to an additional loss of accuracy in the identification of small indels by such methods). Conversely, as the production of longer resequencing reads using NGS technologies becomes more commonplace, the sensitivity of split-mapping methods is expected to increase for small to medium size events and to reduce the impact of repetitive sequences on the performance of all methods. Despite these problems, we note that our analyses of genomic context of predictions and validated predictions suggest that in simple repeats and low-complexity regions, SVM^2 attained higher sensitivity than other methods tested, even for small SV events. The observations that a large number of small SV events detected by Sanger resequencing, but not by PinDel (or 1000 genomes) fall in simple repeat and low-complexity regions, and that a larger proportion of validated SVM^2 than PinDel predictions fall in such regions are interesting. In this light, the similarity of overall 'specificity' between methods when evaluated with the Kidd *et al.* data or with dbSNP and the differences in this metric with respect to the 1000 genomes database is intriguing, particularly given the types of data used to construct these catalogs. Simple repeat/low-complexity regions represent a notable proportion of the 'inaccessible' genome described by the 1000 genomes consortium (2). We suggest that our method, or others based on similar principles, might be of particular use in addressing SV in such regions. Indeed, it is interesting to note that Breakpointer (29), a recently proposed method that incorporates information from read depth, mismatch profiles and split mapping to identify genomic rearrangements also showed an increased sensitivity to SV in repetitive regions with respect to PinDel. However, Breakpointer, unlike PinDel or SVM^2, is apparently not capable of identifying the very smallest (<3 bp) SV events, again emphasizing the value of using complementary approaches dedicated to the detection of different types of events.

We can envisage several potential developments to the approach presented in this study, some of which might be expected to improve the performance with respect to heterozygous SV. First, sequence coverage might be improved by using split mapping in the initial generation of read maps (herein we have used only gapless alignment). Second, additional features, for example the gapless and gapped alignment coverage for each genomic site could be incorporated into the SVM analysis. Another possible step would be to use positional constraints (based on SVM^2 predictions) in split mapping of reads as a post-processing step in establishing additional support for events and in fine mapping positions of SV as currently implemented in Breakpointer (29).

In conclusion, we have shown that inclusion of more detailed information on the local patterns of read mapping can notably enhance the sensitivity of detection of SV events by non-split-mapping methodologies.

Furthermore, we showed that insert-size-based SV detectors such as SVM^2 can complement split-mapping approaches in the localization of ultra-short SV events, particularly those in repetitive and low-complexity regions of the genome.

AVAILABILITY

The SVM^2 software, documentation and example files are available via anonymous ftp from ftp:159.149.109.10/pub/svm2.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Tables 1–4, Supplementary Figures 1–4 and Supplementary Data.

FUNDING

'Ministero dell'Istruzione Università e Ricerca' (MIUR, Italy): 'Laboratorio Internazionale di Bioinformatica', 'Laboratorio di Bioinformatica per la Biodiversità Molecolare' [Project DM19410] and by the Center of Excellence in Genomics (CEGBA, Italy). Funding for open access charge: MIUR/Laboratorio Internazionale di Bioinformatica.

Conflict of interest statement. None declared.

REFERENCES

1. Shendure,J. and Ji,H. (2008) Next-generation DNA sequencing. *Nat. Biotechnol.*, **26**, 1135–1145.
2. 1000 Genomes Project Consortium. (2010) A map of human genome variation from population-scale sequencing. *Nature*, **467**, 1061–1073.
3. Iafrate,A.J., Feuk,L., Rivera,M.N., Listewnik,M.L., Donahoe,P.K., Qi,Y., Scherer,S.W. and Lee,C. (2004) Detection of large-scale variation in the human genome. *Nat. Genet.*, **36**, 949–951.
4. Sebat,J., Lakshmi,B., Troge,J., Alexander,J., Young,J., Lundin,P., Månér,S., Massa,H., Walker,M., Chi,M. *et al.* (2004) Large-scale copy number polymorphism in the human genome. *Science*, **305**, 525–528.
5. Sharp,A.J., Bailey,J.A., Kaul,R., Morrison,V.A., Pertz,L.M., Haugen,E., Hayden,H., Albertson,D., Pinkel,D., Olson,M.V. *et al.* (2005) Fine-scale structural variation of the human genome. *Nat. Genet.*, **37**, 727–732.
6. Buchanan,J.A. and Scherer,S.W. (2008) Contemplating effects of genomic structural variation. *Genet. Med*, **10**, 639–647.

7. McCarroll,S.A., Kuruvilla,F.G., Korn,J.M., Cawley,S., Nemesh,J., Wysoker,A., Shapero,M.H., de Bakker,P.I., Maller,J.B., Kirby,A. *et al.* (2008) Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nat. Genet.*, **40**, 1166–1174.
8. Conrad,D.F., Pinto,D., Redon,R., Feuk,L., Gokcumen,O., Zhang,Y., Aerts,J., Andrews,T.D., Barnes,C., Campbell,P. *et al.* (2010) Origins and functional impact of copy number variation in the human genome. *Nature*, **464**, 704–712.
9. Mills,R.E., Walter,K., Stewart,C., Handsaker,R.E., Chen,K., Alkan,C., Abyzov,A., Yoon,S.C., Ye,K., Cheetham,R.K. *et al.* (2011) Mapping copy number variation by population-scale genome sequencing. *Nature*, **470**, 59–65.
10. Braude,I., Vukovic,B., Prasad,M., Marrano,P., Turley,S., Barber,D., Zielenska,M. and Squire,J.A. (2006) Large scale copy number variation (cnv) at 14q12 is associated with the presence of genomic abnormalities in neoplasia. *BMC Genomics*, **7**.
11. Bijlsma,E.K., Gijsbers,A.C., Schuurs-Hoeijmakers,J.H., van Haeringen,A., Fransen van de Putte,D.E., Anderlid,B.M., Lundin,J., Lapunzina,P., Pérez Jurado,L.A., Delle Chiaie,B. *et al.* (2009) Extending the phenotype of recurrent rearrangements of 16p11.2: deletions in mentally retarded patients without autism and in normal individuals. *Eur. J. Med. Genet.*, **52**, 77–87.
12. McCarthy,S.E., Makarov,V., Kirov,G., Addington,A.M., McClellan,J., Yoon,S., Perkins,D.O., Dickel,D.E., Kusenda,M., Krastoshevsky,O. *et al.* (2009) Microduplications of 16p11.2 are associated with schizophrenia. *Nat. Genet.*, **41**, 1223–1227.
13. Tam,G.W., Redon,R., Carter,N.P. and Grant,S.G. (2009) The role of DNA copy number variation in schizophrenia. *Biol. Psychiatry*, **66**, 1005–1012.
14. Ballif,B.C., Theisen,A., Rosenfeld,J.A., Traylor,R.N., Gastier-Foster,J., Thrush,D.L., Astbury,C., Bartholomew,D., McBride,K.L. *et al.* (2010) Identification of a recurrent microdeletion at 17q23.1q23.2 flanked by segmental duplications associated with heart defects and limb abnormalities. *Am. J. Hum. Genet.*, **86**, 454–461.
15. Clayton-Smith,J., Giblin,C., Smith,R.A., Dunn,C. and Willatt,L. (2010) Familial 3q29 microdeletion syndrome providing further evidence of involvement of the 3q29 region in bipolar disorder. *Clin. Dysmorphol.*, **19**, 128–132.
16. Pinto,D., Pagnamenta,A.T., Klei,L., Anney,R., Merico,D., Regan,R., Conroy,J., Magalhaes,T.R., Correia,C., Abrahams,B.S. *et al.* (2010) Functional impact of global rare copy number variation in autism spectrum disorders. *Nature*, **466**, 368–372.
17. Stankiewicz,P. and Lupski,J.R. (2010) Structural variation in the human genome and its role in disease. *Annu. Rev. Med*, **61**, 437–445.
18. Albers,C.A., Lunter,G., MacArthur,D.G., McVean,G., Ouwehand,W.H. and Durbin,R. (2011) Dindel: accurate indel calls from short-read data. *Genome Res.*, **21**, 961–973.
19. Handsaker,R.E., Korn,J.M., Nemesh,J. and McCarroll,S.A. (2011) Discovery and genotyping of genome structural polymorphism by sequencing on a population scale. *Nat. Genet*, **43**, 269–276.
20. Hajirasouliha,I., Hormozdiari,F., Alkan,C., Kidd,J.M., Birol,I., Eichler,E.E. and Sahinalp,S.C. (2010) Detection and characterization of novel sequence insertions using paired-end next-generation sequencing. *Bioinformatics*, **26**, 1277–1283.
21. Ye,K., Schulz,M.H., Long,Q., Apweiler,R. and Ning,Z. (2009) Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics*, **25**, 2865–2871.
22. Medvedev,P., Stanciu,M. and Brudno,M. (2009) Computational methods for discovering structural variation with next-generation sequencing. *Nat. Methods*, **6(Suppl. 11)**, S13–S20.
23. Alkan,C., Koe,B.P. and Eichler,E.E. (2011) Genome structural variation discovery and genotyping. *Nat. Rev. Genet.*, **12**, 363–376.
24. Yoon,S., Xuan,Z., Makarov,V., Ye,K. and Sebat,J. (2009) Sensitive and accurate detection of copy number variants using read depth of coverage. *Genome. Res*, **19**, 1586–1592.
25. Abyzov,A., Urban,A.E., Snyder,M. and Gerstein,M. (2011) CNVnator: an approach to discover, genotype and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Res.*, **21**, 974–984.
26. Chen,K., Wallis,J.W., McLellan,M.D., Larson,D.E., Kalicki,J.M., Pohl,C.S., McGrath,S.D., Wendl,M.C., Zhang,Q., Locke,D.P. *et al.* (2009) Breakdancer: an algorithm for high-resolution mapping of genomic structural variation. *Nat. Methods*, **6**, 677–681.
27. Lee,S., Hormozdiari,F., Alkan,C. and Brudno,M. (2009) Modil: detecting small indels from clone-end sequencing with mixtures of distributions. *Nat. Methods*, **6**, 473–474.
28. Hormozdiari,F., Hajirasouliha,I., Dao,P., Hach,F., Yorukoglu,D., Alkan,C., Eichler,E.E. and Sahinalp,S.C. (2010) Next-generation Variation Hunter: combinatorial algorithms for transposon insertion discovery. *Bioinformatics*, **26**, i350–i357.
29. Sun,R., Love,M.I., Zemojtel,T., Emde,A.K., Chung,H.R., Vingron,M. and Haas,S.A. Breakpointer: using local mapping artifacts to support sequence breakpoint discovery from single-end reads. *Bioinformatics*, **28**, 1024–1025.
30. Sindi,S.S., Onal,S., Peng,L., Wu,H.T. and Raphael,B.J. (2012) An integrative probabilistic model for identification of structural variation in sequencing data. *Genome Biol*, **13**, R22.
31. Noble,W.S. (2006) What is a support vector machine? *Nat. Biotechnol.*, **24**, 1565–1567.
32. Mills,R.E., Luttig,C.T., Larkins,C.E., Beauchamp,A., Tsui,C., Pittard,W.S. and Devine,S.E. (2006) An initial map of insertion and deletion (INDEL) variation in the human genome. *Genome Res.*, **16**, 1182–1190.
33. Li,R., Yu,C., Li,Y., Lam,T.W., Yiu,S.M., Kristiansen,K. and Wang,J. (2009) SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics*, **25**, 1996–1997.
34. Quinlan,A.R. and Hall,I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinfomatics*, **26**, 841–842.
35. Ching,C.C. and Chin,T.L. (2011) Libsvm: a library for support vector machines. *ACM Trans Intell Syst Technol*, **2**, 1–27.
36. Li,H., Handsaker,B., Wysoker,A., Fennell,T., Ruan,J., Homer,N., Marth,G., Abecasis,G. and Durbin,R. (2009) The sequence alignment/map format and samtools. *Bioinformatics*, **25**, 2078–2079.
37. Bentley,D.R., Balasubramanian,S., Swerdlow,H.P., Smith,G.P., Milton,J., Brown,C.G., Hall,K.P., Evers,D.J., Barnes,C.L., Bignell,H.R. *et al.* (2008) Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, **456**, 53–59.
38. Kidd,J.M., Cooper,G.M., Donahue,W.F., Hayden,H.S., Sampas,N., Graves,T., Hansen,N., Teague,B., Alkan,C., Antonacci,F. *et al.* (2008) Mapping and sequencing of structural variation from eight human genomes. *Nature*, **453**, 56–64.
39. Mills,R.E., Pittard,W.S., Mullaney,J.M., Farooq,U., Creasy,T.H., Mahurkar,A.A., Kemeza,D.M., Strassler,D.S., Ponting,C.P., Webber,C. *et al.* (2011) Natural genetic variation caused by small insertions and deletions in the human genome. *Genome Res.*, **6**, 830–839.
40. Sherry,S.T., Ward,M.H., Kholodov,M., Baker,J., Phan,L., Smigielski,E.M. and Sirotkin,K. (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.*, **29**, 308–311.