Bioinformatics Advance Access published June 11, 2015

SoloDel: A probabilistic model for detecting low-frequent somatic deletions from unmatched sequencing data

Junho Kim[1,2], Sanghyeon Kim[3], Hojung Nam[4], Sangwoo Kim[1,*] and Doheon Lee[2,*]

[1]Severance Biomedical Science Institute, Yonsei University College of Medicine, Seoul 120-752, Korea

[2]Department of Bio and Brain Engineering, KAIST, 291 Daehak-ro, Yuseong-gu, Daejeon 305-701, Korea

[3]Stanley Brain Research Laboratory, Stanley Medical Research Institute, 9800 Medical Center Drive, Rockville, MD 20850

[4]School of Information and Communications, Gwangju Institute of Science and Technology, Gwangju 500-712, Korea

Associate Editor: Prof. Alfonso Valencia

ABSTRACT

Motivation: Finding somatic mutations from massively parallel sequencing data is becoming a standard process in genome-based biomedical studies. There are a number of robust methods developed for detecting somatic single nucleotide variations (SNVs). However, detection of somatic copy number alteration (SCNAs) has been substantially less explored and remains vulnerable to frequently raised sampling issues: low frequency in cell population and absence of the matched control samples.

Results: We developed a novel computational method SoloDel that accurately classifies low-frequent somatic deletions from germline ones with or without matched control samples. We first constructed a probabilistic, somatic mutation progression model that describes the occurrence and propagation of the event in the cellular lineage of the sample. We then built a Gaussian mixture model to represent the mixed population of somatic and germline deletions. Parameters of the mixture model could be estimated using the expectation-maximization (EM) algorithm with the observed distribution of read-depth ratios at the points of discordant-read based initial deletion calls. Combined with conventional structural variation caller, SoloDel greatly increased the accuracy in classifying somatic mutations. Even without control, SoloDel maintained a comparable performance in a wide range of mutated subpopulation size (10% to 70%). SoloDel could also successfully recall experimentally validated somatic deletions from previously reported neuropsychiatric whole genome sequencing data.

Availability and implementation: Java-based implementation of the method is available at http://sourceforge.net/projects/solodel/

Contact: kimjh@biosoft.kaist.ac.kr

1 INTRODUCTION

Advances in next-generation sequencing (NGS) technologies enable us to find novel aspects of disease onsets that could not be detected in the previous chip-based technologies. Finding somatic mutations is one representative NGS analysis, to detect genetic variants that occurred after conception. Various types of somatic mutations such as single nucleotide variations (SNVs) (Gregor, et al., 2013; Lee, et al., 2012), gene-fusions (Maher, et al., 2009; Pflueger, et al., 2011), and transposable element insertions (TE insertions) (Helman, et al., 2014; Lee, et al., 2012) have been focused in a number of disease genome studies, which has led to a successful development of the relevant computational detection methods (Cibulskis, et al., 2013; Kim, et al., 2013; Lee, et al., 2012; Roth, et al., 2012).

Somatic copy number alteration (SCNA) is one of the major genomic aberrations frequently found in cancer (Beroukhim, et al., 2010; Chiang, et al., 2009; Mermel, et al., 2011). Several SCNA detection algorithms have been developed based on read-depth comparison (Boeva, et al., 2012; Boeva, et al., 2011; Koboldt, et al., 2012; Xi, et al., 2010) and/or read-pair analysis (Chen, et al., 2009; Rausch, et al., 2012) between tumor and matched control samples. While these algorithms have successfully discovered a number of disease associated SCNAs, several hurdles still remain. First, many somatic mutations reside only in a small population of the sample to be present in a low allele fraction; this problem is well known as tumor cellularity or tumor subclonality in cancer. Furthermore, non-tumor diseases may harbor similar issues due to somatic mosaicism. For example, remarkably low-frequent (down to a few percent) mutations have been reported in brain diseases such as neuropsychiatric disorders and neurodevelopmental malformations (Evrony, et al., 2012; Lee, et al., 2012; Poduri, et al., 2012; Poduri, et al., 2013). At this range, conventional SCNA detection methods can be seriously compromised because their key evidence (read-depth discrepancy in disease-control pair) is obscured; copy number changes occur only in a subset to make only an insignificant difference in overall read-depth (Fig. 1b). Consequently, these subtle read-depth changes are hardly distinguished from natural variance. Second, matched control samples (e.g. blood, saliva, or normal tissue) are not always available. Genomic study of brain diseases is one typical case; because brain tissues are usually collected posthumously, control samples are no more available once they are missed in the initial design. For example, only 22% of brain non-tumor (and 72% of brain tumor) whole-genome samples deposited in the sequence read archive (SRA) contain matched-control information (Supplementary Table S1). Although less frequent, there is still lack of matched control in general cancer data including 59 case-only studies (2,799 samples) archived in the database of genotypes and phenotypes (dbGaP). This absence of control samples prohibits the use of current algorithms for SCNA detection.

We found the two critical problems, low mutational frequency and absence of control, turn to be a rather solvable probabilistic problem when considered simultaneously. A germline mutation is expected to be observed in the entire population of the sample. So the expected allele frequency of a germline mutation is 50% for heterozygous and 100% for homozygous events. And we expect a broadly shared somatic mutation would represent a similar frequency to germline mutations. On the contrary, if a somatic muta-

*To whom correspondence should be addressed.

tion exists only in a sub-population (e.g. 10% of the sample), the allele frequency of its heterozygous mutation is expected to be much lower (e.g. 5%); these frequencies are hardly expected in germline mutations according to the binomial probability (that one haploid is dominantly sequenced by chance) and can be separated without control.

If we can catalog all existing somatic and germline mutations in the sample, the overall allele frequency distribution can be utilized to de-convolute somatic mutations in a subpopulation from the others. However, the classification of somatic mutation among the all cataloged mutations is also dependent upon the assumed sub-population composition; for example a 1% allele frequency is generally a noise, but can be a successful somatic mutation candidate with a 2% sized sub-population. So an efficient inference should be applied to find the optimal list of somatic mutations with proper composition that maximizes the observation probability. Another problem is interpreting the deviations in read-depth to a deletion call and its frequency. Deletions in rare populations (<5%) are difficult to be distinguished from noises. Likewise, rich populations (>90%) are almost germline-like and hardly separated.

Here, we introduce a novel computational method SoloDel (**So**matic **Lo**w-frequent **Del**etion call model) to find low-frequent somatic deletions with and without matched control data. We first construct a probabilistic model of somatic mutation progression to represent the composition of two subpopulation groups, normal and mutated cells. Two types of model parameters, a rate of somatically affected haploid p_s (=a half of mutated subpopulation size assuming that all somatic deletions are heterozygous) and a proportion of somatic to total deletion λ_s are defined for somatic deletions to show the status of the mutated cell population and the number of somatic deletions. We then fit the constructed model for a given sample with the assumption that the sample consists of normal cells and mutated cells. Optimal model parameters that maximize the observation probability are estimated by applying the expectation-maximization (EM) algorithms for Gaussian mixture model, which represent the mixed allele frequencies of deletions from normal and mutated cells. We accept the mixed model only if the likelihood of the estimated model is higher than that of a simple germline model (subpopulation free). Final somatic deletion candidates that show higher probability to be somatic are selected based on the estimated model.

Using simulated data with various conditions, we demonstrated that SoloDel accurately estimates the mutated subpopulation size with the average error rate of 1%. False rejection of mixture model only happened in near-germline samples. With the combined use of external structural variation caller, we also showed that SoloDel greatly improves the accuracy in terms of overall precision and recall in the entire range of subpopulation size and total number of somatic deletions (100-1000). Moreover, the performance remained consistent without matched control, in a wide range of subpopulation size (10%-70%). Finally, we could successfully recall experimentally validated low-frequent somatic deletions from neuropsychiatric disease samples. This implies that low-frequent somatic deletions can be also recruited using our *in silico* probabilistic model to alleviate high-priced and laborious deletion searching procedures including nested-PCR and single cell sequencing.

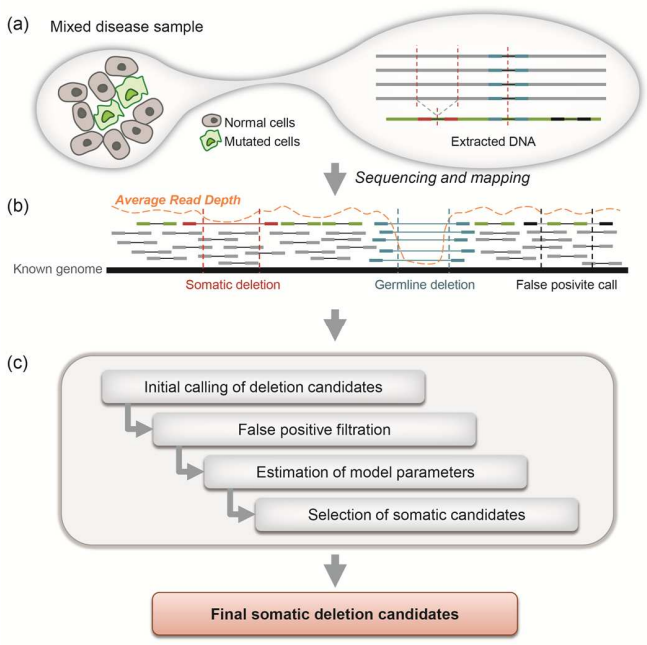

Fig. 1. Overall SoloDel workflow. (a) The mixed disease sample contains two different types of DNA from normal and mutated cells (gray and green lines). Red vertical dotted lines represent break points of deletions. (b) The difference of read-depth hardly shows somatic deletions due to the variance of average depth. Anomaly mapped reads can be generated by three different conditions: (i) somatic deletions (red reads), (ii) germline deletions (blue reads), or (iii) false positive calls (black reads). (c) A generated BAM file follows four main processes. Final somatic deletion candidates with probability scores are provided as a result.

2 METHODS

2.1 Method outline

Overall workflow of SoloDel is shown in Figure 1. The input is sequenced paired-reads from a mixed disease sample, which contains subpopulations of mutated cells. General preprocessing steps including read mapping to the reference genome, sorting, removing duplicate reads, and indexing are required. A resulted BAM file is fed to the following processes.

The first step is an initial calling of deletion candidates. A called candidate can be classified as (i) a germline deletion, (ii) a somatic deletion, or (iii) a false positive. False positives are predicted and filtered out based on the sequence homology in the subsequent step. Somatic deletions are further selected using a somatic progression model with their parameters estimated based on the observed value of called candidates. By using the estimated model, candidates that have higher probability to be somatic than germline are considered to be our final callset.

2.2 Initial calling of deletion candidates

Two major approaches, read-depth analysis (Abyzov, et al., 2011) and read-pair analysis (Chen, et al., 2009) have been generally applied for germline deletion calling from NGS data. Unlike germline deletions, particularly for non-tumor diseases, somatic deletions frequently exist in a small fraction of a cell population. In this case, the performance of read-depth analysis can be seriously compromised due to the insufficient read-depth deviation that is easily obscured by the sequencing and mapping variances (Kim, et al., 2013).

Most of the available read-depth based tools mainly target somatic deletions in cancer (Boeva, et al., 2012; Boeva, et al., 2011; Koboldt, et al., 2012; Krishnan, et al., 2012; Xi, et al., 2010) leaving the aforementioned problem where no clear decline of read-depth is observed. However, a number of anomaly (discordantly) mapped paired-end reads can be obtained with a sufficient sequencing depth to indicate the presence of somatic deletions (Fig. 1b). Therefore, we only used the read-pair analysis method in our initial calling step to minimize potential false negatives that come from the limited read-depth resolution. Of the available read-pair based tool, we applied BreakDancer (Chen, et al., 2009) and DELLY (Rausch, et al., 2012) in SoloDel for their accuracy of breakpoint mapping and popularity of usage.

2.3 Filtration of false positive candidates

Repetitive DNA sequences occasionally disguise the paired-end read mapping to generate false deletion calls (Treangen and Salzberg, 2012). To reduce the false positive calls, we defined the following procedure that assesses the possibility of a deletion call is resulted from sequence homology using Blat alignment tool (Kent, 2002).

Assume that two ends of a paired-end read $p=\{r_1, r_2\}$ were mapped to genomic locations g_1 and g_2 ($g_1 < g_2$) respectively to call a deletion. The gap between g_1 and g_2 is abnormally large, by definition, compared to the expected insert size i of the sequencing data. We captured two read-length sized reference sequences $s_1=[g_1+i, g_1+i+\text{len}(r_1)]$ and $s_2=[g_2-i-\text{len}(r_2), g_2-i]$, which denote the possible alternative mapping sites that make p revert to a concordant read (Supplementary Fig.S1). We compare the captured sequences s_1 and s_2 to their corresponding read sequences r_2 and r_1 to check if they match exactly and continuously in >90% of the regions. Finally, deletion calls that captured sequences are matched to their alternative sites (s_1 matches r_2 or s_2 matches r_1) are filtered out. The remaining calls are considered as a mixture of high confident germline and somatic deletions.

2.4 Somatic deletion generation model

To represent the status of generated deletions of a given sample, we defined a probabilistic model for somatic deletion generation (Fig. 2). In the development of normal tissue, precursor cells are replicated and generate daughter cells that possess identical composition of germline deletions of precursor cells. Those inherited germline deletions are denoted by D_g. However, if a somatic event is occurred during the development, somatic deletions are generated and affect to subordinate cells. These affected cells additionally possess the set of somatic deletions, denoted by D_s.

A somatic event affects the constitution of somatic deletions in two ways. First, the time of a somatic event decides the rate of deletion-harboring haploid, defined as $\mathbf{p} = \{p_g, p_s\}$. The population of the mutated cell that contains somatic deletions is represented by this parameter. Second, the scale of a somatic event decides the proportion of deletion loci, defined as $\boldsymbol{\lambda} = \{\lambda_g, \lambda_s\}$. The number of generated deletions is represented by this parameter.

All cells possess identical composition of germline deletions. For germline deletions, the expected p_g is 0.5 and 1.0 for heterozygous and homozygous deletions, respectively. We hypothesized that the probability of somatic homozygous deletions (two independent somatic deletions occur at the exactly same genomic loci) is ignorable. Therefore, the rate of somatically affected haploid, p_s, is bounded between 0 and 0.5.

The number of heterozygotic germline deletion, N_g, and somatic deletion, N_s, is depicted by the blue and red lines in the genome (Fig. 2). The proportion of germline and somatic deletion loci, λ_g and λ_s, are calculated from each count divided by the total number of deletion loci, N_g+N_s. Through the defined parameters, $\mathbf{p}$ and $\boldsymbol{\lambda}$, the status of generated deletions of a given sample can be represented.

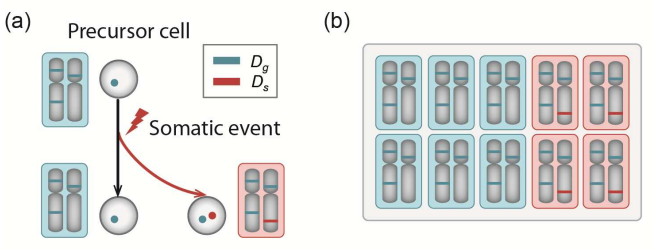

Fig. 2. Defined model of somatic deletion generation. (a) During the normal development process, daughter cells replicated from precursor cell inherit identical composition of germline deletions (blue dots and lines). However, subpopulation of cells can obtain somatic deletions (a red dot and line) by a somatic event during the development. (b) A somatic event affects the composition of somatic deletion into two ways in a fully developed tissue. The time of a somatic event determines the population of mutated cells (the number of cells with red backgrounds), and the scale of a somatic event determines the number of generated somatic deletions (the number of red lines in the genome). Both aspects are reflected by the model parameters – a rate of somatically affected haploid (p_s) and a proportion of somatic to total deletion (λ_s). The figure shows an example of a mixed tissue with p_s=0.2 (mutated subpopulation size is 0.4) and λ_s =0.25.

2.5 Probabilistic estimation of deletion candidates with observed values

For each deletion candidate d_i that passed the sequence homology filter, we can calculate the reduced depth of predicted region (x_i) and the total depth of flanking region (n_i) (Fig. 3a). To select somatic deletion candidates, classification of true somatic deletions from germline heterozygous deletions is indispensible.

For given observed values n_i and x_i, we can consider each value as the sequence of independent trials and the number of successes of the binomial distribution. Then, probability of a deletion with observed values can be estimated by the binomial probability mass function:

$$B(x_i; n_i, p) = \binom{n_i}{x_i} p^{x_i} (1-p)^{n_i - x_i} \tag{1}$$

As d_i can either be a germline heterozygous deletion or a somatic deletion, the success probability p of binomial distribution follows the rate of deletion-harboring haploid, p_g and p_s, for each case. The probability that d_i is from the germline heterozygous deletions or the somatic deletions follows the ratio between the number of germline heterozygous deletions and somatic deletions, λ_g and λ_s. Therefore, the probability of a deletion for given observed values n_i and x_i can be obtained as follows.

$$\begin{aligned} P_{d_i} &= P(x_i \cap D_g) + P(x_i \cap D_s) \\ &= P(x_i \mid D_g) \cdot P(D_g) + P(x_i \mid D_s) \cdot P(D_s) \\ &= B(x_i; n_i, p_g) \cdot \lambda_g + B(x_i; n_i, p_s) \cdot \lambda_s \end{aligned} \tag{2}$$

Given $\theta = \{\mathbf{p}, \boldsymbol{\lambda}\}$, the likelihood function is derived by the product of probabilities of independent deletion candidates with observed values.

$$L(\theta \mid X) = P_\theta(X) = \prod_i \left(\sum_{j \in \{g,s\}} B(x_i; n_i, p_j) \cdot \lambda_j \right) \tag{3}$$

The maximum likelihood estimator (MLE) of $\theta = \{\mathbf{p}, \boldsymbol{\lambda}\}$ can be obtained as follows with the constraints:

$$\hat{\theta} = \arg\max_{\theta} L(\theta \mid X)$$
$$\begin{cases} 0 \le p_s \le 0.5 \\ p_g = 0.5 \\ \lambda_s + \lambda_g = 1 \end{cases} \quad (4)$$

2.6 Estimation of model parameters and selection of somatic candidates

For the estimation of **p** and **λ**, we applied the EM algorithm based on the Gaussian mixture model (Fig. 3b). For each deletion candidate d_i, we first calculated the point estimates $\hat{p}_i$ for the rate of deletion-harboring haploid by dividing observed values. Candidates from germline homozygous deletions that show $\hat{p}_i$ close to 1 are filtered out.

$$\hat{p}_i = x_i / n_i \quad (5)$$

Since we assumed binomial trials for the observed values n_i and x_i, each trial follows the Bernoulli distribution with success probability p. The p can be either p_g or p_s depending on the origin of a deletion, then the point estimate $\hat{p}_i$ is the sample mean of given Bernoulli trials. Therefore, distribution of the point estimates $\hat{p}_i$ for all deletion candidates will be approximately Gaussian by the central limit theorem.

We hypothesized that the point estimates $\hat{p}_i$ are derived from the Gaussian mixture model of two groups, somatic and germline heterozygous deletions.

$$p(x) = \lambda_s N(x \mid p_s, \sigma_s^2) + \lambda_g N(x \mid p_g, \sigma_g^2) \quad (6)$$

To estimate the parameters of Gaussian mixture model, we applied the EM algorithm with the number of components k=2. Estimated mixture model parameters directly represent the parameters of deletion model (**p**, **λ**). Using the estimated parameters, likelihoods of mixture and germline models are calculated as follow:

$$\begin{cases} L(\theta_{mixed} \mid X) = \prod_i \left(\sum_{j \in \{g,s\}} B(x_i; n_i, p_j) \cdot \lambda_j \right) \\ L(\theta_g \mid X) = \prod_i B(x_i; n_i, p_g) \end{cases} \quad (7)$$

If the likelihood of germline model is higher than that of mixture model, we *reject* the mixture model and consider all candidates as germline heterozygous deletions. If the likelihood of mixture model is higher, then we *accept* the mixture model and calculate the somatic candidate score (S_i):

$$S_i = \log\left(\frac{B(x_i; n_i, p_s) \cdot \lambda_s}{B(x_i; n_i, p_g) \cdot \lambda_g} \right) \quad (8)$$

Deletion candidates with $S_i > 0$ are selected as final somatic deletion candidates. In other words, the most likely explanation for the origin of these deletions is a somatic subpopulation, not a highly deviated sequencing from a uniform sample.

Overall processes in SoloDel were implemented as a Java-based single program utilizing open source libraries including Picard, SAMtools (Li, et al., 2009), and other libraries for advanced mathematics.

2.7 Generation of simulation data

To test the performance of SoloDel, we generated multiple sets of simulated data with widely varying core parameters. Datasets were created with various ranges of mutated subpopulation size and the number of generated somatic deletions, which correspond to model parameters p_s and λ_s, respectively. Subpopulation size was varied from 0% to 100% with 10% intervals, and the number of generated somatic deletions was varied from 100 to 1000 with 100 intervals. Each number of homozygous and heterozygous germline deletions was fixed to 1000. A total of 110 simulation datasets were designed to be generated based on the combination of two variables.

For each simulation dataset, we generated two set of diploid genomes, one for normal genome that only contains germline deletions and another for mutated genome with germline and somatic deletions together. All simulated genomes were derived from the reference genome of human chromosome 1 (hg19) by introducing random deletions. The size of introduced deletions was randomly selected between 500bp to 10kb. Generation of simulated genome was implemented by Python.

GemSim (McElroy, et al., 2012) was used to generate paired-end reads from the simulated genomes based on the Illumina paired-end error model. The number of required paired-reads was calculated based on the planned mean coverage (70x). A total of 86,000,000 paired-reads were generated for each simulation dataset. Among them, the relative abundance of the mutated genome was calculated according to each assigned subpopulation

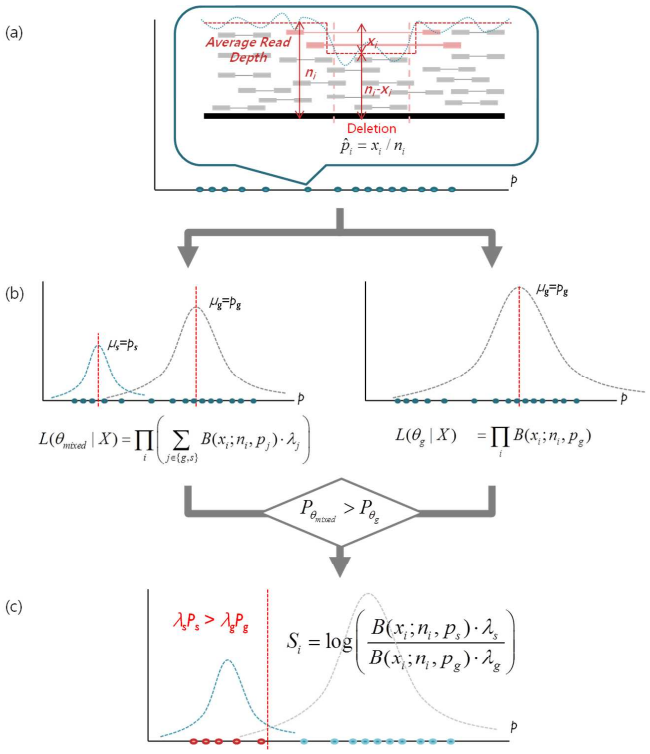

Fig. 3. Procedures for estimation of model parameters and selection of somatic candidates. (a) For each deletion d_i, the point estimate $\hat{p}_i$ for the rate of deletion-harboring haploid is calculated based on the number of mapped read (n_i, x_i). (b) The point estimates $\hat{p}_i$ (blue dots) are hypothesized to be derived from the Gaussian mixture model of two groups, somatic and germline heterozygous deletions (left). The mean value and the weight of each Gaussian distribution indicate p and λ, respectively. Estimation of the parameters is performed by the EM algorithm. Based on the estimated parameters, likelihoods of the mixture model (left) and the germline model (right) are compared. If the likelihood of mixture model is higher, somatic scores (S_i) are calculated. (c) Deletion candidates that $S_i > 0$, which indicate higher probability to be somatic than germline, are selected as final candidates (red dots).

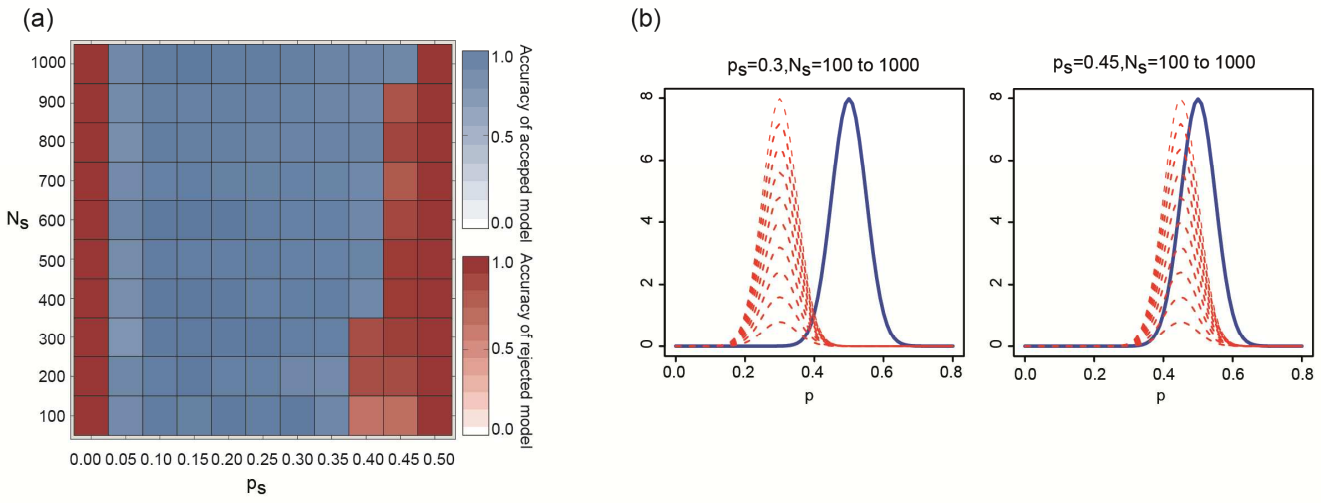

Fig. 4. Estimated results of model parameters. (a) Each cell of the heat map represents a specific type of simulation dataset. The color of each cell indicates the acceptance of mixture model (blue for acceptance, red for rejection), and the gradient represents the accuracy of the estimated parameter p_s. The gradient is set to zero and shown as white color if the deviation between estimated and true p_s is above 0.05. The maximum deviation of estimated p_s was 0.01688, which shows robust performance of parameter estimation. (b) Distribution of somatic and germline deletions are plotted as red and blue lines for the cases with p_s=0.3 and 0.45. Compared to datasets with small subpopulation size (left), most somatic distributions of datasets with large subpopulation size (right) are covered with the germline distribution, which cause the false rejection of mixture model. With the enough number of generated somatic deletions, germline and somatic distributions from the large subpopulation size are also clearly classified by the mixture model of SoloDel.

size. All generated paired-end reads were mapped using Burrows-Wheeler Aligner (BWA) (Li and Durbin, 2010), and merged into one alignment (BAM) file for each dataset.

Merging, sorting, removing duplicates, and indexing of alignment (BAM) file were performed using SAMtools software package (Li, et al., 2009). The resulting BAM files for each dataset were tested to SoloDel and several previous SCNA detection algorithms (Boeva, et al., 2012; Boeva, et al., 2011; Rausch, et al., 2012; Xi, et al., 2010). For the performance measure of SoloDel, two different external initial callers BreakDancer (Chen, et al., 2009) and DELLY (Rausch, et al., 2012) were tested to avoid the performance bias by initial callers. The performances of all methods were compared in precision, recall, and F-score.

3 RESULTS

3.1 Estimated results of model parameters

We first examined the accuracy of parameter estimation for our probabilistic model with simulation data. Estimation results for all simulation datasets from SoloDel with BreakDancer are visualized in the heat map to represent the performance (Fig. 4a). Each cell of the heat map indicates one simulation subtype defined by a combination of p_s and λ_s. To evaluate the accuracy of the acceptance/rejection call for the mixture model, we generated a negative test set from a mutation-less cell population (0%, p_s = 0) and a fully mutated cell population (100%, p_s = 0.5) with no matched control provided; these are the cases for the germline model and should be rejected for the mixture model.

A perfect model should accept the mixture model for all datasets except the negative sets (p_s = 0 and p_s = 0.5). We found SoloDel successfully rejects the mixture model for the negative cases with only reporting a few false rejections at the near germline area (p_s = 0.4 to 0.45). To clarify the false rejection of mixture samples, we plotted the mis-rejected Gaussian mixture distributions (Fig. 4b). With the low numbers of generated somatic deletions (N_s), most somatic distributions were covered with the large germline distribution due to the small difference of mean values between two distributions. Although the mutated subpopulation size is large (p_s = 0.4 and p_s = 0.45), germline and somatic distributions were clearly classified when the dataset possesses a large enough number of somatic deletions. For datasets with small subpopulation size, which are in the main hotspot for SoloDel, mixture models were accepted perfectly for all cases.

The gradient of each heat map cell represents the accuracy of the estimated parameter p_s. As the minimal deviation of true p_s between generated datasets is 0.05, accuracy for cases that deviations between true and estimated p_s above 0.05 are assigned to zero and depicted as white color. If the estimated p_s is exactly matched to the true p_s, the accuracy is assigned to 1. The average of the absolute deviation between true and estimated p_s was less than 0.005 (1% of the mutated subpopulation size), representing the robustness of the parameter estimation of SoloDel. Even with the false rejected datasets, estimations of p_s were successfully achieved regardless of the acceptance of mixture model. Estimated parameters (p_s, λ_s) with true values are shown in Supplementary Table S4. SoloDel with DELLY also showed similar results, which represent the unbiased performance of parameter estimation (Supplementary Fig.S2).

3.2 Performance tests of somatic deletion calling

To evaluate the performance of somatic deletion calling, we ran SoloDel and previous SCNA detection methods to all simulation datasets and predicted somatic deletion candidates for each data. We referred to a recent evaluation study (Alkodsi, et al., 2014) to

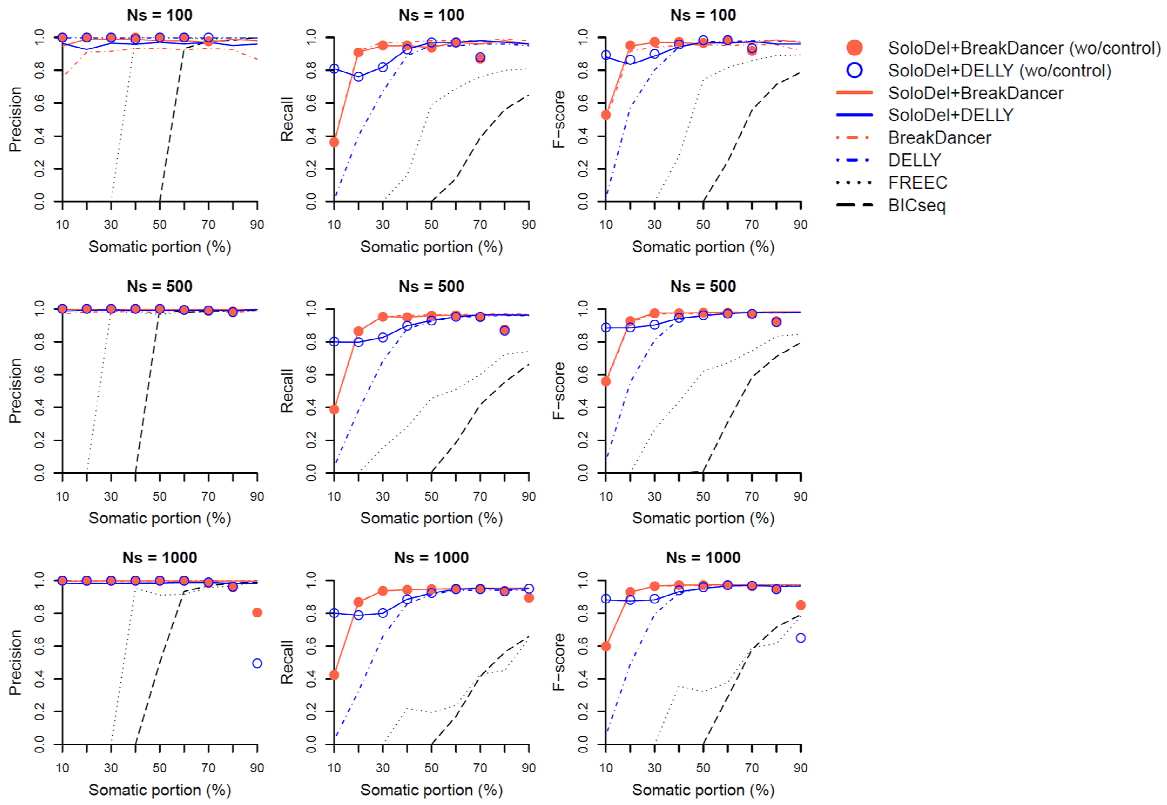

Fig. 5. Performance comparison with the previous methods for somatic deletion detection. Precision, recall, and F-score are depicted for all tested methods with the datasets from three different numbers of somatic deletions (100, 500, and 1000). Performance measurements with matched control data are represented by lines with different colors and shapes for each tested method. Regardless of the type of external initial caller, SoloDel (blue and orange solid lines) outperformed other tools for the most simulated datasets. Performances of SoloDel without matched control data are represented by circles for each initial caller, and showed comparable results to performances with matched control data except for the datasets at a near-germline area. For datasets with small subpopulation size, SoloDel outperformed other tools even without matched control data (blue and orange circles).

select high performing previous methods for comparison including BIC-seq (Xi, et al., 2010) and Control-FreeC (Boeva, et al., 2012; Boeva, et al., 2011). Other high-ranked tools from the referred review have been considered but finally excluded from the comparison; VarScan2 (Koboldt, et al., 2012) and COPS (Krishnan, et al., 2012) were tested but had too low accuracy due to their over-segmentation, and SegSeq (Chiang, et al., 2009) only works on single-end reads. Besides the reviewed tools, BreakDancer and DELLY were also included according to their stated availability of SCNA calling with matched control data. In a test, we considered a result is correct if the genomic coordination of predicted somatic deletion reciprocally overlapped the corresponding true somatic deletion in more than 50%. Based on this criterion, precision, recall, and F-scores were measured. All deletions generated in simulation have been hidden and reserved.

We note two major advantages of SoloDel compared to previous methods: 1) accurate separation of low-frequent somatic deletions and 2) discrimination of somatic deletions from unmatched sequencing data. Since none of the previous methods support the latter, SoloDel's performance in this condition is unrivaled. Therefore, we divided the evaluation process in two steps. First, we used simulation dataset *with* matched control to compare the performances between previous methods and SoloDel (lines in Fig. 5 and Supplementary Fig. S3). And second, we assessed the performances of SoloDel *without* matched control data to measure its robustness (circles in Fig. 5 and Supplementary Fig. S3).

In the test on dataset *with* matched control, SoloDel with read-pair methods outperformed other tools in most conditions (Fig. 5, blue and orange solid lines). As expected, read-depth based methods (BIC-seq and Control-FreeC) hardly detected somatic deletions with small subpopulation size (<30%), even after applying an inherent contamination rate estimation, due to the insufficient read-depth deviations. DELLY (without SoloDel) kept high precision along with the change of subpopulation size, however, a significant drop in recall was observed with small subpopulation size (<30%). We assume that the loss of sensitivity resulted from the absence of subpopulation information (mixture rate) to cause a false filtration of true somatic deletions with a small number of supporting reads. When DELLY was combined with SoloDel, the problem has been successfully addressed (blue lines in Fig. 5). BreakDancer showed almost similar performances to those with SoloDel, but was slightly lower in precision due to a few false positive calls. Application

of SoloDel successfully removed those noises and increased precision (orange lines in Fig. 5). We also confirmed the consistent outperformance of SoloDel regardless of the generated deletion size (Supplementary Fig.S4 and S5).

Finally, we found that SoloDel *without* matched control (Fig 5, blue and orange circles) performed almost comparably (to *with* matched-control) except only at high somatic portions (>80%). Again, note that separation of somatic from germline mutation without control becomes theoretically intractable as being close to 100%. From this point of view, SoloDel's initial rejection of mixture model can be the best estimate of the sample composition. Moreover, when a sufficient number of deletions exist in a sample (Fig 5, below), SoloDel was able to rescue most of true somatic deletions with high precision even at a near-germline sample (>80% somatic portion).

A few recent studies have reported potential somatic mutations from unmatched data by applying a naïve threshold of allele frequency (Jamuar, et al., 2014; Lim, et al., 2015). Although target genes were considerably limited and intensive experimental validations were followed in previous studies, such threshold-based classification possesses several inherent problems. First, an arbitrarily selected threshold does not have logical basis to support its validity, and second, compulsive application of an arbitrary threshold can cause high rate of false positives, especially critical to mutation-less samples. We confirmed superior overall performance of SoloDel compared to the naïve threshold-based method for unmatched data (Supplementary Fig.S6). Besides the outperformance, SoloDel can overcome those problems by providing rationale for threshold estimation based on the somatic progression model and mixture model examination with likelihood comparison. Taken all together, we demonstrated that SoloDel accurately identifies somatic deletions of various mutational frequencies (subpopulation size and total deletion number) and availability of matched control based on the rational estimation processes.

3.3 Performance tests on simulated data with multiple subpopulations

In SoloDel, the estimation of mixture model was performed based on the assumption that the data is derived from two independent Gaussian distributions. However, in the real disease samples, the generation of multiple subpopulations has been continuously reported (Ding, et al., 2012; Gerlinger, et al., 2012; Greaves and Maley, 2012). We thus tested the performance of SoloDel with the data comprised of multiple subpopulations.

We constructed simulation data with multiple subpopulations based on the real tumor case that previously reported (Ding, et al., 2012). In the previous work, authors found the clonal evolution patterns that consist of four subclones in the tumor sample and measured their relative abundance, by analyzing mutant allele frequencies of somatic point mutations from whole genome sequencing data. We reproduced this clonal evolution patterns with measured abundances of subclones to the simulation dataset by introducing somatic deletions (Fig. 6a and Table 1). The number of introduced deletions was decided based on the rate of validated mutations for each subclone in the previous work. Performances of BIC-seq, Control-FreeC, DELLY, and BreakDancer with matched control data were compared to that of SoloDel without matched control for this multiple subclone data.

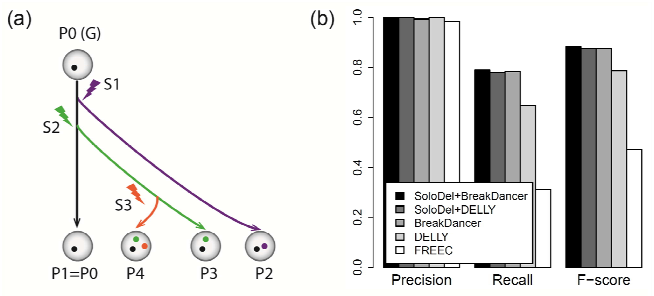

Fig. 6. Performance comparison of somatic deletion detection with multiple subclone data. (a) Four subclones are introduced in the simulation data by three times of subsequent somatic events. Each dot corresponds to the somatic deletions from the specific somatic events. (b) Compared with Control-FreeC, BreakDancer, and DELLY, SoloDel showed better performances for all except one measurement (recall of BreakDancer) regardless of the type of initial caller.

Table 1. Composition information of subclones in the simulation data

Group	P1	P2	P3	P4
Composition of deletion	G	G+S1	G+S2	G+S2+S3
N_g	450	450	450	450
N_s	0	250	150	300
p	0.5	0.2656	0.1707	0.0255

Population rates of subclones (p) were obtained from Ding *et al.* (2012)

BIC-seq failed to report any somatic deletions from every subclones. Although the data consist of multiple subpopulations, SoloDel showed acceptable performance with the estimation based on the mixture of two subgroups (somatic and germline). In the performance comparison, overall measurements of SoloDel outperformed those of Control-FreeC and DELLY regardless of the type of initial caller (Fig. 6b). Although Control-FreeC showed good performance in precision, most somatic deletions from subclone P3 and P4 were missed. DELLY detected substantial number of somatic deletions from P3, but lost all deletions from P4. Combined with SoloDel, DELLY recovered a part of somatic deletions from subclone P4 that consists of less than 5% of cell population. This result indicates the importance of appropriate model estimation for the detection of somatic deletion with small subpopulation size. Only BreakDancer showed comparable results, but performances of SoloDel were achieved without matched control information. With the multiple subclone data, we confirmed again the competitiveness of SoloDel compared to previous methods.

3.4 Somatic deletion calling on neuropsychiatric data

Detecting low-frequent somatic deletions in the real data is challenging problem. Most previous studies have searched and validated the existence of somatic deletions based on the SNP arrays, which cannot capture somatic deletions with small subpopulation size (Lee, et al., 2012; O'Huallachain, et al., 2012). A recent study

Table 2. Validated somatic deletions in the previous study and their classified results from SoloDel

ID	Chr	Start	End	Size	Estimated allele frequency ($\hat{p}_i$)	Classified result from SoloDel
C13	chr6	136641919	136642449	530	0.0062	somatic
C13	chr15	39652090	39652798	708	0.0778	somatic
C21	chr7	26214791	26217983	3192	0.2805	somatic
C21	chr12	102903681	102904789	1108	0.3718	germline
C16	chr7	6986591	6992106	5515	0.2494	somatic
C17	chr12	102903625	102904795	1170	0.2883	somatic

reports the detection of low-frequent somatic deletions in neuropsychiatric samples with the conservative approach based on the manual inspection (Kim, et al., 2014). Although they only focused on a small number of somatic candidates with high probability, a few of them were experimentally validated by Sanger sequencing and single cell sequencing of deletion breakpoints.

We applied SoloDel to the real whole-genome sequencing data from the previous study, comprised of two schizophrenic brain and two normal brain data with the absence of matched control information. Mixture models were accepted for all samples, indicating the presence of mutated subpopulation with somatic deletions. A total number of 287 and 238 somatic deletion candidates for normal brains and 382 and 346 candidates for schizophrenic brains were called by SoloDel, compared to the predicted number of 29 and 18 candidates for normal brains and 15 and 18 candidates for schizophrenic brains in the previous study. Although called results of SoloDel may include a certain amount of false positives, much higher number of calls than previous study was expected according to the declared risk of excessive false negatives due to the intensive minimization of false positive rate (Kim, et al., 2014). Differed from the result of previous work, more number of somatic deletion candidates were called from the schizophrenic cases, which may suggest the relationship between the number of occurred somatic deletion and the onset of disease.

We then examine the called candidates to confirm somatic deletions that were validated in the previous study. Table 2 shows the called results for the previously validated candidates. Among the six validated somatic deletions, five were successfully called as somatic by SoloDel. One exceptional candidate was classified as germline, however, mis-classification was due to the corrected penalty of high variance of read-depth at the predicted region of deletion. Validated somatic deletions covered a broad range of somatic portion, supporting unbiased confirmation of called results from SoloDel.

4 DISCUSSION

In this article, we developed a novel computational method SoloDel for detecting low-frequent somatic deletions with and without matched control samples. We defined a probabilistic model for somatic deletion generation and estimated model parameters using EM algorithms to provide the criteria of classifying somatic deletion candidates. Compared to previous methods, we showed outstanding performances of detecting low-frequent somatic deletions with various simulation datasets. With the real sequencing data, we successfully recalled experimentally validated somatic deletions confirmed by the previous study, supporting the reliable performance.

Recent studies have been increasingly reporting the contribution of low-frequent somatic mutations to the onset of various diseases including neuropsychiatric diseases (Lee, et al., 2012; Lim, et al., 2015; Poduri, et al., 2012; Poduri, et al., 2013; Shirley, et al., 2013). However, confirmed list of low-frequent somatic mutations are still limited due to the lack of detecting methods, insufficiency of matched control information, and difficulty of experimental validation. Advances of new techniques such as single cell sequencing and targeted ultra-depth sequencing will allow us to fill up the catalog of low-frequent somatic mutations. Decline of sequencing cost and growth of interest for the somatic analysis will increase the datasets with matched control and improve the quality of somatic call. With enough numbers of secured gold standards, the development of computational tools for accurate detection of somatic mutations will be followed.

There are several issues to enhance the detecting ability of SoloDel. To detect low-frequent somatic deletions as intended, enough coverage of whole genome sequencing data is indispensable. Initial calling of deletion candidates that completely depend on the external callers is another issue to improve. Consideration of polyploidy that affects the estimation of subpopulation size may also elevate the performance of SoloDel by using GC content normalization (Boeva, et al., 2011). In spite of the issues for improvement, SoloDel showed considerable results compared to the conventional methods even with the harsh conditions, which makes it applicable for identifying low-frequent somatic deletions of many challenging cases.

ACKNOWLEDGEMENTS

The authors acknowledge the support of neuropsychiatric sequencing data at Stanley Medical Research Institute.

Funding: This work was supported by the Bio-Synergy Research Project (NRF-2012M3A9C4048758) of the Ministry of Science, ICT and Future Planning through the National Research Foundation and a faculty research grant of Yonsei University College of Medicine for (6-2014-0067).

Conflict of Interest: none declared.

REFERENCES

Abyzov, A., et al. CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. Genome research 2011;21(6):974-984.

Alkodsi, A., Louhimo, R. and Hautaniemi, S. Comparative analysis of methods for identifying somatic copy number alterations from deep sequencing data. Briefings in bioinformatics 2014.

Beroukhim, R., et al. The landscape of somatic copy-number alteration across human cancers. Nature 2010;463(7283):899-905.

Boeva, V., et al. Control-FREEC: a tool for assessing copy number and allelic content using next-generation sequencing data. Bioinformatics 2012;28(3):423-425.

Boeva, V., et al. Control-free calling of copy number alterations in deep-sequencing data using GC-content normalization. Bioinformatics 2011;27(2):268-269.

Chen, K., et al. BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. Nature methods 2009;6(9):677-681.

Chiang, D.Y., et al. High-resolution mapping of copy-number alterations with massively parallel sequencing. Nature methods 2009;6(1):99-103.

Cibulskis, K., et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nature biotechnology 2013;31(3):213-219.

Ding, L., et al. Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. Nature 2012;481(7382):506-510.

Evrony, G.D., et al. Single-neuron sequencing analysis of L1 retrotransposition and somatic mutation in the human brain. Cell 2012;151(3):483-496.

Gerlinger, M., et al. Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. New England Journal of Medicine 2012;366(10):883-892.

Greaves, M. and Maley, C.C. Clonal evolution in cancer. Nature 2012;481(7381):306-313.

Gregor, A., et al. De novo mutations in the genome organizer CTCF cause intellectual disability. American journal of human genetics 2013;93(1):124-131.

Helman, E., et al. Somatic retrotransposition in human cancer revealed by whole-genome and exome sequencing. Genome research 2014.

Jamuar, S.S., et al. Somatic mutations in cerebral cortical malformations. New England Journal of Medicine 2014;371(8):733-743.

Kent, W.J. BLAT--the BLAST-like alignment tool. Genome research 2002;12(4):656-664.

Kim, J., et al. Somatic deletions implicated in functional diversity of brain cells of individuals with schizophrenia and unaffected controls. Scientific reports 2014;4:3807.

Kim, S., et al. Virmid: accurate detection of somatic mutations with sample impurity inference. Genome biology 2013;14(8):R90.

Koboldt, D.C., et al. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome research 2012;22(3):568-576.

Krishnan, N.M., et al. COPS: a sensitive and accurate tool for detecting somatic copy number alterations using short-read sequence data from paired samples. PloS one 2012;7(10):e47812.

Lee, E., et al. Landscape of somatic retrotransposition in human cancers. Science 2012;337(6097):967-971.

Lee, J.H., et al. De novo somatic mutations in components of the PI3K-AKT3-mTOR pathway cause hemimegalencephaly. Nature genetics 2012;44(8):941-945.

Li, H. and Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics 2010;26(5):589-595.

Li, H., et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics 2009;25(16):2078-2079.

Lim, J.S., et al. Brain somatic mutations in MTOR cause focal cortical dysplasia type II leading to intractable epilepsy. Nature medicine 2015;21(4):395-400.

Maher, C.A., et al. Transcriptome sequencing to detect gene fusions in cancer. Nature 2009;458(7234):97-101.

McElroy, K.E., Luciani, F. and Thomas, T. GemSIM: general, error-model based simulator of next-generation sequencing data. BMC genomics 2012;13:74.

Mermel, C.H., et al. GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. Genome biology 2011;12(4):R41.

O'Huallachain, M., et al. Extensive genetic variation in somatic human tissues. Proceedings of the National Academy of Sciences of the United States of America 2012;109(44):18018-18023.

Pflueger, D., et al. Discovery of non-ETS gene fusions in human prostate cancer using next-generation RNA sequencing. Genome research 2011;21(1):56-67.

Poduri, A., et al. Somatic activation of AKT3 causes hemispheric developmental brain malformations. Neuron 2012;74(1):41-48.

Poduri, A., et al. Somatic mutation, genomic variation, and neurological disease. Science 2013;341(6141):1237758.

Rausch, T., et al. DELLY: structural variant discovery by integrated paired-end and split-read analysis. Bioinformatics 2012;28(18):i333-i339.

Roth, A., et al. JointSNVMix: a probabilistic model for accurate detection of somatic mutations in normal/tumour paired next-generation sequencing data. Bioinformatics 2012;28(7):907-913.

Shirley, M.D., et al. Sturge-Weber syndrome and port-wine stains caused by somatic mutation in GNAQ. The New England journal of medicine 2013;368(21):1971-1979.

Treangen, T.J. and Salzberg, S.L. Repetitive DNA and next-generation sequencing: computational challenges and solutions. Nature reviews. Genetics 2012;13(1):36-46.

Xi, R., et al. BIC-seq: a fast algorithm for detection of copy number alterations based on high-throughput sequencing data. Genome biology 2010;11(Suppl 1):O10.