Published online 28 January 2017

Nucleic Acids Research, 2017, Vol. 45, No. 10 ***e76***
doi: 10.1093/nar/gkx024

MosaicHunter: accurate detection of postzygotic single-nucleotide mosaicism through next-generation sequencing of unpaired, trio, and paired samples

August Yue Huang[1,2,†], Zheng Zhang[1,3,†], Adam Yongxin Ye[1,4,5,†], Yanmei Dou[1,2,†], Linlin Yan[1], Xiaoxu Yang[1], Yuehua Zhang[6] and Liping Wei[1,*]

[1]Center for Bioinformatics, State Key Laboratory of Protein and Plant Gene Research, School of Life Sciences, Peking University, Beijing 100871, People's Republic of China, [2]National Institute of Biological Sciences, Beijing 102206, People's Republic of China, [3]School of Life Sciences, Tsinghua-Peking Joint Center for Life Sciences, Tsinghua University, Beijing 100084, People's Republic of China, [4]Peking-Tsinghua Center for Life Sciences, Beijing, People's Republic of China, [5]Academy for Advanced Interdisciplinary Studies, Peking University, Beijing, People's Republic of China and [6]Peking University First Hospital, Peking University, Beijing 100034, People's Republic of China

Received August 18, 2016; Revised December 24, 2016; Editorial Decision December 31, 2016; Accepted January 26, 2017

ABSTRACT

Genomic mosaicism arising from postzygotic mutations has long been associated with cancer and more recently with non-cancer diseases. It has also been detected in healthy individuals including healthy parents of children affected with genetic disorders, highlighting its critical role in the origin of genetic mutations. However, most existing software for the genome-wide identification of single-nucleotide mosaicisms (SNMs) requires a paired control tissue obtained from the same individual which is often unavailable for non-cancer individuals and sometimes missing in cancer studies. Here, we present MosaicHunter (http://mosaichunter.cbi.pku.edu.cn), a bioinformatics tool that can identify SNMs in whole-genome and whole-exome sequencing data of unpaired samples without matched controls using Bayesian genotypers. We evaluate the accuracy of MosaicHunter on both simulated and real data and demonstrate that it has improved performance compared with other somatic mutation callers. We further demonstrate that incorporating sequencing data of the parents can be an effective approach to significantly improve the accuracy of detecting SNMs in an individual when a matched control sample is unavailable. Finally, MosaicHunter also has a paired mode that can take advantage of matched control samples when available, making it a useful tool for detecting SNMs in both non-cancer and cancer studies.

INTRODUCTION

Genomic mosaicism describes the presence of cells with varied genomes within one individual (1,2). Mutations may escape the DNA repair system during postzygotic cell divisions in early development and aging and lead to genomic mosaicism at multiple scales including substitutions and indels of only a few base pairs, gains and losses of copy number, and large-scale chromosomal alterations (3–7). Although the exact occurrence rates of the different types of postzygotic mutations in the general population remain largely unknown, previous studies concerning cancer somatic mutations (8) and familial *de novo* mutations (9) have demonstrated that single-nucleotide substitution is the most predominant type of DNA alteration. The mosaicism of single-nucleotide substitutions (single-nucleotide mosaicism, or SNM) has long been known to play critical roles in many types of cancer (10,11). In recent years, an increasing number of non-cancer diseases have been identified as resulting from SNMs (12–14). In addition, SNMs have been detected in clinically unremarkable individuals (6) and shown to accumulate in multiple types of cells under natural selection (15–17). Certain mutant alleles of the mosaic sites in healthy parents might be transmitted to offspring and lead to severe genetic diseases (6,18,19). Thus the importance of genomic mosaicism has been increasingly recognized in human genetics research in the study of the etiology of non-cancer genetic disorders as well as in the origin and transmission of genetic mutations.

Taking advantage of next-generation sequencing (NGS) platforms, many algorithms have been developed to identify SNMs through comparisons of sequencing data between matched tissue pairs (20–23) and successfully used in many cancer studies (24,25). However, the application

*To whom correspondence should be addressed. Tel: +86 10 62755206; Fax: +86 10 62764970; Email: weilp@mail.cbi.pku.edu.cn
†These authors contributed equally to this work as first authors.

of these methods is unfeasible when paired control samples are unavailable, which is the case in most non-cancer studies and even in some cancer studies. Several other tools such as SNVer (26) and LoFreq (27) were developed to identify the presence of mutant alleles in pooled or individual NGS data, but they lacked the power to distinguish postzygotic SNMs from inherited heterozygous sites. Recently, Smith *et al.* presented SomVarIUS (28) as another control-free postzygotic mutation caller for tumor samples, but its performance has not been benchmarked in non-cancer samples which are expected to have significantly lower mutation rate. We had previously reported the first bioinformatics pipeline to identify SNMs from next-generation whole-genome sequencing (WGS) data of unpaired control-free samples from healthy non-cancer individuals, but it has the limitation of being not applicable to whole-exome sequencing (WES) data, not capable of incorporating related data when available, and being slow (6).

Compared with WGS, WES has the benefit of focusing on coding regions of the genome. However, compared with the binomial expectations that are generally accepted in analysis of WGS data, WES data show non-negligible capturing bias and over-dispersion in the distribution of alternative allele fractions (29), which obstruct the distinction of SNMs from heterozygous sites, and thus require a different statistical model. Although a few studies have reported the identification of clinically relevant SNMs from non-cancer WES data (30,31), their scope was limited to mutations in only one or a few candidate genes, and they could not be extended to an exome-wide search due to low specificity.

The main challenge of identifying SNMs in WGS or WES data of unpaired control-free non-cancer samples is achieving high enough precision and specificity to make validation feasible. Despite the power of NGS, the imperfections in NGS library preparation, base-calling and alignment introduce a large number of technical artifacts that are difficult to be distinguished from true mutations. This is known to cause a large number of false positives in the identification of cancer somatic mutations (22,32). Because the occurrence rates of postzygotic mutations in non-cancer individuals were about one to three orders of magnitude lower than that in cancer samples and that of germline variations (6,33,34), the large number of false positives poses even bigger challenges. For instance, attempts to use NGS to find postzygotic mutations between monozygotic twins often fail to validate the *in silico* candidate mutations (35,36). Reumers *et al.* managed to identify and validate true genetic differences from the whole-genomes of non-cancer monozygotic twin by employing 12 different types of error filters, however, the precision was as low as 0.25% (37). In order to elucidate the genomic patterns of postzygotic mutations in healthy individuals, it is necessary to gather a clean set of true positives. Thus, it is critical to significantly reduce the number of false positives to make experimental validations feasible, and hence methods that can identify postzygotic mutations with high precision and specificity are needed (32,38).

Here, we describe MosaicHunter, a Java-based computational tool for accurate identification of SNMs in both WGS and WES data without requiring matched control tissues, using Bayesian genotypers supplemented with a series of stringent error filters to remove systematic errors in NGS data to significantly reduce false positive rates. As it was shown in previous studies that incorporating sequencing data from parents may improve genotyping accuracy (39,40), we developed a new joint Bayesian model to incorporate parental data in SNM identification. We also developed a paired mode that can take advantage of matched control samples when available. The performance of MosaicHunter was benchmarked by both simulated and real sequencing datasets and compared with other existing methods, demonstrating the unique advantages of this method for accurately identifying SNMs in both cancer and non-cancer applications.

MATERIALS AND METHODS

Overview of MosaicHunter

The flowchart of MosaicHunter is shown in Figure 1A. In summary, MosaicHunter incorporated a Bayesian-based mosaic genotyper and a series of empirical error filters. The Bayesian genotyper was able to calculate the posterior probabilities of mosaic genotype and three germline genotypes (reference homozygous, heterozygous and alternative homozygous), by integrating base-calling errors, random sampling variations and population allele frequencies annotated in dbSNP. The error filters could further remove false positives caused by systematic errors in base-calling and read alignment as well as other types of genomic variants such as indels and structural variations. In addition to the binomial model for analyzing WGS data (6), the Bayesian genotyper provided a new beta-binomial model designed for WES data. MosaicHunter can run in 'single', 'trio' or 'paired' modes to handle sequencing data from unpaired, familial trio, or paired samples. The Bayesian genotyper for the 'single' mode of WGS data analysis was described previously. The other Bayesian models are described in detail below.

Beta-binomial model for WES data

To model the notable capturing bias and over-dispersion of alternative allele fractions in WES data, we changed the prior distribution of theoretical allele fraction θ for heterozygous sites, from the spike at $\theta = 0.5$, to a fitted beta distribution, i.e.

$$\theta|G = \text{heterozygous} \sim \text{Beta}(\alpha, \beta) \Leftrightarrow \text{P}(\theta|G = \text{heterozygous}) = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{\text{beta}(\alpha, \beta)}$$

Correspondingly, the inference part of heterozygous sites was subsequently changed to

$$\begin{aligned}\text{P}(r|G = \text{heterozygous}, d) &= \int_0^1 \theta^r (1-\theta)^{(d-r)} \text{P}(\theta|G = \text{heterozygous})\, \text{d}\theta \\ &= \frac{1}{\text{beta}(\alpha, \beta)} \int_0^1 \theta^{r+\alpha-1}(1-\theta)^{d-r+\beta-1}\, \text{d}\theta \\ &= \frac{\text{beta}(\alpha + r, \beta + d - r)}{\text{beta}(\alpha, \beta)}\end{aligned}$$

The prior beta distribution $\theta \sim \text{Beta}(\alpha, \beta)$ has properties

$$\begin{aligned}\text{E}[\theta] &= \frac{\alpha}{\alpha+\beta} \\ \text{Var}[\theta] &= \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}\end{aligned}$$

For the mutant allele count $r \sim \text{Binomial}(d; \theta)$, assuming that the beta prior of the theoretical mutant allele fraction

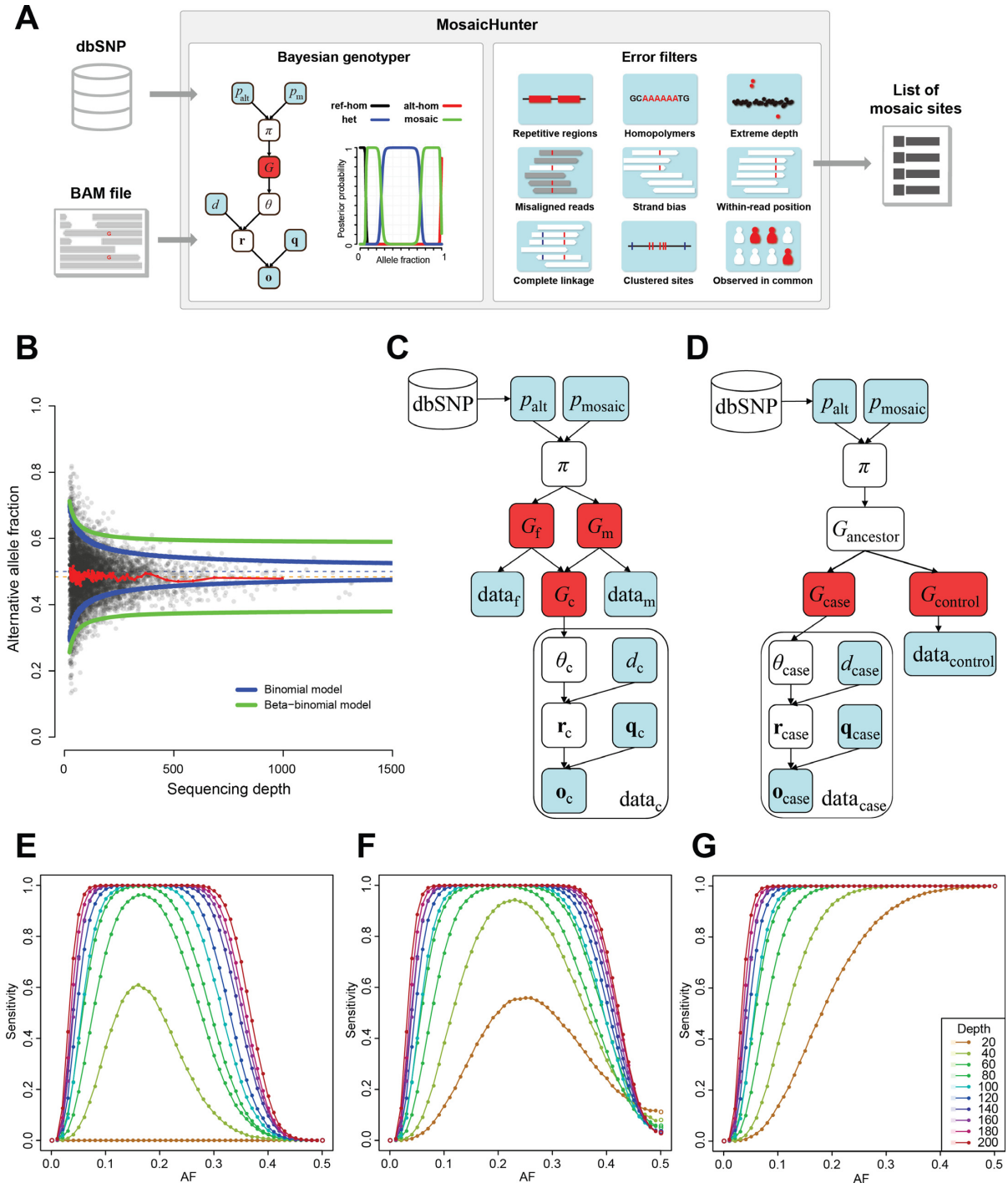

Figure 1. Overview of MosaicHunter. (**A**) Framework for the detection of single-nucleotide mosaicisms (SNMs) from a single unpaired sample. The candidate SNMs were first identified using a Bayesian genotyper and subsequently filtered using a series of stringent filters. ref-hom: homozygous for reference allele; alt-hom: homozygous for alternative allele; het: heterozygous. (**B**) The capturing bias and over-dispersion in WES data correction by the beta-binomial model. The blue and green lines represent the 95% confidence intervals of the original binomial and fitted beta-binomial models for heterozygous sites. The red line denotes the mean alternative allele fractions of heterozygous sites with varied depths which clearly deviated from 0.5. (**C** and **D**) The extended Bayesian models of MosaicHunter for incorporating familial trio data (C) and paired data from the same individual (D). In the Bayesian genotyper, *G* denotes the genotype state, π denotes the prior probabilities of each genotype, and d, **q**, **o** denote the depth, base qualities, and bases for calculating likelihood from observed sequencing data. (**E-G**) Depth-dependent sensitivity of the single (E), trio (F) and paired modes (G) of the Bayesian genotyper in MosaicHunter on simulated data.

θ does not change with the sequencing depth d, the expectation and variance of the mutant allele fraction r/d can be deduced as

$$\begin{aligned}\mathrm{E}[r/d] &= \mathrm{E}[\theta] \\ \mathrm{Var}[r/d] &= \frac{(d-1)\,\mathrm{Var}[\theta] + \mathrm{E}[\theta] - (\mathrm{E}[\theta])^2}{d} \\ &\approx \mathrm{Var}[\theta] + \frac{\mathrm{E}[\theta](1-\mathrm{E}[\theta])}{d} \quad (\text{when } d \gg 1)\end{aligned}$$

To estimate the parameters α and β, MosaicHunter first fitted the linear regression $\mathrm{Var}[r/d] \sim 1/d$ among the called heterozygous sites. For those sequencing depths which had <50 heterozygous sites, we grouped sites with similar sequencing depths. $\mathrm{E}[\theta]$ was estimated from the average observed allele fraction $\mathrm{E}[r/d]$, and $\mathrm{Var}[\theta]$ was subsequently calculated as the intercept of the line which had slope $\mathrm{E}[\theta](1 - \mathrm{E}[\theta])$ and intersected the fitted line at depth 80. The fitted α and β were directly calculated from $\mathrm{E}[\theta]$ and $\mathrm{Var}[\theta]$ according to the properties of beta distribution.

Because the capturing bias and over-dispersion varied among different runs of WES data, the estimation of α and β of the beta prior was performed for each exome.

To further reduce false positive SNMs in WES data, in addition to the series of empirical error filters which were described previously (6), we implemented and applied two more new filters: one is a multifactorial filter to distinguish systematic errors from true mutations based on their sequence context, strand bias and base quality (41), and the other is a mapping quality filter that rejects sites for which the mapping quality distributions between the major and minor alleles were significantly different (Wilcoxon rank-sum test's P-value < 0.05).

The Bayesian model for trio data

When the sequencing datasets of both parents and their child were available, the genotype relationship between these datasets could be modeled as a conditional probability factor $\mathrm{P}(G_c|G_f, G_m)$ and incorporated into the Bayesian model (Figure 1C), where G_c, G_f and G_m are the genotypes of the child, father and mother, respectively. Subsequently, the genotypes of the trio could be jointly inferred, as shown in the formula below:

$$\begin{aligned}&\mathrm{P}(G_c, G_f, G_m|\mathrm{data}_c, \mathrm{data}_f, \mathrm{data}_m) \\ &\propto \mathrm{P}(G_f)\,\mathrm{P}(G_m)\,\mathrm{P}(G_c|G_f, G_m)\mathrm{P}(\mathrm{data}_c|G_c)\,\mathrm{P}(\mathrm{data}_f|G_f)\,\mathrm{P}(\mathrm{data}_m|G_m)\end{aligned}$$

where the parental priors $\mathrm{P}(G_f)$ and $\mathrm{P}(G_m)$ and the likelihoods $\mathrm{P}(\mathrm{data}_c|G_c)$, $\mathrm{P}(\mathrm{data}_f|G_f)$ and $\mathrm{P}(\mathrm{data}_m|G_m)$ were calculated as previously described (6). The added trio factor $\mathrm{P}(G_c|G_f, G_m)$ was specified according to Mendelian inheritance rules and the occurrence of *de novo* and mosaic mutations. Here, we denote R_d as the *de novo* mutation rate and R_m as the mosaicism occurrence rate, set by default as 10^{-8} and 10^{-7}, respectively. The transmission of the genotypes from parents to their child could be formulated into two steps: (i) generating gamete alleles from the genotype of the parents; and (ii) combining two gamete alleles to form the genotype of the child.

The probability to generate gamete alleles from the genotype of each parent was specified as follows: (i) if the parent's genotype is homozygous, then the probability of the corresponding gamete allele is set to $1 - R_d$, and the probabilities for the other three alleles are set to $R_d/3$; (ii) if the parent's genotype is heterozygous, the probabilities of the two constitutional gamete alleles are set to $1/2 - R_d/3$, and the probabilities of the other two alleles are set to $R_d/3$; and (iii) if the parent's genotype is mosaic, then the probabilities of the major and minor gamete alleles are set to $\mathrm{AF}_{maj} \times (1 - 2R_d/3)$ and $\mathrm{AF}_{min} \times (1 - 2R_d/3)$, where AF_{maj} and AF_{min} denote the major and minor allele fractions of the mosaic site, and the probabilities of the other two alleles are still set to $R_d/3$.

The probability of the genotype of the child was specified as follows: (i) if the child's genotype is not mosaic and is exactly the combination of the gamete alleles from the two parents, then the probability is set to $1 - R_m$; (ii) if the child's genotype is not mosaic but not exactly the combination of the gamete alleles from the two parents, then the probability is set to zero; and (iii) if the child's genotype is mosaic, then the probability is set to R_m.

The joint posterior distribution of trio genotypes was calculated based on the Bayesian rule and marginalized to obtain the marginal posterior probabilities of genotypes for each individual. All sites with novel mutations (i.e. sites where the mutant allele was present in the child but absent in both parents) could be further genotyped as mosaic or *de novo* heterozygous based on the likelihood obtained from the sequencing data for the child. Because of the postzygotic origin of mosaicism, we only considered candidate sites where both parents were genotyped as homozygous for the major allele of the child's mosaicism.

The Bayesian model for paired data

When a paired control sample from the same individual was sequenced, we introduced a latent variable, the genotype of the ancestor, into the original Bayesian model and introduced the genotype change rate (Figure 1D). The inference formula can be formulated as

$$\begin{aligned}&\mathrm{P}(G_{case}, G_{control}|\mathrm{data}_{case}, \mathrm{data}_{control}) \propto \sum_{G_{ancestor}} \mathrm{P}(G_{ancestor}) \\ &\times \mathrm{P}(G_{case}|G_{ancestor})\mathrm{P}(G_{control}|G_{ancestor})\mathrm{P}(\mathrm{data}_{case}|G_{case})\,\mathrm{P}(\mathrm{data}_{control}|G_{control})\end{aligned}$$

where the prior $\mathrm{P}(G_{ancestor})$ and the likelihood $\mathrm{P}(\mathrm{data}_{case}|G_{case})$, $\mathrm{P}(\mathrm{data}_{control}|G_{control})$ were calculated as previously described (6). The change rate factor $\mathrm{P}(G_{case}|G_{ancestor})$ and $\mathrm{P}(G_{control}|G_{ancestor})$ for the remaining two genotypes were specified as

$$\mathrm{P}(G_{to}|G_{from}) = \begin{array}{c} \\ \text{ref}-\text{hom} \\ \text{heterozygous} \\ \text{alt}-\text{hom} \\ \text{mosaic} \end{array} \begin{array}{c} \begin{array}{cccc} \text{ref}-\text{hom} & \text{heterozygous} & \text{alt}-\text{hom} & \text{mosaic} \end{array} \\ \begin{pmatrix} 1-R_d-R_m & R_d & 0 & R_m \\ R_d & 1-2\cdot R_d-R_m & R_d & R_m \\ 0 & R_d & 1-R_d-R_m & R_m \\ R_m & R_m & R_m & 1-3\cdot R_m \end{pmatrix} \end{array}$$

After the joint posterior distribution $\mathrm{P}(G_{case}, G_{control}|\mathrm{data}_{case}, \mathrm{data}_{control})$ was calculated, the probability that the case and control samples have different genotypes was considered as the probability of a postzygotic mutation occurring between the case and control samples, summing up all posterior probabilities $\mathrm{P}(G_{case} \neq G_{control})$.

Theoretical generation of sequencing bases

To compare the performance of the 'single', 'trio' and 'paired' Bayesian models in MosaicHunter, we theoretically

generated sequencing bases of sites from a binomial distribution with varied sequencing depths (20–200) and expected alternative allele fractions (0–0.5), and then changed reference allele to alternative allele and alternative allele to reference allele by a Bernoulli process with the sequencing error rate 10^{-3} (fixed baseQ to 30). The sites with expected alternative allele fraction 0 or 0.5 were considered as reference homozygous or heterozygous genotype, respectively, whereas the sites with other allele fractions were considered as true SNMs. For the 'trio' or 'paired' mode, we also generated sequencing bases in parental or control samples where the expected alternative allele fraction was set to 0. Each simulation was done 10^7 times. The posterior probability threshold to determine a positive mosaic site was set to the default 0.05 in each mode of MosaicHunter.

Simulated non-cancer datasets for benchmarking

The simulated non-cancer WGS and WES datasets were generated *in silico* by mixing the sequencing data from NA12878 and that from an unrelated individual (ACC1-II-1 for the WGS dataset and NA12889 for the WES dataset) (6), with average depth of 72X and 258X for the WGS and WES datasets, respectively. The genotyping files were compared between the two individuals, and the genomic positions where NA12878 was homozygous for the reference allele and the unrelated individual was heterozygous were identified. To exclude potential genotyping errors, only the sites located in non-repetitive regions were considered. Because the genders of the two individuals were different, the sites located on the sex chromosomes were excluded. To mimic the postzygotic origination of the SNMs, only the sites genotyped as homozygous for the reference allele in both parents of NA12878 were considered in subsequent analyses. For each of the five alternative allele fractions tested (0.05, 0.1, 0.2, 0.3 and 0.4), we sampled ∼7000 sites genome-wide for the WGS datasets and ∼250 sites within the capture regions for the WES datasets. For each site with adequate sequencing depths, the paired-end reads of NA12878 overlapping with the site were randomly replaced with reads from the unrelated individual at the same site, following binomial sampling with the alternative allele fraction and sequencing depth. As shown in Supplementary Figures S1 and S2, the distribution of the simulated polymorphic sites followed the expected distribution of stochastic sampling of real reads in the WGS and WES datasets. The sensitivity of identifying SNMs at each allele fraction was subsequently calculated as the proportion of polymorphic sites identified from the simulated dataset. The data of NA12878 were utilized as the control for the 'paired' mode of MosaicHunter, Varscan 2 and MuTect. The data of NA12891 and NA12892 were treated as the father and mother for the 'trio' mode of MosaicHunter.

According to a previously described strategy (6), the specificity was estimated by using the WGS and WES read libraries of NA12878. After excluding *de novo* mutations identified from the genotyping files of the NA12878 trio, all the identified postzygotic mutations were considered as false positives, and the depth-dependent specificities were calculated for homozygous and heterozygous sites separately. For the 'single' and 'trio' mode of MosaicHunter as well as SomVarIUS, Solexa-18483 and library 1 listed in Supplementary Table S2 were used for WGS and WES benchmarking, respectively. For the 'paired' mode of MosaicHunter, Varscan 2, and MuTect, another read libraries constructed from the same individual NA12878 (Solexa-18484 for WGS benchmarking and library 2 for WES benchmarking) were treated as the control datasets (Supplementary Table S2).

Precision was subsequently calculated based on the sensitivity and specificity estimated above, and the occurrence rates of germline heterozygous sites (1.2×10^{-3} per bp) and postzygotic mutations (4.4×10^{-7} per bp) were set based on estimates from previous studies (6,42).

TCGA and ICGC cancer datasets for benchmarking

The performance of MosaicHunter in cancer studies was evaluated using two different synthetic cancer datasets from TCGA and ICGC. For the TCGA dataset, the 58–71X WGS data obtained from the breast cancer cell lines HCC1954 and the paired normal control cell line HCC1954BL provided through TCGA Mutation/Variation Calling Benchmark 4. The BAM files of the two cell lines were sub-sampled and mixed with each other to construct two tumor datasets with different tumor purities, 30% tumor *vs* 70% normal (T30N70) and 50% tumor versus 50% normal (T50N50), with an average depth of ∼83×. For the ICGC dataset, we downloaded the ∼40× synthetic sequencing data of paired tumor and normal samples from ICGC-TCGA DREAM Somatic Mutation Calling Challenge. The synthetic data were generated using BAMSurgeon by introducing synthetic mutations with varied allele fractions (50%, 33% and 20%) into the WGS data of HCC1143BL (43). The 'ground truth' list of mutations as well as the candidate lists called by MuTect (22) and Varscan 2 (20) were also downloaded. To achieve a higher sequencing depth for benchmarking the control-free methods, we also generated a merged BAM file with ∼80× depth by combing the tumor and normal sequencing data, where the mutant allele fractions were dropped to 25%, 17.5% and 10%, respectively.

The genotypes of each sample were separately called and filtered using GATK (version 1.6) (44), and CNVnator (version 0.2.7) (45) and BIC-seq (version 2.1.1) (46) were applied for calling copy number variations. The sites that were heterozygous in the tumor sample and homozygous for the reference allele in the normal sample were subsequently considered as postzygotic mutations. To remove potential technical artifacts resulting from base-calling or alignment errors, the sites located in or near repetitive regions, homopolymers, indels and copy number variations were excluded.

Because some parameters of MosaicHunter, including the prior probability of the mosaic genotype in the Bayesian genotyper and the maximum distance in the filter of clustered sites, were originally designed for non-cancer sequencing data and might not be suitable for cancer samples, we estimated the precision and sensitivity with varied values of the two parameters in the TCGA cancer dataset. A maximum distance of 2000 bp (5000 bp as default), showing the highest sensitivity and acceptable specificity, was set for

cancer studies (Supplementary Figure S3), consistent with prior knowledge that cancer samples have a higher density of postzygotic mutations (47).

Whole-genome and whole-exome sequencing

Genomic DNA was extracted from the peripheral blood samples of ACC1-II-1, DS1-II-2, DS1-III-1, AU1-II-1 and their parents. ACC1-II-1 was sequenced at ∼90× average sequencing depth and his parents were sequenced at 30–54× average sequencing depth (Table 1). The exomes of the trios of DS1-II-2 and DS1-III-1 were captured using the Agilent SureSelect Human All Exon 71M kit and sequenced at ∼400× average depth each. The exomes of trio of AU1-II-1 were processed using the Agilent SureSelect Human All Exon 50M kit and sequenced at ∼150× depth each (Table 1). All sequencing was performed on an Illumina HiSeq2000 platform using 100 bp paired-end reads. The clinical histories of all four trios showed no symptoms of cancer or other overgrowth disorders.

Identification of SNMs and performance comparison

Sequencing reads in FASTQ format were aligned to the GRCh37 reference genome using BWA (48) (version 0.6.1). The aligned BAM files were pre-processed as previously described (6), including the removal of duplicated and error-prone reads, indel realignments, and base-quality recalibrations. The 'single', 'trio' and 'paired' modes of MosaicHunter (version 1.0) were used to call postzygotic SNMs from the processed BAM files. For comparison, postzygotic mutations were also called using SomVarIUS (28) (version 1.1), Varscan 2 (20) (version 2.2.11) and MuTect (22) (version 1.1.4), with default settings and 'best-practice' pipeline suggested in their websites. The pipeline of each software includes the steps of both raw genotyping and subsequent filtering.

Experimental validation of SNMs

The candidate SNMs identified from the WGS and WES datasets of the samples were validated by direct Sanger sequencing and PASM (49), a more sensitive amplicon-based resequencing method.

Direct Sanger sequencing was first applied to heterozygous sites in which the mutant allele was inherited from either parent. Subsequently, all the remaining candidates were verified using the standard workflow of PASM. Except for sites for which primers were not able to be designed, the genomic regions flanking the candidate loci were captured after 35 cycles of PCR with an annealing temperature of 59.5°C (Supplementary Table S4), followed by agarose gel electrophoresis and extraction using the Qiagen QIAquick gel extraction kit (Qiagen). Subsequently, the extracted amplicons were barcoded and sequenced using Ion Torrent (Thermo Fisher), according to the manufacturer's instructions. The reads were aligned against the hg19 reference genome using Torrent Suite (version 4.4.2) and subsequently pileuped using Samtools (version 1.2) (50). The hierarchical Bayesian model was applied to estimate the mosaic allele fraction, considering sequencing errors and random variations of binomial sampling. An SNM site is considered true when the estimated mosaic allele fraction is between 3% and 40% in the offspring (mosaic genotype) and less than 3% in both parents (homozygous genotype).

Data access

For benchmarking the performance of MosaicHunter in non-cancer individuals, the WGS and WES datasets of CEU trios and an unrelated individual for generating simulated sequencing data are available at the URLs listed in Supplementary Table S2. The TCGA cancer dataset were downloaded from TCGA Mutation/Variation Calling Benchmark 4 following the instructions at https://gdc.cancer.gov/resources-tcga-users/. For the ICGC cancer dataset, the sequencing and genotyping files of Synthetic Dataset 3 were downloaded from ICGC-TCGA DREAM Somatic Mutation Calling Challenge (43). The WGS and WES datasets of trios of non-cancer samples have been deposited to the Sequencing Read Archives under accession number SRP028833.

RESULTS

Description of MosaicHunter's Bayesian genotyper and error filters

MosaicHunter is summarized in Figure 1A, described below, and detailed in Materials and Methods. We had previously reported the single-sample unpaired WGS mode of MosaicHunter. It calculated the posterior probabilities of the mosaic genotype and three inherited genotypes under a Bayesian model (6). The prior probability incorporated the population allele frequency information from dbSNP. The likelihood of each genotype was calculated by binomial sampling of reads for each allele and modeling sequencing substitution error based on base quality (Figure 1A). We also developed a series of stringent filters to remove systematic sequencing errors at the levels of genomic region, read alignment, and nucleotide site (Figure 1A), which we had demonstrated as effective (6).

Compared to the previous version which can only analyze WGS data of unpaired samples, the new version of MosaicHunter has the significantly enhanced flexibility of being able to analyze WGS as well as WES data in 'single', 'trio', and 'paired' modes. First, to extend MosaicHunter to handle capture-based WES data, we began with investigating the extent of capturing bias and over-dispersion in five WES datasets. As shown in Supplementary Figure S4, deviation from the binomial expectation was observed in all the five WES datasets (but not in the WGS dataset), confirming the existence of notable capturing bias and over-dispersion in WES data and the necessity of a new model. Hence, we replaced the binomial model in our previous genotyper with a new beta-binomial model in order to better fit the WES data (Materials and Methods). After using the new beta-binomial model, the proportion of heterozygous sites located outside the 95% confidence intervals decreased from 17% to 4.4% in WES data (Figure 1B), closer to the theoretical expectation of 5%, demonstrating that the new model indeed had better performance for distinguishing SNMs from heterozygous sites in WES data.

Table 1. Validation rates of MosaicHunter on whole-genome sequencing (WGS) and whole-exome sequencing (WES) data of control-free unpaired samples ('single' mode) and familial trio samples ('trio' mode)

Sequencing type	Sample ID	Sequencing depth	'Single' mode validation rate	Sequencing depth of father	Sequencing depth of mother	'Trio' mode validation rate
WGS	ACC1-II-1	87×	31.8% (7/22)	54×	30×	100% (8/8)
WES[a]	DS1-II-2	403×	33.3% (1/3)	375×	419×	100% (1/1)
WES[a]	DS1-III-1	398×	33.3% (1/3)	375×	403×	100% (1/1)
WES[b]	AU1-II-1	169×	25% (1/4)	119×	130×	40% (2/5)

[a]Captured using the Agilent SureSelect Human All Exon 71M kit.
[b]Captured using the Agilent SureSelect Human All Exon 50M kit.

Although the main purpose of developing MosaicHunter was to detect SNMs without needing control samples, we also extended its Bayesian model to allow for the incorporation of sequencing data from additional available samples to improve accuracy, including blood samples from parents or paired tissue samples from the same individual. As illustrated in Figure 1C, in the 'trio' mode of MosaicHunter, sequencing data of parents were modeled in the Bayesian genotyper by linking the genotype likelihoods of the parents and their offspring with a factor of transmission conditional probability, formulated according to Mendelian inheritance rules and estimated mutation rates (Materials and Methods). To incorporate data from paired tissue sample of the same individual, in the 'paired' mode of MosaicHunter we introduced a latent state of ancestral genotype and extended the Bayesian model (Figure 1D and Materials and Methods). Using simulated sequencing bases with varied sequencing depths and mutant allele fractions, we showed that the 'trio' and 'paired' modes achieved higher sensitivity to detect true mosaicisms, especially for SNMs with allele fractions greater than 0.3 (Figure 1E-G), and higher specificity to remove false positives arisen from homozygous and heterozygous sites (Supplementary Table S1). Our results confirmed that the extended 'trio' and 'paired' Bayesian models incorporating data from additional samples performed better than the 'single' model, and that all three models could benefit from increased sequencing depth (Figure 1E-G and Supplementary Table S1).

MosaicHunter takes the aligned reads in BAM format (50) files as input. It is implemented in JAVA. A typical analysis of a 90× WGS datasets takes ∼20 h using one core of a 2.66 GHz Intel Xeon processor, compared to over 500 h taken by our previous version implemented in Perl (6). MosaicHunter is freely available for non-commercial use at http://mosaichunter.cbi.pku.edu.cn.

Evaluation and comparison on simulated benchmarking non-cancer datasets

We first evaluated MosaicHunter on *in silico* simulated non-cancer WGS and WES datasets. Using the WGS and WES datasets of sample NA12878 from the 1000 Genomes Project, we generated SNMs with varied mutant allele fractions (0.05, 0.1, 0.2, 0.3 and 0.4, respectively) by randomly replacing the sequencing reads of sample NA12878 with those of an unrelated healthy individual (ACC1-II-1 for WGS and NA12889 for WES) following a binomial sampling process at autosomal sites where NA12878 was homozygous for the reference allele and the second individual was heterozygous (Supplementary Table S2 and Materials and Methods). False positive rates were estimated by counting mis-identified postzygotic SNMs from homozygous or heterozygous sites of NA12878 (Materials and Methods).

Results on simulated WGS datasets are shown in Figure 2A–C. As shown in Figure 2A, the 'single' mode of MosaicHunter achieved a precision of over 50% in identifying SNMs with minor allele fractions between 0.05 and 0.3, although the precision was lower for minor allele fractions of 0.4. In comparison, SomVarIUS achieved a precision <1%. One unique feature of MosaicHunter is that it can utilize WGS or WES data from parents to improve the accuracy of SNM identification. Utilizing sequencing data from the parents of sample NA12878, the 'trio' mode of MosaicHunter achieved a significantly improved precision of over 75% for SNMs with allele fractions between 0.05 and 0.4. Using the original NA12878 dataset as control, the 'paired' mode of MosaicHunter achieved a precision of 95% for mutations with allele fractions >0.05. In comparison, Varscan 2 (20) and MuTect (22) achieved precisions of 30% and 5%, respectively. Furthermore, as shown in Figure 2B and C, all three modes of MosaicHunter achieved lower false positive rates at both homozygous sites (Figure 2B) and heterozygous sites (Figure 2C) and outperformed SomVarIUS, Varscan 2 and MuTect.

Results on simulated WES datasets are shown in Figure 2D–F. We first demonstrated that the beta-binomial model of MosaicHunter reduced false positive rates more efficiently than the original binomial model, particularly for false positives arising from heterozygous sites (Supplementary Figure S5). Even without using paired control samples, the 'single' mode of MosaicHunter achieved a precision ranging from 79% to 94% in the identification of SNMs with allele fractions of 0.05–0.3 (Figure 2D), outperforming SomVarIUS and achieving comparable performance to Varscan 2 and MuTect which used the paired control dataset (average precision = 76% and 93%, respectively) (Figure 2D). When utilizing parental data, the average precision of the 'trio' mode of MosaicHunter increased to 95% for SNMs with minor allele fraction between 0.05 and 0.4 (Figure 2D). When utilizing paired control data, the average precision of the 'paired' mode of MosaicHunter further increased to 99% for SNMs with minor allele fraction between 0.05 and 0.4 (Figure 2D), surpassing the precision of Varscan 2 and MuTect. Furthermore, all three modes of MosaicHunter achieved comparable or lower false positive rate than SomVarIUS, Varscan 2, and MuTect (Figure 2E and F).

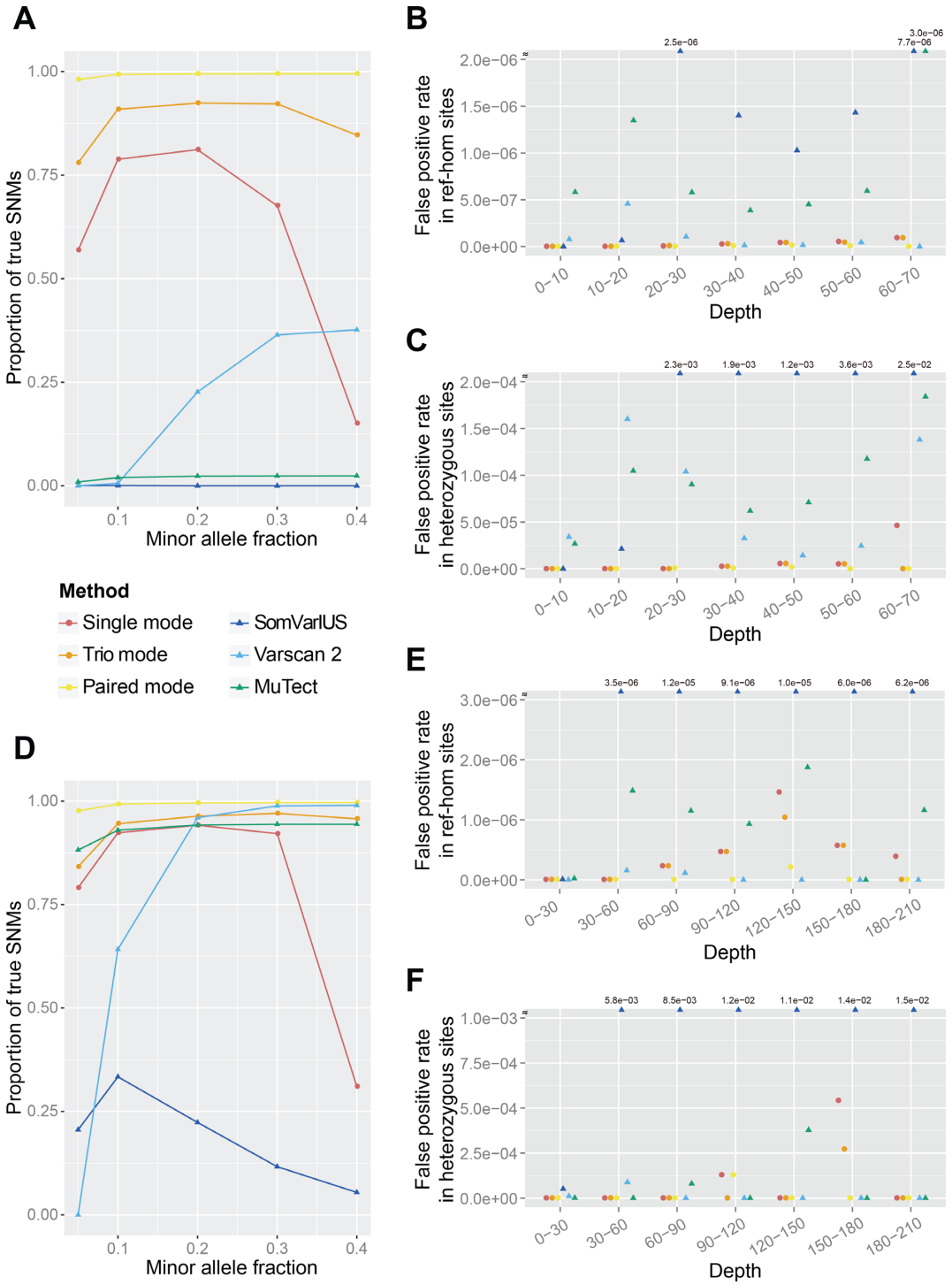

Figure 2. Comparison of the precision and false positive rate of the identification of SNMs with varied allele fractions using MosaicHunter, SomVarIUS, Varscan 2 and MuTect demonstrated that MosaicHunter had better performance on both WGS and WES data. (**A**) Proportion of true SNMs among all sites identified in WGS data; (**B**) false positive rate among simulated reference homozygous (ref-hom) sites in WGS data; (**C**) false positive rate among simulated heterozygous sites in WGS data; (**D**) proportion of true SNMs among all sites identified in WES data; (**E**) false positive rate among simulated reference homozygous (ref-hom) sites in WES data; (**F**) false positive rate among simulated reference heterozygous sites in WES data.

Evaluation of identifying somatic mutations in simulated unpaired and paired cancer genomes

To assess the performance of MosaicHunter in cancer studies, we firstly applied MosaicHunter to WGS datasets from a breast cancer cell line and its paired normal control that are part of the TCGA Mutation/Variation Calling Benchmark (Materials and Methods), a widely-used benchmarking dataset for cancer-related bioinformatics tools (23,51). Considering that tumor samples are typically a mixture of tumor cells and normal cells, two simulated tumor datasets were generated with tumor purities of 30% and 50% (labeled as T30N70 and T50N50), respectively. In cancer studies where the paired normal control tissue samples are unavailable or inappropriate, MosaicHunter can be an important tool for control-free identification of somatic mutations in tumor samples. As illustrated in Figure 3A, the 'single' mode of MosaicHunter identified 1457 true somatic mutation sites with only 26 false positives from the T30N70 dataset in a control-free manner. In comparison SomVarIUS identified 840 true positives with 888 false positives. When utilizing the paired normal control dataset, the 'paired' mode of MosaicHunter further reduced the number of false positives to four and achieved comparable sensitivity and higher specificity than Varscan 2 and MuTect (Figure 3A). Evaluation on the T50N50 dataset confirmed MosaicHunter's performances (Figure 3A).

Because the multi-clonal origin of tumor samples was common in many cancer types (16,52), we next employed MosaicHunter to a simulated cancer WGS dataset from ICGC-TCGA DREAM Somatic Mutation Calling Challenge in which the allele fraction of synthetic mutations varied from 20%, 33% to 50% (43). As expected, without the help of matched control samples, the somatic mutations with 50% allele fractions were not able to be distinguished from germline sites (Figure 3B and C, solid lines). However, such somatic mutations could be identified if the roughly equal amount of sequencing data from both the tumor and normal samples were pooled altogether as the input for the 'single' mode of MosaicHunter (Figure 3B and C, dotted lines). Our results suggested, in both cases, the 'single' mode of MosaicHunter had better sensitivity and precision than SomVarIUS, especially for the mutations with relatively higher allele fractions (Figure 3B and C). When the sequencing data from the paired control sample was incorporated, MosaicHunter achieved comparable precision but a little lower sensitivity against Varscan2 and MuTect (Figure 3B and C). In summary, our two benchmarking assessments demonstrated that MosaicHunter can identify somatic mutations in cancer samples with varied mutant allele fractions with or without matched control tissues.

Experimental validation of SNMs identified in control-free unpaired non-cancer samples

We further evaluated and experimentally validated the results of MosaicHunter on new WGS and WES data of real samples. We first confirmed that this newly implemented version of MosaicHunter achieved similar accuracy to our previously reported version when analyzing unpaired WGS data. We used our previously sequenced WGS data from the peripheral blood sample of a clinically unremarkable individual without clinical history of cancer (ACC1-II-1, Table 1) (6). Out of 22 candidate SNMs identified by MosaicHunter, seven (31.8%) were validated as true by PGM Amplicon Sequencing of Mosaicism (PASM), an amplicon-based ultra-high sequencing method (49) (Supplementary Table S3 and Materials and Methods), an accuracy similar to that of the previous version (6). The low number of SNMs is consistent with the fact that the sequenced individual has no clinical history of cancer.

To evaluate MosaicHunter's accuracy of identifying postzygotic SNMs in WES data without using paired control samples, we used Illumina HiSeq platform to sequence the whole exomes from peripheral blood samples from three unrelated individuals who did not have any clinical history of cancer (samples DS1-II-2, DS1-III-1, and AU1-II-1, Table 1). The sequencing depth ranged from 169× to 403×. We applied the 'single' mode of MosaicHunter to the WES datasets. Out of a total of 10 candidate SNMs, three (30%, Table 1) were validated by PASM (49) (Supplementary Table S3 and Materials and Methods). The allelic fractions of the validated SNMs were 17.6%, 18.0% and 12.5%.

For benchmarking, we compared with SomVarIUS (28), the only other available software for control-free SNM detection, by applying it to the same WGS and WES datasets. SomVarIUS reported 25 848 candidate SNMs from the WGS data and an average of 69 candidate SNMs from the WES datasets, with mosaic mutant allele fraction >0.05. These were orders of magnitude larger than previous estimation (6), suggesting a high false positive rate. Furthermore, despite the huge number of reported SNM candidates, only two of the seven validated SNMs in WGS dataset and none of the three validated SNMs in WES dataset (Supplementary Table S2) were identified by SomVarIUS.

Experimental validation of SNMs identified in non-cancer familial trio samples

We sequenced the whole genomes and exomes of peripheral blood DNA from the parents of ACC1-II-1, DS1-II-2, DS1-III-1 and AU1-II-1 by Illumina HiSeq platform (Table 1). When the parental data was incorporated, MosaicHunter achieved increased sensitivity as well as precision for both the WGS dataset, with eight SNMs validated among eight identified (100% precision), and the WES datasets, with four SNMs validated among seven identified (57.1% precision) (Table 1). The precisions on the DS1-II-2 and DS1-III-1 WES datasets were significantly higher than that on the AU1-II-1 WES dataset likely because they were sequenced at higher depths and using a newer version of the capture kit which resulted in less over-dispersion (Supplementary Figure S4). Our results demonstrated the power of using parental data to significantly improve SNM detection accuracy when no paired control tissues are available. To our knowledge, no other existing methods can incorporate familial information in control-free SNM detection.

DISCUSSION

Emerging evidence has demonstrated the previously neglected contribution of postzygotic mutations in the etiol-

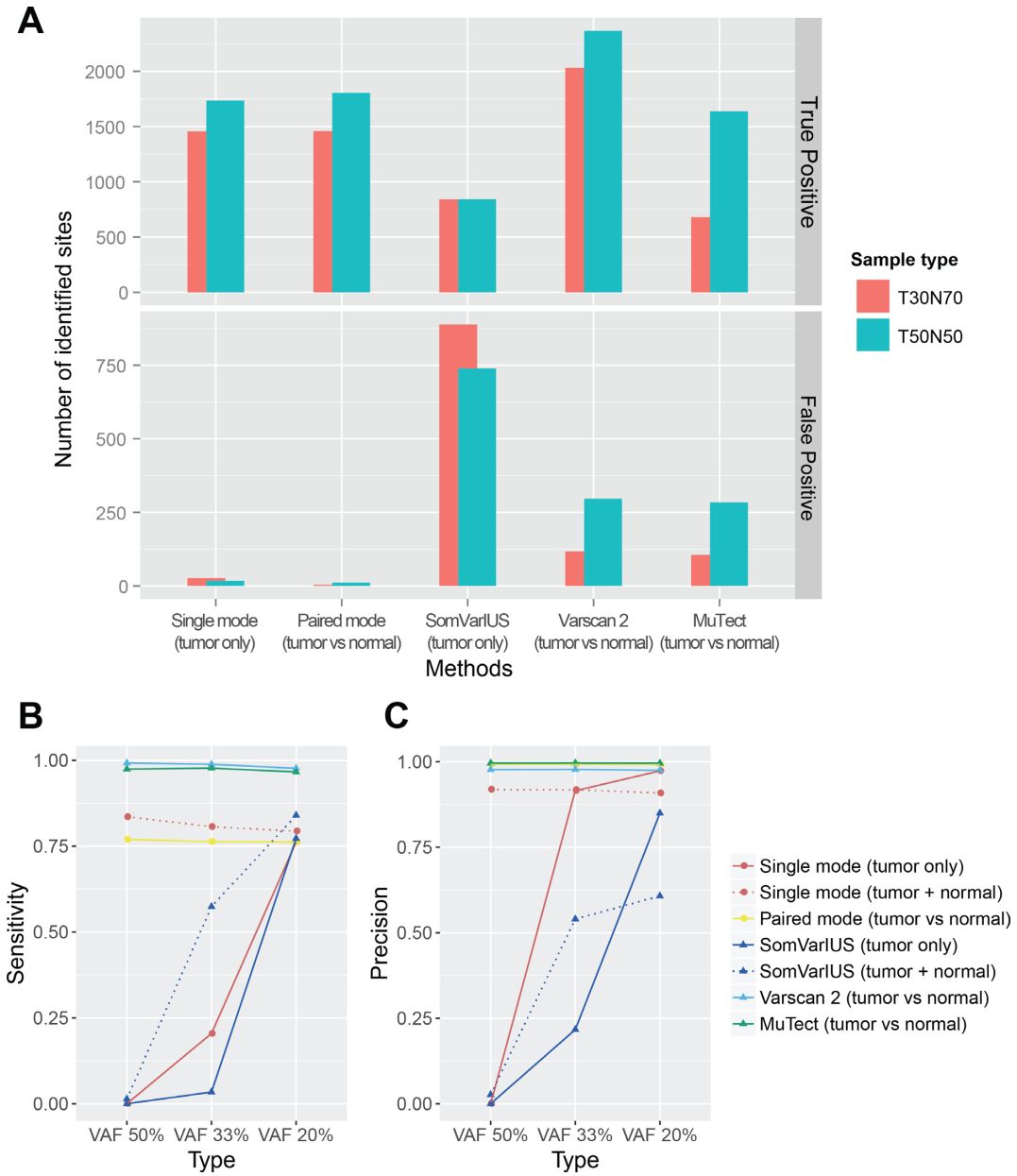

Figure 3. Performance of MosaicHunter in the TCGA and ICGC cancer datasets. (**A**) Number of true and false positive SNMs identified from the TCGA cancer dataset using MosaicHunter, SomVarIUS, Varscan 2 and MuTect. T30N70 and T50N50 denote sequencing data simulating 30% and 50% tumor cell purity, respectively. (**B** and **C**) Comparison of the sensitivity (B) and precision (C) of MosaicHunter, SomVarIUS, Varscan 2 and MuTect from the ICGC cancer dataset, with three different sub-clonal allele fractions of synthetic mutations. For the 'single' mode of MosaicHunter and SomVarIUS, solid lines denote using the original ~40× tumor sequencing data as input, whereas dotted lines denote using the combined ~80× sequencing data of both the tumor and normal samples as input.

ogy of non-cancer diseases (53–56). Within a healthy individual, postzygotic mutations identified in different samples of the same individual have broadened the concept of 'a personal genome' to 'the personal genomes' (2,4,6). The origins of more and more '*de novo*' mutations identified in offspring have been traced back to postzygotic mutations in their parents (57). In all these cases, the identification of postzygotic mutations faces the challenge of not having paired control tissues. Here, we demonstrated that MosaicHunter can identify SNMs from whole-genome and whole-exome data in the absence of paired control samples from the same individual, making it a useful tool for these non-cancer studies as well as some cancer studies when matched normal control samples are unavailable or difficult to obtain.

The increasing throughput and decreasing cost of next-generation sequencing technologies over the past decade have made the genome-wide identification of postzygotic mutations possible. Compared to WGS, WES is more cost-effective because it is enriched for functional mutations and usually has higher sequencing depth (58). However, the notable capturing bias and over dispersion in WES data posed serious challenges for identifying SNMs. We demonstrated

that the integration of a new beta-binomial model into the Bayesian genotyper can effectively handle the capturing bias and over-dispersion in WES data and accurately identify SNMs in the data. Since the parameters of the beta-binomial model vary across different capturing kits and sequencing platforms, MosaicHunter enables users to pre-estimate these parameters from the heterozygous sites in customized WES or other capturing-based target panel re-sequencing data.

Our simulation results demonstrated that increasing sequencing depth and base quality can significantly increase the accuracy of SNM identification (Figure 1E–G and Supplementary Table S1). For control-free non-cancer samples, the sensitivities to detect SNMs with various allele fractions approximately doubled when the depth increased from 40× to 60× (Figure 1E). As shown in our cancer benchmarking data, MosaicHunter performed generally better in the ∼65× TCGA dataset than the ∼40× ICGC dataset (Figure 3). Thus, we suggest a minimal average depth of 60× for identifying SNMs from control-free sequencing data. For cancer studies, the tumor purity is another key factor critical for MosaicHunter's performance. In the 'single' mode, MosaicHunter achieved the highest sensitivity when detecting SNMs with allele fractions ranging from 0.15 to 0.2 (Figures 1E and 3). Assuming that the majority of cancer mutations are heterozygous in tumor cells, this range of allele fraction denotes tumor purity between 30% and 40%.

If a postzygotic mutation affects germ cells, there is a chance that the mutant allele may be transmitted to the offspring and lead to a heterozygous genotype. As shown in Huang *et al.* (6), the heterozygous missense mutations in the *SCN1A* gene of two children with Dravet syndrome were inherited from postzygotic SNMs in their respective parents. MosaicHunter may be used to identify such parental SNMs and contribute to more sensitive genetic counseling. At the same time, as we have shown in this study, when paired control samples from the same individual are not available, incorporating parental sequencing data can significantly improve the performance of SNM identification in the offspring. Collecting and sequencing parental samples is already a routine practice in the studies of *de novo* mutations and inherited mutations in genetic disorders (59). Our results advocate the routine collection of parental samples in broader areas of medical genetics, human genetics and genomics studies.

To achieve higher precision that could allow the identification and validation of postzygotic mutations in non-cancer samples, MosaicHunter incorporated a series of stringent error filters to remove numerous sequencing errors, which inevitably limited the sensitivity to detect SNMs. In the human genome, about half of the DNA sequence is comprised of various types of repetitive elements, and the highly homologous DNA sequences post challenges on genome assembly and read alignment, which hinders not only the identification but also the validation of mutations in these regions (60). The repetitive regions are filtered out in the default pipeline of MosaicHunter. In the non-repetitive regions up to ∼50% of the true SNMs would be missed due to the stringent error filters in our default pipeline (Supplementary Table S5). To allow users to achieve their own desired balance between sensitivity and specificity, the MosaicHunter software allows users to easily customize the on/off and parameters of each filter.

In this study, we identified and validated 8 SNMs in the WGS dataset and 1–2 SNMs in each WES dataset of non-cancer individuals. Based on the sensitivity of MosaicHunter, the occurrence rate of postzygotic SNMs with allele fractions greater than 0.05 was about 1×10^{-8} to 1×10^{-7} per nucleotide in healthy individuals, which was comparable to the rate of *de novo* mutations (61) but greatly lower than the rate of somatic mutations in tumors (62). A systematic profiling of multiple non-cancer tissues from more individuals would be required for a more accurate estimation of the baseline postzygotic mutation rate. As NGS technologies become cheaper, more powerful, and more accurate, they will enable us to identify more SNMs in non-cancer individuals and elucidate their genomic patterns, which may shed light on exploring the etiology of genetic diseases, understanding the origin of *de novo* mutations, and providing baseline mutation profiles for future cancer studies.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We are grateful to Drs Yu Shyr, Jinzhu Jia and Cheng Li for their valuable suggestions and discussions about the statistical models.

FUNDING

National Natural Science Foundation of China [31530092, 81171221]; Ministry of Science and Technology 863 Grant [2015AA020108]; Peking University Clinical Cooperation '985 Project' [PKU-2014-1-1, PKU-2013-1-06]. Funding for open access charge: National Natural Science Foundation of China [31530092, 81171221]; Ministry of Science and Technology 863 Grant [2015AA020108]; Peking University Clinical Cooperation '985 Project' [PKU-2014-1-1, PKU-2013-1-06].

Conflict of interest statement. None declared.

REFERENCES

1. Lupski,J.R. (2013) Genome mosaicism—one human, multiple genomes. *Science*, **341**, 358–359.
2. De,S. (2011) Somatic mosaicism in healthy human tissues. *Trends Genet.*, **27**, 217–223.
3. Biesecker,L.G. and Spinner,N.B. (2013) A genomic view of mosaicism and human disease. *Nat. Rev. Genet.*, **14**, 307–320.
4. O'Huallachain,M., Karczewski,K.J., Weissman,S.M., Urban,A.E. and Snyder,M.P. (2012) Extensive genetic variation in somatic human tissues. *Proc. Natl. Acad. Sci. U.S.A.*, **109**, 18018–18023.
5. Behjati,S., Huch,M., Boxtel,R.V., Karthaus,W., Wedge,D.C., Tamuri,A.U., Martincorena,I., Petljak,M., Alexandrov,L.B., Gundem,G. *et al.* (2014) Genome sequencing of normal cells reveals developmental lineages and mutational processes. *Nature*, **513**, 422–425.
6. Huang,A.Y., Xu,X., Ye,A.Y., Wu,Q., Yan,L., Zhao,B., Yang,X., He,Y., Wang,S., Zhang,Z. *et al.* (2014) Postzygotic single-nucleotide mosaicisms in whole-genome sequences of clinically unremarkable individuals. *Cell Res.*, **24**, 1311–1327.

7. Upton,K.R., Gerhardt,D.J., Jesuadian,J.S., Richardson,S.R., Sanchez-Luque,F.J., Bodea,G.O., Ewing,A.D., Salvador-Palomeque,C., Knaap,M.S.V.D., Brennan,P.M. *et al.* (2015) Ubiquitous L1 mosaicism in hippocampal neurons. *Cell*, **161**, 228–239.
8. Greenman,C., Stephens,P., Smith,R., Dalgliesh,G.L., Hunter,C., Bignell,G., Davies,H., Teague,J., Butler,A., Stevens,C. *et al.* (2007) Patterns of somatic mutation in human cancer genomes. *Nature*, **446**, 153–158.
9. Besenbacher,S., Liu,S., Izarzugaza,J.M.G., Grove,J., Belling,K., Bork-Jensen,J., Huang,S., Als,T.D., Li,S., Yadav,R. *et al.* (2015) Novel variation and de novo mutation rates in population-wide de novo assembled Danish trios. *Nat. Commun.*, **6**, 5969.
10. Watson,I.R., Takahashi,K., Futreal,P.A. and Chin,L. (2013) Emerging patterns of somatic mutations in cancer. *Nat. Rev. Genet.*, **14**, 703–718.
11. Futreal,P.A., Coin,L., Marshall,M., Down,T., Hubbard,T., Wooster,R., Rahman,N. and Stratton,M.R. (2004) A census of human cancer genes. *Nat. Rev. Cancer*, **4**, 177–183.
12. Erickson,R.P. (2014) Recent advances in the study of somatic mosaicism and diseases other than cancer. *Curr. Opin. Genet. Dev.*, **26**, 73–78.
13. Lindhurst,M.J., Sapp,J.C., Teer,J.K., Johnston,J.J., Finn,E.M., Peters,K., Turner,J., Cannons,J.L., Bick,D., Blakemore,L. *et al.* (2011) A mosaic activating mutation in AKT1 associated with the proteus syndrome. *N. Engl. J. Med.*, **365**, 611–619.
14. Poduri,A., Evrony,G.D., Cai,X. and Walsh,C.A. (2013) Somatic mutation, genomic variation, and neurological disease. *Science*, **341**, 43.
15. Martincorena,I., Roshan,A., Gerstung,M., Ellis,P., Loo,P.V., McLaren,S., Wedge,D.C., Fullam,A., Alexandrov,L.B., Tubio,J.M. *et al.* (2015) High burden and pervasive positive selection of somatic mutations in normal human skin. *Science*, **348**, 880–886.
16. Genovese,G., Kähler,A.K., Handsaker,R.E., Lindberg,J., Rose,S.A., Bakhoum,S.F., Chambert,K., Mick,E., Neale,B.M., Fromer,M. *et al.* (2014) Clonal hematopoiesis and blood-cancer risk inferred from blood DNA sequence. *N. Engl. J. Med.*, **371**, 2477–2487.
17. Jaiswal,S., Fontanillas,P., Flannick,J., Manning,A., Grauman,P.V., Mar,B.G., Lindsley,R.C., Mermel,C.H., Burtt,N., Chavez,A. *et al.* (2014) Age-related clonal hematopoiesis associated with adverse outcomes. *N. Engl. J. Med.*, **371**, 2488–2498.
18. Campbell,I.M., Stewart,J.R., James,R.A., Lupski,J.R., Stankiewicz,P., Olofsson,P. and Shaw,C.A. (2014) Parent of origin, mosaicism, and recurrence risk: probabilistic modeling explains the broken symmetry of transmission genetics. *Am. J. Hum. Genet.*, **95**, 345–359.
19. Rahbari,R., Wuster,A., Lindsay,S.J., Hardwick,R.J., Alexandrov,L.B., Turki,S.A., Dominiczak,A., Morris,A., Porteous,D., Smith,B. *et al.* (2016) Timing, rates and spectra of human germline mutation. *Nat. Genet.*, **48**, 126–133.
20. Koboldt,D.C., Zhang,Q., Larson,D.E., Shen,D., McLellan,M.D., Lin,L., Miller,C.A., Mardis,E.R., Ding,L. and Wilson,R.K. (2012) VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.*, **22**, 568–576.
21. Roth,A., Ding,J., Morin,R., Crisan,A., Ha,G., Giuliany,R., Bashashati,A., Hirst,M., Turashvili,G., Oloumi,A. *et al.* (2012) JointSNVMix: amprobabilistic model for accurate detection of somatic mutations In normal/tumour paired next generation sequencing data. *Bioinformatics*, **28**, 907–913.
22. Cibulskis,K., Lawrence,M.S., Carter,S.L., Sivachenko,A., Jaffe,D., Sougnez,C., Gabriel,S., Meyerson,M., Lander,E.S. and Getz,G. (2013) Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat. Biotechnol.*, **31**, 213–219.
23. Usuyama,N., Shiraishi,Y., Sato,Y., Kume,H., Homma,Y., Ogawa,S., Miyano,S. and Imoto,S. (2014) HapMuC: somatic mutation calling using heterozygous germ line variants near candidate mutations. *Bioinformatics*, **30**, 3302–3309.
24. Nik-Zainal,S., Alexandrov,L.B., Wedge,D.C., Loo,P.V., Greenman,C.D., Raine,K., Jones,D., Hinton,J., Marshall,J., Stebbings,L.A. *et al.* (2012) Mutational processes molding the Genomes of 21 breast cancers. *Cell*, **149**, 979–993.
25. Lawrence,M.S., Stojanov,P., Mermel,C.H., Robinson,J.T., Garraway,L.A., Golub,T.R., Meyerson,M., Gabriel,S.B., Lander,E.S. and Getz,G. (2014) Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature*, **505**, 495–501.
26. Wei,Z., Wang,W., Hu,P., Lyon,G.J. and Hakonarson,H. (2011) SNVer: a statistical tool for variant calling in analysis of pooled or individual next-generation sequencing data. *Nucleic Acids Res.*, **39**, e132.
27. Wilm,A., Aw,P.P.K., Bertrand,D., Yeo,G.H.T., Ong,S.H., Wong,C.H., Khor,C.C., Petric,R., Hibberd,M.L. and Nagarajan,N. (2012) LoFreq: a sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets. *Nucleic Acids Res.*, **40**, 11189–11201.
28. Smith,K.S., Yadav,V.K., Pei,S., Pollyea,D.A., Jordan,C.T. and De,S. (2016) SomVarIUS: somatic variant identification from unpaired tissue samples. *Bioinformatics*, **32**, 808–813.
29. Ramu,A., Noordam,M.J., Schwartz,R.S., Wuster,A., Hurles,M.E., Cartwright,R.A. and Conrad,D.F. (2013) DeNovoGear: de novo indel and point mutation discovery and phasing. *Nat. Methods*, **10**, 985–987.
30. Pagnamenta,A.T., Lise,S., Harrison,V., Stewart,H., Jayawant,S., Quaghebeur,G., Deng,A.T., Murphy,V.E., Akha,E.S., Rimmer,A. *et al.* (2012) Exome sequencing can detect pathogenic mosaic mutations present at low allele frequencies. *J. Hum. Genet.*, **57**, 70–72.
31. Tapper,W.J., Foulds,N., Cross,N.C.P., Aranaz,P., Score,J., Hidalgo-Curtis,C., Robinson,D.O., Gibson,J., Ennis,S., Temple,I.K. *et al.* (2014) Megalencephaly syndromes: exome pipeline strategies for detecting low-level mosaic mutations. *PLoS One*, **9**, e86940.
32. Robasky,K., Lewis,N.E. and Church,G.M. (2014) The role of replicates for error mitigation in next-generation sequencing. *Nat. Rev. Genet.*, **15**, 56–62.
33. Li,R., Montpetit,A., Rousseau,M., Wu,S.Y.M., Greenwood,C.M.T., Spector,T.D., Pollak,M., Polychronakos,C. and Richards,J.B. (2014) Somatic point mutations occurring early in development: a monozygotic twin study. *J. Med. Genet.*, **51**, 28–34.
34. King,D.A., Jones,W.D., Crow,Y.J., Dominiczak,A.F., Foster,N.A., Gaunt,T.R., Harris,J., Hellens,S.W., Homfray,T., Innes,J. *et al.* (2015) Mosaic structural variation in children with developmental disorders. *Hum. Mol. Genet.*, **24**, 2733–2745.
35. Petersen,B.-S., Spehlmann,M.E., Raedler,A., Stade,B., Thomsen,I., Rabionet,R., Rosenstiel,P., Schreiber,S. and Franke,A. (2014) Whole genome and exome sequencing of monozygotic twins discordant for Crohn's disease. *BMC Genomics*, **15**, 564.
36. Magne,F., Serpa,R., Vliet,G.V., Samuels,M.E. and Deladoëy,J. (2015) Somatic mutations are not observed by exome sequencing of lymphocyte DNA from monozygotic twins discordant for congenital hypothyroidism due to thyroid dysgenesis. *Hormone Res. Paediatr.*, **83**, 79–85.
37. Reumers,J., Rijk,P.D., Zhao,H., Liekens,A., Smeets,D., Cleary,J., Loo,P.V., Bossche,M.V.D., Catthoor,K., Sabbe,B. *et al.* (2012) Optimized filtering reduces the error rate in detecting genomic variants by short-read sequencing. *Nat. Biotechnol.*, **30**, 61–68.
38. Gundry,M. and Vijg,J. (2012) Direct mutation analysis by high-throughput sequencing: From germline to low-abundant, somatic variants. *Mutat. Res.*, **729**, 1–15.
39. Santoni,F.A., Makrythanasis,P., Nikolaev,S., Guipponi,M., Robyr,D., Bottani,A. and Antonarakis,S.E. (2014) Simultaneous identification and prioritization of variants in familial, de novo, and somatic genetic disorders with VariantMaster. *Genome Res.*, **24**, 349–355.
40. Cleary,J.G., Braithwaite,R., Gaastra,K., Hilbush,B.S., Inglis,S., Irvine,S.A., Jackson,A., Littin,R., Nohzadeh-Malakshah,S., Rathod,M. *et al.* (2014) Joint variant and De Novo mutation identification on pedigrees from high-throughput sequencing data. *J. Comput. Biol.*, **21**, 405–419.
41. Meacham,F., Boffelli,D., Dhahbi,J., Martin,D.I., Singer,M. and Pachter,L. (2011) Identification and correction of systematic error in high-throughput sequence data. *BMC Bioinformatics*, **12**, 451.
42. Consortium,T.G.P. (2010) A map of human genome variation from population-scale sequencing. *Nature*, **467**, 1061–1073.
43. Ewing,A.D., Houlahan,K.E., Hu,Y., Ellrott,K., Caloian,C., Yamaguchi,T.N., Bare,J.C., P'ng,C., Waggott,D., Sabelnykova,V.Y. *et al.* (2015) Combining tumor genome simulation with crowdsourcing to benchmark somatic singlenucleotide-variant detection. *Nat. Methods*, **12**, 623–630.

44. DePristo,M.A., Banks,E., Poplin,R., Garimella,K.V., Maguire,J.R., Hartl,C., Philippakis,A.A., Angel,G.d., Rivas,M.A., Hanna,M. *et al.* (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.*, **43**, 491–498.
45. 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.
46. Xi,R., Hadjipanayis,A.G., Luquette,L.J., Kim,T.-M., Lee,E., Zhang,J., Johnson,M.D., Muzny,D.M., Wheeler,D.A., Gibbs,R.A. *et al.* (2011) Copy number variation detection in whole-genome sequencing data using the Bayesian information criterion. *Proc. Natl. Acad. Sci. U.S.A.*, **108**, E1128–E1136.
47. Roberts,S.A. and Gordenin,D.A. (2014) Hypermutation in human cancer genomes: footprints and mechanisms. *Nat. Rev. Cancer*, **14**, 786–800.
48. Li,H. and Durbin,R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754–1760.
49. Xu,X., Yang,X., Wu,Q., Liu,A., Yang,X., Ye,A.Y., Huang,A.Y., Li,J., Wang,M., Yu,Z. *et al.* (2015) Amplicon resequencing identified parental mosaicism for approximately 10% of 'de novo' SCN1A mutations in children with dravet syndrome. *Hum. Mutat.*, **36**, 861–872.
50. Li,H., Handsaker,B., Wysoker,A., Fennell,T., Ruan,J., Homer,N., Marth,G., Abecasis,G., Durbin,R. and Subgroup,G.P.D.P. (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.
51. Favero,F., Joshi,T., Marquard,A.M., Birkbak,N.J., Krzystanek,M., Li,Q., Szallasi,Z. and Eklund,A.C. (2015) Sequenza: allele-specific copy number and mutation profiles from tumor sequencing data. *Ann. Oncol.*, **26**, 64–70.
52. Gerlinger,M., Horswell,S., Larkin,J., Rowan,A.J., Salm,M.P., Varela,I., Fisher,R., McGranahan,N., Matthews,N., Santos,C.R. *et al.* (2014) Genomic architecture and evolution of clear cell renal cell carcinomas defined by multiregion sequencing. *Nat. Genet.*, **46**, 225–233.
53. Nota,B., Hamilton,E.M., Sie,D., Ozturk,S., Dooren,S.J.M.V., Ojeda,M.R.F., Jakobs,C., Christensen,E., Kirk,E.P., Sykut-Cegielska,J. *et al.* (2013) Novel cases of D-2-hydroxyglutaric aciduria with IDH1 or IDH2 mosaic mutations identified by amplicon deep sequencing. *J. Med. Genet.*, **50**, 754–759.
54. Jamuar,S.S., Lam,A.T.N., Kircher,M., D'Gama,A.M., Wang,J., Barry,B.J., Zhang,X., Hill,R.S., Partlow,J.N., Rozzo,A. *et al.* (2014) Somatic mutations in cerebral cortical malformations. *N. Engl. J. Med.*, **371**, 733–743.
55. Freed,D. and Pevsner,J. (2016) The contribution of mosaic variants to autism spectrum disorder. *PLoS Genet.*, **12**, e1006245.
56. Tyburczy,M.E., Dies,K.A., Glass,J., Camposano,S., Chekaluk,Y., Thorner,A.R., Lin,L., Krueger,D., Franz,D.N., Thiele,E.A. *et al.* (2015) Mosaic and intronic mutations in TSC1/TSC2 explain the majority of TSC patients with no mutation identified by conventional testing. *PLoS Genet.*, **11**, e1005637.
57. Acuna-Hidalgo,R., Bo,T., Kwint,M.P., Vorst,M.V.D., Pinelli,M., Veltman,J.A., Hoischen,A., Vissers,L.E.L.M. and Gilissen,C. (2015) Post-zygotic point mutations are an underrecognized source of De Novo genomic variation. *Am. J. Hum. Genet.*, **97**, 67–74.
58. Maxmen,A. (2011) Exome sequencing deciphers rare diseases. *Cell*, **144**, 635–637.
59. Ronemus,M., Iossifov,I., Levy,D. and Wigler,M. (2014) The role of de novo mutations in the genetics of autism spectrum disorders. *Nat. Rev. Genet.*, **15**, 133–141.
60. Treangen,T.J. and Salzberg,S.L. (2012) Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nat. Rev. Genet.*, **13**, 36–46.
61. Kong,A., Frigge,M.L., Masson,G., Besenbacher,S., Sulem,P., Magnusson,G., Gudjonsson,S.A., Sigurdsson,A., Jonasdottir,A., Jonasdottir,A. *et al.* (2012) Rate of de novo mutations and the importance of father's age to disease risk. *Nature*, **488**, 471–475.
62. Alexandrov,L.B., Nik-Zainal,S., Wedge,D.C., Aparicio,S.A.J.R., Behjati,S., Biankin,A.V., Bignell,G.R., Bolli,N., Borg,A., Børresen-Dale,A.-L. *et al.* (2013) Signatures of mutational processes in human cancer. *Nature*, **500**, 415–421.