Bioinformatics, 35(4), 2019, 571–578
doi: 10.1093/bioinformatics/bty677
Advance Access Publication Date: 2 August 2018
Original Paper

OXFORD

Genome analysis

Detection of de novo copy number deletions from targeted sequencing of trios

Jack M. Fu[1], Elizabeth J. Leslie[2], Alan F. Scott[3], Jeffrey C. Murray[4], Mary L. Marazita[5], Terri H. Beaty[6], Robert B. Scharpf[7] and Ingo Ruczinski[1,*]

[1]Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, 21205 MD, USA, [2]Department of Human Genetics, Emory University, Atlanta, 30322 GA, USA, [3]Department of Medicine, Johns Hopkins School of Medicine, Baltimore, 21205 MD, USA, [4]Department of Pediatrics, Carver College of Medicine, University of Iowa, Iowa City, 52242 IA, USA, [5]Department of Oral Biology, University of Pittsburgh, Pittsburgh, 15213 PA, USA, [6]Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Baltimore, 21205 MD, USA and [7]Department of Oncology, Johns Hopkins School of Medicine, Baltimore, 21205 MD, USA

*To whom correspondence should be addressed.

Associate Editor: John Hancock

Received on February 28, 2018; revised on July 25, 2018; editorial decision on July 26, 2018; accepted on August 1, 2018

Abstract

Motivation: De novo copy number deletions have been implicated in many diseases, but there is no formal method to date that identifies de novo deletions in parent-offspring trios from capture-based sequencing platforms.

Results: We developed Minimum Distance for Targeted Sequencing (MDTS) to fill this void. MDTS has similar sensitivity (recall), but a much lower false positive rate compared to less specific CNV callers, resulting in a much higher positive predictive value (precision). MDTS also exhibited much better scalability.

Availability and implementation: MDTS is freely available as open source software from the Bioconductor repository.

Contact: ingo@jhu.edu

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

1 Introduction

Copy number variants (CNVs) are a major contributor of genome variability in humans (Zarrei *et al.*, 2015), and frequently underlie the etiology of disease (Gilissen *et al.*, 2014; Pinto *et al.*, 2010; Walsh *et al.*, 2008; Wellcome Trust Case Control Consortium, 2010; Zhang and Lupski, 2015). *De novo* CNVs, especially *de novo* deletions, are of interest as they have the potential to play a functional role in the genesis of a disease phenotype (Georgieva *et al.*, 2014; Glessner *et al.*, 2014; van Bon *et al.*, 2016; Veltman and Brunner, 2012). Over the last decade, next generation sequencing (NGS) has become routine and widespread (Metzker, 2010; Shendure and Ji, 2008), permitting the assessment of CNVs based on hundreds of millions of short reads observed in each sample. The advantages of some types of NGS approaches for CNV assessment, compared to single nucleotide polymorphism (SNP) arrays, may include higher and more uniform coverage, better quantitation yielding more precise estimates of DNA copy number, and higher resolution for break point detection (Alkan *et al.*, 2011; Meyerson *et al.*, 2010). Computational methods to detect CNVs from NGS short reads can generally be categorized into approaches based on discordant read mapping, split read mapping, read depth, *de novo* assembly, or a combination of these approaches (Zhao *et al.*, 2013). Due to the differences in the attempted capture, methodologies for whole genome sequencing (WGS), whole exome sequencing (WES) and targeted sequencing (TS) platforms differ substantially, with TS and WES platforms primarily relying on read depth (Tattini *et al.*, 2015).

A large number of methods for detecting CNVs in independent samples are available for all types of NGS data (Bansal *et al.*, 2014;

Bellos *et al.*, 2014; Bellos and Coin, 2014; Fromer and Purcell, 2014; Krumm *et al.*, 2012; Kuilman *et al.*, 2015; Li *et al.*, 2012; Nord *et al.*, 2011; Packer *et al.*, 2016; Talevich *et al.*, 2016). However, there is no method to date that identifies *de novo* CNVs in parent-offspring trios from capture-based TS and WES platforms. For WGS platforms, the software TrioCNV jointly calls CNVs in parent-offspring trios (Liu *et al.*, 2016) using a hidden Markov model (HMM) with 125 possible underlying states to segment the sequencing data (5 possible underlying states per sample: two-copy deletion, one-copy deletion, normal, one-copy duplication, multiple-copy duplication). Its performance in TS or WES platforms however is not well described. In CANOES, also HMM based, inference for *de novo* copy number deletions in TS and WES data is obtained post-hoc by comparing single-sample derived CNV calls. For each sample of the trio, the observed read counts are modeled using negative binomial distributions, and the respective variances are estimated using a regression-based approach based on selected reference samples (Backenroth *et al.*, 2014). However, such approaches do not fully leverage the Mendelian relationship between parents and offspring to delineate *de novo* CNVs. The loss of statistical power for delineating *de novo* CNVs by post-hoc methods has been demonstrated previously in CNV calls from SNP array data (Scharpf *et al.*, 2012; Wang *et al.*, 2008).

The motivating example in this manuscript is a targeted re-sequencing study we recently carried out in 1409 Asian and European case-parent trios ascertained by non-syndromic orofacial cleft probands, targeting 13 regions previously implicated in candidate genes and genome-wide association studies (GWASs) (Leslie *et al.*, 2015). The study successfully confirmed 48 *de novo* nucleotide mutations, and provided strong evidence for several specific alleles as contributory risk alleles for non-syndromic clefting in humans. Choosing two of these *de novo* nucleotide variants for functional assays, we showed one mutation in PAX7 disrupted the DNA binding of the encoded transcription factor, while the other mutation disrupted the activity of a neural crest enhancer downstream of FGFR2 (Leslie *et al.*, 2015). However, for the majority of trios, we were not able to identify a genetic cause underlying the proband's oral cleft. Since *de novo* deletions have previously been shown to underlie oral cleft risk (Salahshourifar *et al.*, 2012; Tan *et al.*, 2013; Younkin *et al.*, 2014), we speculated that in addition to *de novo* nucleotide variants, *de novo* deletions in the 13 targeted regions also contribute to clefting for some of our trio's probands.

In this manuscript, we present a novel method to delineate *de novo* deletions from TS of trios. We propose a novel capture-based definition of targets (using median read depth as the defining metric for bins underlying the algorithm, instead of using a uniform number of base pairs), normalize copy number counts using the entire study population, and utilize a 'minimum distance' statistic based on normalized read count summaries, aiming to further reduce shared sources of technical variation between offspring and parents within a trio. We characterize the sensitivity, specificity and positive predictive value (PPV) of MDTS on simulated data to benchmark its performance relative to the closest existing methods TrioCNV (Liu *et al.*, 2016) and CANOES (Backenroth *et al.*, 2014). We show that properly addressing the capture in TS data is critical, and thus, methods specifically developed for WGS data (e.g. TrioCNV) do not perform well for TS data. Compared to CNV callers designed for capture based sequencing data that do not exploit the family design (e.g. CANOES), MDTS has similar sensitivity but a much lower false positive rate, resulting in a much higher PPV. In the analysis of the 6.7 Mb TS oral cleft data, which identified one *de novo* deletion in the gene TRAF3IP3 (a suspected regulator of IRF6), MDTS also exhibited much better scalability.

2 Materials and methods

Our novel MDTS introduces two novel algorithmic aspects. First, MDTS employs bins of varying sizes based on read depth (Fig. 1, A–D) as compared to the common standards of using either uniform, non-overlapping bins defined by the number of nucleotide base pairs (default in TrioCNV: tiled, non-overlapping 200 bp bins) or probe-based coordinates (default in CANOES: the genomic coordinates of the designed capture baits). Second, MDTS fully exploits the trio design to infer *de novo* deletions (Fig. 1, E–G), as compared to processing the trio samples separately and carrying out post-hoc inference. To demonstrate that both of these algorithmic features are important in the delineation of *de novo* deletions, and to quantify their relative contributions to sensitivity, specificity and PPV, we compare the default implementations of MDTS and CANOES in the following section, plus MDTS based on the 'probe-based' bins (MDTS:p) and CANOES based on the dynamically sized 'MDTS bins' (CANOES:b).

2.1 Sample and target region selection

The original study population included 1409 case-parent trios comprised of 4227 individuals of Asian or European ancestry from Europe, the United States, China and the Philippines (Leslie *et al.*, 2015, Supplementary Table S1). Thirteen genomic regions spanning 6.7 Mb were selected for sequencing based on prior association and/or linkage studies, targeting both coding and non-coding sequence at each locus (Leslie *et al.*, 2015, Table 1).

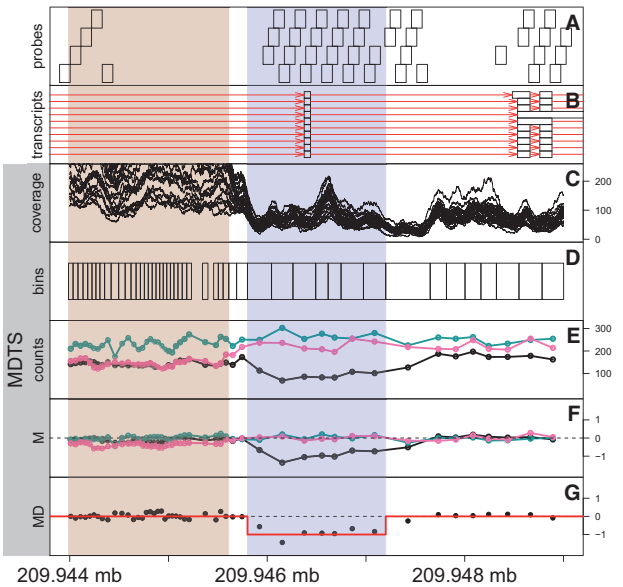

Fig. 1. Schematic flowchart of the MDTS method, from bin to CNV delineation. (**A**) Design probes in the genomic region between 209.944 and 209.948 Mb of chromosome 1. The probes are approximately 120 bp long, and often overlap by 60 bp. (**B**) Transcripts (red lines) from the GencodeV27 annotation. Ten transcripts of TRAF3IP3 contain the exon (white boxes) in the region shaded blue. (**C**) Basepair coverage (read depth) derived from the 25 samples randomly selected to calculate MDTS bins. The region indicated by the rose color was flagged by MDTS for high variability. (**D**) MDTS bins calculated from read depth, leading to wider bins when coverage is low (and vice versa). (**E**) Read depths for the MDTS bins among the three DS10826 family members (proband in black). (**F**) Normalized counts (M-scores) for the three DS10826 family members. (**G**) The minimum distance for family DS10826, and the outcome from CBS segmentation (red line), inferring a candidate *de novo* deletion (Color version of this figure is available at *Bioinformatics* online.)

Table 1. MDTS inferred de novo deletions in the oral cleft data

	Start locus	End locus	Size	Family	MD
Chromosome 1	209 945 655	209 947 210	1556	DS10826	−0.90
Chromosome 8	129 614 522	129 616 078	1557	DS12329	−0.82
Chromosome 8	130 113 612	130 132 753	19 142	DS11025	−0.88

The region on chromosome 1 (top row) is a genuine *de novo* heterozygous deletion of approximately 1556 base pairs in the proband of family DS10826, inferred using the minimum distance and supported by aberrantly spaced reads (Fig. 3, left column). The region of about 1557 base pairs near 129.6 Mb on chromosome 8 (middle row) likely is a false positive identification, inferred based on read depth and the minimum distance, but not supported by aberrantly spaced reads (Fig. 3, right column). The region of about 19 kb near 130.1 Mb on chromosome 8 (bottom row) stems from an unusual Mendelian event in family DS11025 outside a copy number polymorphism (Fig. 4, left column). MD: average minimum distance in the respective regions.

2.2 Library preparation, sequencing and alignment

Multiplexed libraries were constructed with 1 μg of native genomic DNA according to standard Illumina protocol with modifications as follows, described in Leslie *et al.* (2015): (i) DNA was fragmented with a Covaris E220 DNA Sonicator (Covaris) to range in size between 100 and 400 bp; (ii) Illumina adaptor-ligated library fragments were amplified in four 50 ml PCR reactions for 18 cycles; and (iii) solid phase reversible immobilization (SPRI) bead cleanup was used for enzymatic purification throughout the library process, as well as final library size selection targeting 300–500 bp fragments. NimbleGen custom target probes were designed to the target region and hybrid capture on pools of 96 indexed samples per capture was performed. Each capture pool was sequenced on two lanes of Illumina HiSeq for an average of ∼40 Gb per lane or ∼835 Mb per sample. 100 bp paired-end reads were mapped to the GRCh37-lite reference sequence by BWA v.0.5.912 (Li and Durbin, 2009).

2.3 MDTS algorithm

2.3.1 Definition of bins and M scores

Due to the prevalence of off-target capture and heterogeneity of coverage within targeted regions, we utilized an empirical approach to define the MDTS bins for computing read depth. Specifically, we randomly sampled 25 subjects and calculated the coverage statistics in each sample across the autosomes. A set of contiguous proto-regions were identified as the set of all basepairs where at least one of the samples had observed coverage of 10× or more. The size of the bins were a consequence of the minimum coverage parameter in MDTS. As proto-regions harbored substantial heterogeneity in size depending on both probe density and capture efficiency, the final bins were generated by sequentially partitioning the proto-regions into smaller, non-overlapping regions where the median number of reads across the 25 subsamples in each region reached the desired minimum coverage. Setting this parameter to 160 separated the copy neutral state from a rare heterozygous deletions spanning a single bin by more than 6 standard deviations, assuming a Poisson model for the counts. Bins were excluded if the average mappability of a bin was less than 0.75, or if the average GC content was outside a 'normal' range defined as [0.15, 0.85]. Subsequently, the number of reads overlapping the bins were counted for all samples. The raw count data were organized in a 'bin by sample' matrix. We applied a $\log_2(\text{count}+1)$ transformation to reduce skew. Each cell of the matrix was centered by the column and row medians, in that order. The resulting scores for each sample were further adjusted for average GC content and 100mer mappability of their respective bins, using a locally weighted scatterplot smoother (loess) fit to produce *M* scores, a relative measures of DNA copy number, with an expected value of 0 for a copy-neutral DNA segment, and -1 for a single copy deletion (unless there is a CNP).

2.3.2 Minimum distance

To infer *de novo* deletions we utilize the Minimum Distance statistic, previously defined for SNP array data (Scharpf *et al.*, 2012). In brief, at each bin we considered the difference in *M* scores between the offspring (O) and the father (F), calculated as $M_O - M_F$, and denote this difference as δ_F. We calculated the equivalent distance of offspring and mother, and denote this difference as δ_M. The Minimum Distance between parents and offspring at a bin is defined as the smaller of those two differences when comparing their absolute values:

$$d = \arg\min_{\delta \in \{\delta_F, \delta_M\}} |\delta| \quad (1)$$

2.3.3 Filtering and segmentation

Of the 1409 families, 383 were removed prior to MDTS bin calculation for experimental design insufficiencies. For these families, the family members were either run in different batches, or did not pass basic quality control as noted by the reporting lab. An additional 8 families were excluded from the analysis based on Minimum Distances summary statistics (lag10 auto-correlation > 0.4 and/or variance > 0.05). Circular Binary Segmentation (CBS) (Olshen *et al.*, 2004; Venkatraman and Olshen, 2007), implemented in the Bioconductor package `DNAcopy`, was used for each targeted region to segment the Minimum Distances across the bins for each trio. CBS computes a permutation reference distribution of the input Minimum Distances to infer change points for copy number. As this is a random process by default, we fixed a seed `set.seed(137)` in R to ensure reproducibility of our results. We required the minimum number of bins in any segment to be at least 3. In general, default input parameters were used, except using $\alpha = 0.001$ as the minimum significance required in the CBS t-tests to infer a change point. Further, we allowed change points to be undone when the difference in means was less than 4 standard deviations (`undo.splits='sdundo'` in conjunction with `undo.SD=4`). Candidate *de novo* heterozygous deletions were identified as regions where the segmented Minimum Distance was within 0.3 of the theoretical value of –1. To reduce the likelihood of false positives based on failures in the normalization process (caused by the existence of CNPs or technical anomalies), regions of high variability were identified as bins where more than 5% of samples had *M* scores outside the interval [−0.5, 0.5]. MDTS reported *de novo* deletions only when more than half of the bins in the candidate region were not flagged in such a manner.

2.4 Alternative approaches

2.4.1 CANOES

This algorithm was designed for capture-based WES and TS data, but the statistical inference does not explicitly take the familial relationship into account. Assessment of *de novo* copy number events in CANOES is based on a post-hoc comparison of the inferred copy number states of the individual samples. The default binning scheme in the algorithm utilizes the bait design coordinates, but MDTS bins can also be used as input. A simple modification had to be made to the CANOES R code, publicly available at `www.columbia.edu/~ys2411/canoes/`, to make it scalable for our simulation study and oral cleft data analysis. For large sample sizes (here, $n = 3054$ in the oral cleft study) the calculation of

the $n \times n$ covariance matrix between bin read counts of samples to locate reference samples for a given individual is computationally very intensive. In the original R code this is carried out for each sample (within the `for() loop`), but actually has to be carried out only once (outside the `for() loop`).

2.4.2 TrioCNV

In comparison to CANOES, this algorithm explicitly models the proband-parent trio relationship, however was designed for WGS data (i.e. non-capture based sequencing data). The default binning scheme for the inference is based on subdividing the genome into non-overlapping 200 bp windows. We restricted these bins to those in the 6.7 Mb targeted for sequencing (Leslie *et al.*, 2015). In the simulation study and the oral cleft data analysis we used the TrioCNV default parameters, with two exceptions: We reduced the value for the argument `min_distance`, which specifies the distance between adjacently called CNVs to be merged, from the default 10 000 to 1000. We also changed the value for the argument `gc_bin_size`, from its default value of 1 to 2. This value determines the grouping of bins for the estimation of the emission probabilities in the Hidden Markov Model. The default value of 1 did not produce a sufficient number of bins for certain GC values in the capture based data, resulting in JAVA runtime errors thrown.

2.5 Simulation study

We sampled with replacement 1000 case-parent trios from the 1018 families that passed QC. For each instance, we simulated read data based on the TS data for that trio. We first sampled 10 non-overlapping regions among MDTS regions that passed the normalization criterion described above. Of the 10 regions, 5 were designated to harbor *de novo* deletions and 5 were designated to harbor inherited deletions of sizes 250, 500, 1000, 2000 and 4000 bp. The 5 *de novo* deletion spike-ins were achieved by randomly and independently dropping reads that overlapped the selected regions with probability 0.5 in the BAM file of the proband. The 5 inherited deletions were generated by randomly and independently dropping reads overlapping the respective regions with probability 0.5 in the BAM files of the proband and one parent. Split reads were not simulated as all methods compared here are based on read-depth. We compared the performances of MDTS, CANOES and TrioCNV, using default and alternative binning schemes. Specifically, we assessed the performances of MDTS with default read-depth based bins (MDTS), MDTS with bins based on bait design coordinates as defined in CANOES (MDTS:p), CANOES with MDTS bins (CANOES:b), CANOES with default bins (CANOES), TrioCNV with MDTS bins (TrioCNV:b) and TrioCNV with restricted genomic bins as described above (TrioCNV). For CANOES and CANOES:p, the CNV calling was carried out for each family member. Inferred deletions in the proband found to be at least 25% covered by a called deletion in at least one of the parents were deemed to be inherited, otherwise deletions in the proband were considered *de novo*. The spiked-in *de novo* and inherited deletions were considered called if 25% of the deletion was covered by candidates reported. Alternative thresholds of $> 0\%$ (any overlap) and 50% (at least half of the deletion was identified) were also considered.

3 Results

3.1 Simulation study

MDTS and CANOES produced somewhat similar results for sensitivity (recall) among *de novo* deletions of 1 kb or larger, while CANOES had better sensitivity for very small *de novo* deletions. As expected, the algorithms using the smaller ('probe-based') bins faired slightly better for small *de novo* deletions, while using read depth based bins ('MDTS bins') had higher sensitivity for 1 kb *de novo* deletions or larger (Fig. 2A, Supplementary Table S1). These findings remained the same under other definitions of 'overlap' between called and simulated deletions (Supplementary Fig. S1). Very pronounced differences were observed with regards to the number of false positive identifications. Depending on size, up to 8% of inherited deletions were incorrectly identified as *de novo* by CANOES using the default 'probe-based bins' (increasing to about 15% for CANOES:b, i.e. when using 'MDTS bins'), while MDTS was extremely robust towards this type of mistake. This was also true when 'probe-based bins' were used in the MDTS algorithm (e.g. MDTS:p), highlighting the importance of fully exploiting the trio design when inferring *de novo* deletions (Fig. 2B, Supplementary Table S2).

In addition, MDTS incorrectly identified 2 small *de novo* deletions of 374 and 637 base pairs in this simulation study, while CANOES yielded 2967 false positives with a median width of 361 base pairs (ranging from 121 to 24 474 base pairs). This number was reduced to 114 false positives when our proposed read depth based bins ('MDTS bins') were used in the CANOES algorithm (e.g. CANOES:b), but these inferred deletions were generally larger in size with a median width of 2485 base pairs and a range of 206 to

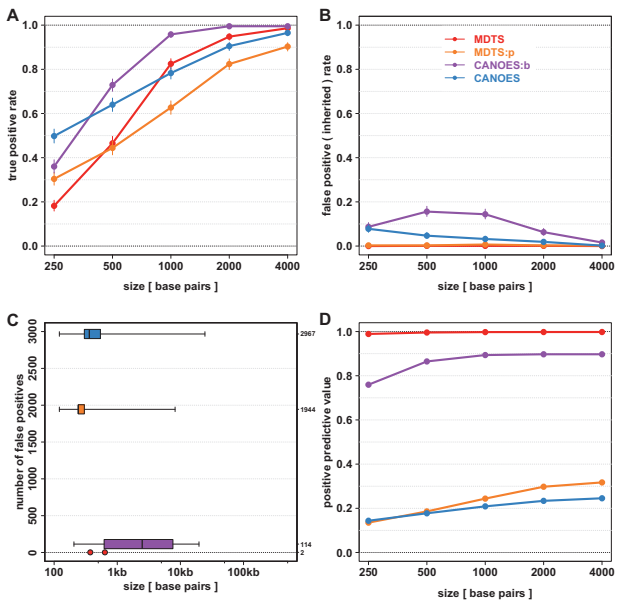

Fig. 2. Simulation results to assess sensitivity, specificity and positive predictive value of four different algorithms to infer de novo deletions. True positive rate (sensitivity, y-axis) among 1000 iterations for simulated *de novo* deletions of various sizes (x-axis). Point estimates are shown as circles together with Binomial 95% confidence intervals. (**B**) False positive rate (specificity) among 1000 iterations for simulated inherited deletions of various sizes. (**C**) Number of additional false positive identifications from the simulation experiment (y-axis), with length distribution on the logarithmic scale (x-axis) shown as boxplots. MDTS with the newly defined bins only produced two additional false positives, which are shown as points. (**D**) Positive predictive value based on the true positive rate in panel (**A**) and the false positives in panel (C). Colors indicate the algorithms. MDTS and CANOES refer to the respective algorithms as implemented, MDTS:p refers to MDTS based on 'probe-based bins', CANOES:b refers to CANOES based on the non-uniform read depth based 'MDTS bins' (Color version of this figure is available at *Bioinformatics* online.)

19 709 base pairs. The importance of using read depth based bins in the algorithms to control false positive identifications was evident, as MDTS built on probe-based coverage (MDTS:p) also faired a lot worse than MDTS (Fig. 2C, Supplementary Table S3). These differences in the numbers of false positive identifications observed among these algorithms also resulted in substantial differences when estimating the PPV. The almost complete absence of false positive identifications in MDTS resulted in PPVs approaching 100%, while CANOES did not exceed 25% even for the large *de novo* deletions. CANOES:b on the other hand achieved about 90% PPV, highlighting the importance of using read depth based bins (Fig. 2D, Supplementary Table S4).

As expected, TrioCNV did not perform well in the simulation study due to its design for WGS (i.e. non-capture) data. TrioCNV with default 200 bp genomic bins was unable to detect any *de novo* deletions, and TrioCNV with MDTS bins only achieved at most 2% sensitivity even for the larger deletions.

3.2 Oral cleft case study

Of the full complement of 4227 samples, 3054 samples in 1018 case-parent trios passed sequencing quality control metrics. Among these families, the MDTS binning procedure generated 25 305 bins, spanning just over 6.3 Mb of the targeted 6.7 Mb autosomal region. The bins ranged in size from 19 to 2956 bp, with a median size of 220 bp. The mean bin coverage ranged from 24× to 305× across these samples, with a median of 66× and a median absolute deviation (standard deviation) of 14.0 (17.8) (Supplementary Fig. S2).

MDTS identified three candidate *de novo* deletions (Table 1). The first candidate spanned a 1.6 kb segment on chromosome 1 with an average Minimum Distance of −0.90 across 7 bins, and was strongly supported as a *de novo* deletion by the presence of improperly paired reads spanning this segment (Fig. 3, left column). The average read depth for the proband in that region was 714, while a read depth of 1318 was expected for a copy neutral state. This finding was further corroborrated by whole genome sequencing data available for this trio (Supplementary Fig. S3). The second candidate region spanned a 1.6 kb segment on chromosome 8 with an average Minimum Distance of −0.82 across 7 bins. The average read depth for the proband in that region was 740, which compared to an expected read depth of 1380 for a copy neutral state, suggesting this proband carried a heterozygous deletion. In contrast to the region on chromosome 1 however, no improperly paired reads spanning this segment were observed, rendering this finding somewhat less conclusive. Thus, this region could also represents a false positive identification (Fig. 3, right column). Mendelian inconsistencies among trio genotypes can also indicate a *de novo* deletion (Ting *et al.*, 2006, 2007), while heterozygous genotypes in the proband provide strong evidence against *de novo* deletions, however neither result was observed in this short 1.6 kb region for family DS12329 (only 1 variant was reported in the vcf files as 0/0, 0/1, 0/1 for the child and the parents, respectively). The third candidate region spanned a 19 kb segment on chromosome 8, with an average Minimum Distance of −0.88 across 74 bins. The apparent deletion in the proband of family DS11025 however was not *de novo*, but inherited from a parent with zero copies (Fig. 4, left column). This represents a rather uncommon occurrence, as homozygous deletions typically are only observed for copy number polymorphisms (Fig. 4, right column), while the 19 kb segment on chromosome 8 was only observed for this one family. In total, MDTS detected and flagged two copy number polymorphic regions, a 7.1 kb segment

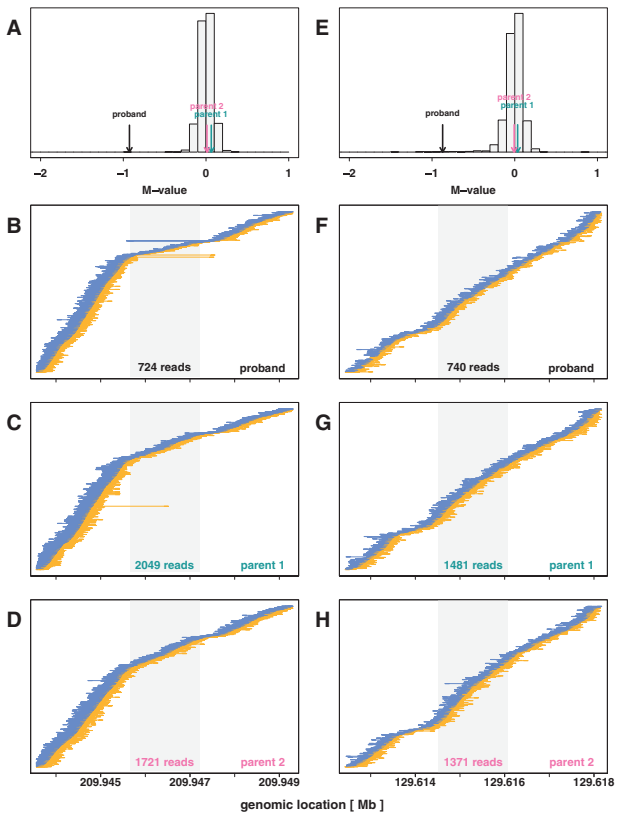

Fig. 3. Data underlying inferred de novo heterozygous deletions in two probands. **[Left Column]** Evidence for a *de novo* heterozygous deletion on chromosome 1 for the proband in family DS10826. (**A**) The average of the M scores of the proband (-0.93, black arrow), the parents (0.06 and 0.01, green and pink arrows, respectively) and all other subjects (gray histogram) between loci 209 945 655 and 209 947 210 on chromosome 1. The proband's average of the M scores near −1, compared to the values near zero for all other samples including the parents, is consistent with a *de novo* deletion of one allele in this region. (**B–D**) Read-pairs observed among the members of family DS10826 near the region with the *de novo* heterozygous deletion. The read-pair locations, mapped to the hg19 reference genome, are shown as thick ends connected by thin lines (positive strands shown in yellow, negative strands shown in blue), and sorted by beginning location of mate 1 of the read-pair (e.g. yellow lines are left aligned, blue lines are right aligned). Read-pairs mapped far apart, apparent as a long line, are indicative of a deletion between the ends. A Z-shaped signature of read pairs flanked by such discordant reads as seen for the proband is strong evidence for a 1-copy DNA deletion. The gray region in these panels indicate the inferred 1556 bp heterozygous *de novo* deletion region in the proband's genome. The number at the bottom of the grey regions in each panel indicates the total number of reads mapped to the inferred *de novo* deletion. **[Right Column]** A possible false positive identification of a *de novo* heterozygous deletion on chromosome 8 for the proband in family DS12329. (**E**) The average of the M scores of the proband (−0.87), the parents (0.035 and −0.007, green and pink arrows, respectively), and all other subjects (gray histogram) between loci 129 614 522 and 129 616 078 on chromosome 8. The proband's average of the M scores near −1, compared to the values near zero for all other samples including the parents, is consistent with a *de novo* deletion of one allele in this region. (**F–H**) Read-pairs observed among the members of family DS12329 near the region with the inferred *de novo* heterozygous deletion. The absence of discordant reads and the Z-shaped signature is evidence against a 1-copy DNA deletion (Color version of this figure is available at *Bioinformatics* online.)

on chromosome 1 and a 3.2 kb segment on chromosome 8 (Supplementary Table S5).

CANOES also identified the true *de novo* deletion in the proband of family DS10826, did not identify the inherited deletion in family DS11025, and did not report the inconclusive MDTS identification in

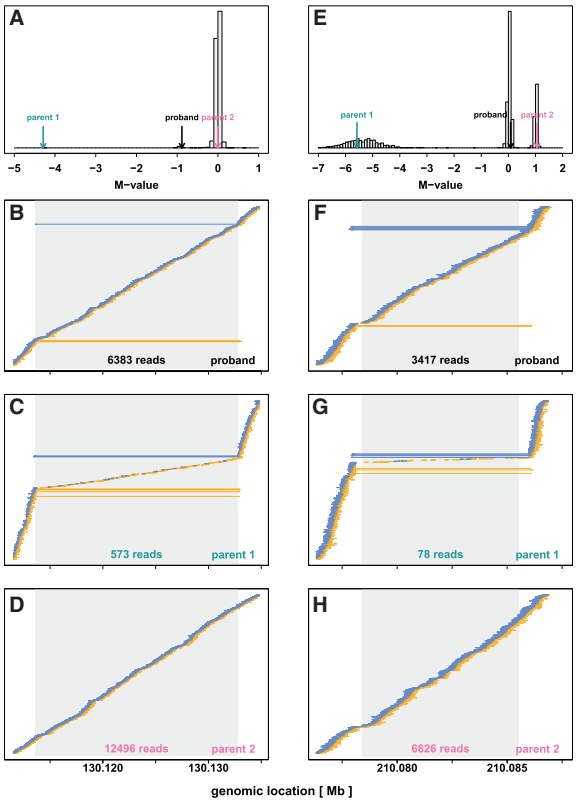

Fig. 4. Examples of Mendelian events with a heterozygous deletion in the proband. **[Left Column]** A rare Mendelian inheritance event observed on chromosome 8 in family DS11025. (**A**) The average of the M scores for the proband (−0.88, black arrow) and the parents (−4.3 and −0.01, green and pink arrows, respectively), and all other subjects (gray histogram) between loci 130 113 612 and 130 132 753 on chr8. This is consistent with a heterozygous deletion for the proband, inheriting one copy of the allele from the copy-neutral parent 2, and the deletion from parent 1 showing a homozygous deletion. (**B–D**) Read-pairs observed among the members of family DS11025 near the region with the inferred Mendelian inheritance event, using the same plotting approach as described in the Figure 3 legend. The Z-shaped signature of a substantial number of read pairs flanked by aberrantly spaced reads seen for the proband again is evidence for a 1-copy (heterozygous) deletion. The Z-shaped signature sandwiching very few (presumably incorrectly mapped) reads for parent 1 is evidence for a 2-copy (homozygous) deletion. The read pairs for parent 2 show a copy-neutral state. The gray region in these panels indicate the inferred 18 956 bp inherited deletion region. The number at the bottom of the grey regions in each panel indicates the total number of reads mapped to the inferred *de novo* deletion. **[Right Column]** A Mendelian inheritance event observed at a copy number polymorphic region on chromosome 1 in family DS11230. (**E**) The average of the M scores for the proband (0.084, black arrow) and the parents (−5.58 and 1.06, green and pink arrows, respectively), and all other subjects (gray histogram) between loci 210 078 417 and 210 085 527 on chr1. This again is consistent with a heterozygous deletion for the proband, inheriting one copy of the allele from the copy-neutral parent 2, and the deletion from parent 1 showing a homozygous deletion. Due to the polymorphic nature of this region, the initial median normalization failed to correctly center the copy neutral state at zero, which was subsequently inferred by the post-segmentation filter. (**F–H**) Read-pairs observed among the members of family DS11230 near the region with the inferred Mendelian inheritance event, supporting the inferred 7111 bp heterozygous (homozygous) deletion in the proband (parent 1) (Color version of this figure is available at *Bioinformatics* online.)

family DS12329. Consistent with the general findings in the simulation study, CANOES also reported a large number of additional *de novo* deletions. In the targeted 6.7 Mb region—representing only 0.2% of the genome—the algorithm identified an additional 2969 *de novo* deletions among the 1018 families, i.e. about 3 *de novo* deletions per trio on average. Among those 2969 identifications, 2702 had a Minimum Distance (calculated from probe-based coverage) outside the [−1.3, −0.7] interval, not consistent with *de novo* deletions (Supplementary Fig. S4). The remaining 267 reported *de novo* deletions with Minimum Distances in the interval [−1.3, −0.7] were small (median width 361 bp), and none had improper read-pairs spanning the length of the indicated deletion (Supplementary Fig. S5). CANOES:b utilizing the MDTS determined bins had the same calls as CANOES reported above for families DS10826, DS11025 and DS12329, but only returned 79 additional *de novo* deletions (though only 28 of those overlapped with any of the 2969 deletions identified by CANOES). Among those 79 identified deletions, 67 had average Minimum Distances outside the [−1.3, −0.7] interval, and so were inconsistent with *de novo* deletions (Supplementary Fig. S6). Among the remaining 12 apparent *de novo* deletions (median width 619 bp) with Minimum Distances in the interval [−1.3, −0.7] one is actually an inherited homozygous deletion (Supplementary Fig. S7), while the other 11 are located in flagged regions of highly variable normalized depth of coverage (Supplementary Fig. S8).

TrioCNV with default bins (tiled non-overlapping 200 bp bins within the targeted region) did not report any *de novo* deletions among these 1018 families. In particular, the algorithm failed to identify the true *de novo* deletion on chromosome 1 of the DS10826 proband. TrioCNV with MDTS bins did identify 24 *de novo* deletions, however, 23 of those were actually inherited deletions (Mendelian events) within the chromosome 1 copy number polymorphism (Supplementary Fig. S9). The remaining inferred *de novo* deletion supported by only one bin had a Minimum Distance of −0.94, but no improperly mapped read-pairs spanning the deletion which would support a true *de novo* deletion (Supplementary Fig. S10). This version of the algorithm also failed to identify the *de novo* deletion on chromosome 1 of the DS10826 proband.

3.3 Scalability

MDTS completed the analysis of the 1018 oral cleft trios in under 29 h using a single core, peaking at 15 G of memory in the binning step (Supplementary Tables S6 and S7). The run time was cut to less than 6 h when employing the distributed computing option with 15 cores, albeit at the cost of increasing the peak memory usage to 160 G during the counting step. For CANOES, even after editing the supplied R code (which resulted in an almost ten-fold speed-up of the inference), this analysis still required 1310 h of CPU times for a single core, but only 14 G of memory. TrioCNV, using default parameters except for the distance between adjacent CNVs to be merged and the GC content bin range (see Section 2) had a comparable computational footprint to MDTS, requiring 34 CPU hours and 11 G of memory to complete the analysis. The usage of MDTS bins slightly reduced the run time for TrioCNV and cut the CANOES CPU time about in half, though the latter was still an order of magnitude slower than MDTS and TrioCNV. MDTS based on probe based bins (MDTS:p) required additional CPU time for the inference compared to the default (MDTS), presumably due to the auto-correlation of the Minimum Distance estimates (resulting from the overlapping design probes) passed to CBS, making break point selection more challenging.

4 Discussion

In this manuscript we presented the Minimum Distance for Targeted Sequencing (MDTS) approach for delineating *de novo*

copy number deletions simultaneously across multiple trios from TS data. In a simulation study, our approach had a sensitivity competitive with existing methods, but to our knowledge, MDTS is the first caller that rarely generates any false positives. In our simulation study, this approach resulted in a positive predictive value of nearly 100%. We showed this improvement is largely owed to two novel algorithmic features. MDTS employs non-uniformly sized bins based on read depth instead of using either uniform, non-overlapping bins defined by the number of nucleotide base pairs or coordinates of capture probes, and further, MDTS fully exploits the trio design by using a 'minimum distance' statistic to quantify differences in read depths between the offspring and the parents, thereby reducing shared sources of technical variation. We note similar results (equal sensitivity but much improved specificity) were observed for detection of *de novo* deletions based on SNP array data when the Minimum Distance approach was employed, and compared to the results from the trio based PennCNV algorithm (Scharpf *et al.*, 2012; Wang *et al.*, 2008). Summarizing the trio data at each locus (probe for SNP arrays or bins for sequencing data) and segmenting these statistics resulted in an estimating procedure with much lower dimensionality than that of a HMM (as used for example in CANOES and TrioCNV). A smaller parameter space is less likely to over-fit, and to generate false positive identifications. Further, fitting a HMM induces an empirical process governing the rate and lengths of these deletions, which may not be realistic as *de novo* deletions are very rare, and could be very small or very large. It should also be noted that MDTS was designed with the sole intent to detect *de novo* deletions in trios, and thus, is much more limited in scope than other CNV callers such as CANOES and TrioCNV (although in principle the MDTS algorithm could also be adapted to detect *de novo* amplifications by applying a positive threshold to the segmented means).

Split reads provide additional compelling evidence for the presence of a copy number deletions, and allow for base pair resolution of break point detection. However, mapping split reads is computationally infeasible for larger deletions unless a candidate has already been identified, and thus, methods based on read depth bins are usually employed to find larger deletions. MDTS is such a method primarily based on read depth, and similar to other read-depth based CNV callers, MDTS has problems identifying very small deletions. In our simulation study, MDTS nonetheless achieved greater than 80% sensitivity for *de novo* deletions of 1 kb, and virtually 100% sensitivity for *de novo* deletions of 5 kb. We have also implemented functionality allowing for post-hoc inspection of the read ensemble mapped to a region around any putative deletion. In particular the presence of a Z-shaped signature of read pairs flanked by discordant reads–as seen in the suspected IRF6 regulator for the proband of family DS10826–provides further support for a deletion, and uses information in addition to read depth alone. As the MDTS specificity is very high and *de novo* deletions are rare, the number of candidate deletions to be inspected is low. We queried BAM files to locate split reads that are in the vicinity of a putative deletion. We used SAMtools (`samtools.sourceforge.net`) to extract split alignments and BLAT (`genome.ucsc.edu/cgi-bin/hgBlat`) to re-align un-mapped sequences, but were unsuccessful in locating supporting split reads. Thus, no attempts were made to employ LUMPY, arguably the most common CNV caller currently used, to call *de novo* deletions in our data, as its performance heavily relies on such split reads (Layer *et al.*, 2014). Further, LUMPY was intended for WGS data and does not account for family structure, thus being less applicable for comparison than TrioCNV and CANOES. Lastly, LUMPY depends on an external read depth caller, which we provide here for TS data in trios.

We also applied our method to 1305 case-parent trios with 6.7 Mb of TS data of regions previously implicated in oral cleft. We detected one *de novo* deletion in the gene TRAF3IP3 on chromosome 1q32 in a Caucasian proband with a cleft lip. TRAF3IP3 is adjacent to IRF6, a gene known to be causal for Van der Woude syndrom (a Mendelian malformation syndrome). Finding only one *de novo* deletion is not too surprising though, as these events are rare, and the MDTS sensitivity is high for deletions larger than 1 kb. However, in contrast to single nucleotide variants (Jónsson *et al.*, 2017), exact *de novo* mutation rates for copy number variants have not been reported widely. Acuna-Hidalgo *et al.* (2016) estimate one event in 50–100 meiosis for large *de novo* CNVs (in excess of 100 kb), but do not give estimates for smaller CNVs citing technical limitations in detecting such events with current short-read sequencing technology. MDTS also returned a second candidate *de novo* region, spanning a 1.6 kb segment on chromosome 8. This call was supported by a roughly 50% observed decrease in read depth in this region, in contrast to the region on chromosome 1 however, no improperly paired reads spanning this segment were observed. As no split reads were observed either, an equally confident call whether or not this region harbored a true *de novo* deletion in the proband was not possible. In contrast, one rare inherited deletion identified by MDTS was strongly supported by the observed read depths and improperly paired reads, in addition to two copy number polymorphic regions. It is noteworthy that these two *de novo* deletions as well as the rare inherited deletion identified by MDTS (Table 1) were adjacent to known CNPs on chromosomes 1 and 8, respectively (Supplementary Table S5).

Both CANOES and CANOES:b also identified the true *de novo* deletion in the proband of family DS10826, but did not identify the inherited deletion in family DS11025, and did not report the questionable *de novo* deletion in family DS12329. TrioCNV on the other hand did not perform well due to its design for WGS (i.e. non-capture) data. In our simulation study, CANOES had almost identical sensitivity to MDTS for *de novo* deletions 1 kb or larger, which was pushed even higher when using the MDTS bins based on read depth in that algorithm (CANOES:b). In conjunction with a much smaller false positive rate observed (and thus much higher PPV), CANOES:b generally outperformed CANOES in detecting *de novo* deletions (a small caveat however is that CANOES:b was more likely to classify inherited deletions as *de novo*). The reduced number of 'hits' from CANOES using our bins compared to the default bins is likely due to our bins avoiding areas where baits were designed, but actual capture was poor. The size of MDTS bins is controlled by selecting the median number of reads while delineating bins across the initial subsample. Currently, the median is chosen to be 160 reads, yielding a median bin size of 220 bp. Thus, if detection of smaller *de novo* deletions was a priority, smaller bins could be chosen by decreasing the set median (which would come at the expense of specificity, naturally).

Scalability of an algorithm is always a concern when working with genomic sequencing data. Even for TS data, CPU demand can be excessive when many samples (or here, many trios) are jointly analyzed. MDTS exhibited much better scalability than CANOES. The oral cleft data analysis was not computationally feasible with the original CANOES code, but we were able to substantially speed up that algorithm by moving a variance-covariance estimation step outside the loop over all trios. Despite running an order of magnitude faster with this tweak, CANOES was still more than an order of magnitude slower than MDTS, and about two orders of

magnitude slower than MDTS run multi-threaded. In our opinion it is likely that CANOES was simply not designed with the scale of our oral cleft dataset in mind.

Funding

This work was supported by the National Institutes of Health (NIH) grant R03 DE-02579 to IR and RBS. Data collection was supported by NIH grants R01-DE016148 to MLM and R37-DE008559 to JCM. Sequencing of the oral cleft trios was supported by NIH grant U01 HG-005925 to JCM.

Conflict of Interest: none declared.

References

Acuna-Hidalgo,R. *et al*. (2016) New insights into the generation and role of de novo mutations in health and disease. *Genome Biol*., **17**, 241.

Alkan,C. *et al*. (2011) Genome structural variation discovery and genotyping. *Nat. Rev. Genet*., **12**, 363–376.

Backenroth,D. *et al*. (2014) CANOES: detecting rare copy number variants from whole exome sequencing data. *Nucleic Acids Res*., **42**, e97.

Bansal,V. *et al*. (2014) Outlier-based identification of copy number variations using targeted resequencing in a small cohort of patients with tetralogy of fallot. *PLoS One*, **9**, e85375.

Bellos,E. and Coin,L.J.M. (2014) cnvOffSeq: detecting intergenic copy number variation using off-target exome sequencing data. *Bioinformatics (Oxford, England)*, **30**, i639–i645.

Bellos,E. *et al*. (2014) cnvCapSeq: detecting copy number variation in long-range targeted resequencing data. *Nucleic Acids Res*., **42**, e158.

Fromer,M. and Purcell,S.M. (2014) Using XHMM software to detect copy number variation in whole-exome sequencing data. *Curr. Protoc. Hum. Genet*., **81**, 7.23.1–7.2321.

Georgieva,L. *et al*. (2014) De novo CNVs in bipolar affective disorder and schizophrenia. *Hum. Mol. Genet*., **23**, 6677–6683.

Gilissen,C. *et al*. (2014) Genome sequencing identifies major causes of severe intellectual disability. *Nature*, **511**, 344–347.

Glessner,J.T. *et al*. (2014) Increased frequency of de novo copy number variants in congenital heart disease by integrative analysis of single nucleotide polymorphism array and exome sequence data. *Circ. Res*., **115**, 884–896.

Jónsson,H. *et al*. (2017) Parental influence on human germline de novo mutations in 1,548 trios from iceland. *Nature*, **549**, 7673519–7673522.

Krumm,N. *et al*. (2012) Copy number variation detection and genotyping from exome sequence data. *Genome Res*., **22**, 1525–1532.

Kuilman,T. *et al*. (2015) CopywriteR: DNA copy number detection from off-target sequence data. *Genome Biol*., **16**, 49.

Layer,R.M. *et al*. (2014) LUMPY: a probabilistic framework for structural variant discovery. *Genome Biol*., **15**, R84.

Leslie,E.J. *et al*. (2015) Identification of functional variants for cleft lip with or without cleft palate in or near PAX7, FGFR2, and NOG by targeted sequencing of GWAS loci. *Am. J. Hum. Genet*., **96**, 397–411.

Li,H. and Durbin,R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754–1760.

Li,J. *et al*. (2012) CONTRA: copy number analysis for targeted resequencing. *Bioinformatics (Oxford, England)*, **28**, 1307–1313.

Liu,Y. *et al*. (2016) Joint detection of copy number variations in parent-offspring trios. *Bioinformatics (Oxford, England)*, **32**, 1130–1137.

Metzker,M.L. (2010) Sequencing technologies – the next generation. *Nat. Rev. Genet*., **11**, 31–46.

Meyerson,M. *et al*. (2010) Advances in understanding cancer genomes through second-generation sequencing. *Nat. Rev. Genet*., **11**, 685–696.

Nord,A.S. *et al*. (2011) Accurate and exact CNV identification from targeted high-throughput sequence data. *BMC Genomics*, **12**, 184.

Olshen,A.B. *et al*. (2004) Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics*, **5**, 557–572.

Packer,J.S. *et al*. (2016) CLAMMS: a scalable algorithm for calling common and rare copy number variants from exome sequencing data. *Bioinformatics*, **32**, 133–135.

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

Salahshourifar,I. *et al*. (2012) Mutation screening of IRF6 among families with non-syndromic oral clefts and identification of two novel variants: review of the literature. *Eur. J. Med. Genet*., **55**, 389–393.

Scharpf,R.B. *et al*. (2012) Fast detection of de novo copy number variants from SNP arrays for case-parent trios. *BMC Bioinformatics*, **13**, 330.

Shendure,J. and Ji,H. (2008) Next-generation DNA sequencing. *Nat. Biotechnol*., **26**, 1135–1145.

Talevich,E. *et al*. (2016) CNVkit: genome-wide copy number detection and visualization from targeted DNA sequencing. *PLoS Comput. Biol*., **12**, e1004873.

Tan,E.C. *et al*. (2013) De novo 2.3 Mb microdeletion of 1q32.2 involving the Van der Woude Syndrome locus. *Mol. Cytogenet*., **6**, 31.

Tattini,L. *et al*. (2015) Detection of genomic structural variants from next-generation sequencing data. *Front. Bioeng. Biotechnol*., **3**, 92.

Ting,J.C. *et al*. (2006) Analysis and visualization of chromosomal abnormalities in SNP data with SNPscan. *BMC Bioinformatics*, **7**, 25.

Ting,J.C. *et al*. (2007) Visualization of uniparental inheritance, mendelian inconsistencies, deletions, and parent of origin effects in single nucleotide polymorphism trio data with snptrio. *Hum. Mutat*., **28**, 1225–1235.

van Bon,B.W.M. *et al*. (2016) Disruptive de novo mutations of DYRK1A lead to a syndromic form of autism and ID. *Mol. Psychiatry*, **21**, 126–132.

Veltman,J.A. and Brunner,H.G. (2012) De novo mutations in human genetic disease. *Nat. Rev. Genet*., **13**, 565–575.

Venkatraman,E.S. and Olshen,A.B. (2007) A faster circular binary segmentation algorithm for the analysis of array CGH data. *Bioinformatics (Oxford, England)*, **23**, 657–663.

Walsh,T. *et al*. (2008) Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia. *Science (New York, N.Y.)*, **320**, 539–543.

Wang,K. *et al*. (2008) Modeling genetic inheritance of copy number variations. *Nucleic Acids Res*., **36**, e138.

Wellcome Trust Case Control Consortium (2010) Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. *Nature*, **464**, 713–720.

Younkin,S.G. *et al*. (2014) A genome-wide study of de novo deletions identifies a candidate locus for non-syndromic isolated cleft lip/palate risk. *BMC Genetics*, **15**, 24.

Zarrei,M. *et al*. (2015) A copy number variation map of the human genome. *Nat. Rev. Genet*., **16**, 172–183.

Zhang,F. and Lupski,J.R. (2015) Non-coding genetic variants in human disease. *Hum. Mol. Genet*., **24**, R102–R110.

Zhao,M. *et al*. (2013) Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives. *BMC Bioinformatics*, **14**, S1.