Detecting rare copy number variants from Illumina genotyping arrays with the CamCNV pipeline: Segmentation of $z$-scores improves detection and reliability

Joe Dennis1 | Logan Walker2 | Jonathan Tyrer3

Kyriaki Michailidou1,4,5 | Douglas F. Easton1

1Department of Public Health and Primary Care, Centre for Cancer Genetic Epidemiology, University of Cambridge, Cambridge, UK
2Department of Pathology and Biomedical Science, University of Otago, Christchurch, New Zealand
3Department of Oncology, Centre for Cancer Genetic Epidemiology, University of Cambridge, Cambridge, UK
4Biostatistics Unit, The Cyprus Institute of Neurology and Genetics, Nicosia, Cyprus
5Cyprus School of Molecular Medicine, Nicosia, Cyprus

Correspondence
Joe Dennis, Department of Public Health and Primary Care, Centre for Cancer Genetic Epidemiology, University of Cambridge, Cambridge CB1 8RN, UK. Email: jgd29@cam.ac.uk

Funding information
Cancer Research UK, Grant/Award Numbers: C1287/A10118, C1287/A16563; National Institutes of Health, Grant/Award Numbers: U19 CA148065, X01HG007492; Canadian Institutes of Health Research, Grant/Award Number: GPH-129344

Abstract
The intensities from genotyping array data can be used to detect copy number variants (CNVs) but a high level of noise in the data and overlap between different copy-number intensity distributions produces unreliable calls, particularly when only a few probes are covered by the CNV. We present a novel pipeline (CamCNV) with a series of steps to reduce noise and detect more reliably CNVs covering as few as three probes. The pipeline aims to detect rare CNVs (below 1% frequency) for association tests in large cohorts. The method uses the information from all samples to convert intensities to $z$-scores, thus adjusting for variance between probes. We tested the sensitivity of our pipeline by looking for known CNVs from the 1000 Genomes Project in our genotyping of 1000 Genomes samples. We also compared the CNV calls for 1661 pairs of genotyped replicate samples. At the chosen mean $z$-score cut-off, sensitivity to detect the 1000 Genomes CNVs was approximately 85% for deletions and 65% for duplications. From the replicates, we estimate the false discovery rate is controlled at $\sim$10% for deletions (falling to below 3% with more than five probes) and $\sim$28% for duplications. The pipeline demonstrates improved sensitivity when compared to calling with PennCNV, particularly for short deletions covering only a few probes. For each called CNV, the mean $z$-score is a useful metric for controlling the false discovery rate.

KEYWORDS
CNVs, genotyping arrays
1 INTRODUCTION

Genotyping experiments with large case-control studies have been effective at identifying single-nucleotide polymorphisms (SNPs) associated with the disease. Associations with common copy number variants (CNVs) can also be detected by the imputation of their genotypes using reference datasets with sequence data. The 1000 Genomes Project reported that 73% of structural variants with greater than 1% frequency are in strong linkage disequilibrium ($r^2 > 0.6$) with nearby SNPs, although the proportion falls to 44% for common duplications (Sudmant et al., 2015). For example, at two genome-wide association study loci associated with breast cancer, the APOBEC locus (Michailidou et al., 2017; Nik-Zainal et al., 2014) and 2q35 (Wyszynski et al., 2016), imputed deletions tagged by SNPs are associated with risk, and the CNV has been identified as the probable causal variant in each case. However, with current imputation reference panel sizes, it is not possible to reliably impute rare CNVs. With increasing sample sizes genotyped on arrays, there is an opportunity to detect novel associations with rare CNVs that potentially have large effects on disease risk.

Disease associations for rare CNVs detected using genotyping arrays have typically involved large CNVs, often covering multiple genes, for example, those associated with developmental disorders (Jacquemont et al., 2011) and schizophrenia (Rees et al., 2014). Evidence from clinical tests using gene panels suggests that intragenic CNVs account for a significant proportion of pathogenic variants, between 4.7% and 35% depending on the clinical area (Truty et al., 2018). There is also evidence that intronic CNVs can affect the regulation of genes where the exons are highly conserved across evolution (Rigau et al., 2019). Such CNVs are likely to cover only a small number of probes on genotyping arrays.

Several methods have been developed for calling CNVs from the intensities of genotyping arrays. The most commonly used method is PennCNV (Wang et al., 2007), which implements a Hidden Markov Model to find increases or decreases in the intensities of probes, ordered by genomic position, for a single sample. Other methods such as DNAcopy (Venkatraman & Olshen, 2007) use circular binary segmentation (CBS) to detect these shifts in intensity by sample. Methods such as CNVTools (Barnes et al., 2008) and cnvHap (Coin et al., 2010) model the intensities from all samples at CNV loci but are designed to analyse more common CNVs.

The pipeline (CamCNV; Figure 1) we describe here uses DNAcopy but also aims to incorporate the information from all samples. It works on a similar principle to the MeZOD method (Kirov et al., 2011; McCarthy et al., 2009) to detect rare outliers in the distribution of intensities. The method excludes probes in common CNV regions and assumes that at each probe tested the vast majority of samples will have two copies of DNA and their intensities will form a normal distribution. Single copy deletions or three-copy duplications will form part of separate distributions that overlap with the two-copy distribution. Converting the intensity measurements to z-scores and segmenting using DNAcopy detects runs of probes that are potential CNVs for each sample. Our method draws on previous work (Cooper et al., 2015) to remove noise and batch effects from the intensities with a principal component adjustment (PCA). We provide estimates for the sensitivity using known CNVs in 1000 Genomes samples and for the false discovery rate (FDR) from a large number of replicates.

2 METHODS/PIPELINE

2.1 Principal component adjustment

We first export the log R ratio (LRR) intensities for each sample at each probe from Illumina’s Genome Studio.
software into a matrix ordered by probe position. To reduce batch effects in the intensity measurements, we follow the advice of Cooper et al. (2015) and perform a PCA on the LRR. This should also adjust for genomic waves (Dískin et al., 2008). We exclude from the matrix SNPs that fail genotyping quality control (QC) and use the “thin” function from the R bigpca package (Cooper) to select 15% of SNPs most representative of the principal components. We use the function “quick.elbow” to decide how many principal components lie above the “elbow” of the PCA plot and adjust the entire LRR matrix for all SNPs for those PCs using the bigpca PC.correct function. Figures S1–S4 show an example principal component scree plot and the variance of the LRR before and after PCA.

2.2 | z-Score calculation

Using all samples we calculate the means and standard deviations for the LRR at each probe and hence convert the LRRs to z-scores as:

\[
\frac{(SampleLRR - MeanLRRforprobe)}{StandardDeviationofLRRforprobe}
\]

2.3 | Probe and loci exclusions

Before calling CNVs we exclude probes likely to produce unreliable calls. To exclude probes likely to represent technical failures at the chip genotyping level we exclude those not clustered by the Illumina Gentrain algorithm (Gentrain score < 0.15) and low-intensity probes (Illumina metric AB R Mean < 0.2). As our pipeline is designed to detect rare CNVs, we exclude probes within CNVs reported by the 1000 Genomes Project as having greater than 1% frequency. We observed some false-positive deletion calls at probes within a few base pairs of another probe, presumably because competition for DNA fragments lowers the intensity. To guard against this, we exclude probes within three base pairs of another probe. Finally, we exclude noisier probes with high LRR variance (choosing a threshold of 0.12 – two SDs above the mean variance for all probes). This should exclude probes that fail genotyping in a large enough subset of samples to skew the z-scores for all samples. In total, we exclude ~10% of probes (Table S1). After CNV calling, we exclude CNVs falling within 1 Mb of the telomeres and centromeres as defined in the “gap” database table from the UCSC Genome Browser. We also exclude CNVs within immune-related loci (major histocompatibility complex, T-cell receptor, and immunoglobulin heavy chain), as CNVs at these loci may occur somatically in blood.

2.4 | CBS of z-scores

The z-scores are saved into an R data frame for each chromosome ordered by genomic position with the excluded probes removed. We use the R package DNACopy (Olshen et al., 2004; Venkatraman & Olshen, 2007) to segment the z-scores for each sample individually by chromosome using CBS. After using the “smooth” function to detect outliers and smooth the data, we call the segments using the segment function. The R commands used for sample s in z-score data frame z_df are:

```r
CNA. object <- CNA(cbind(z_df$s), z_df$schr, z_df $positions, data. type="logratio", sampleid=colnames (z_df$s), presorted = TRUE).
smoothed. CNA. object <- smooth. CNA(CNA. object).
these_segments <- segment(smoothed. CNA. object, undo. splits="sdundo", undo. SD=2, verbose=0).
```

This produces a list of segments for each sample for each chromosome with the start and end positions, the number of probes, and the segmentation mean for each segment. Before deciding the final cut-off values we label as potential CNVs segments with a mean z-score below −2 or above +1.5, covering at least three probes. Rare larger CNVs may bias our comparisons so we exclude segments above an arbitrary upper length limit of 3 Mb or 200 probes.

2.5 | LRR shift scores per sample

As an additional check, we calculate the shift in the original LRR (before the PCA) between the baseline for the sample and the mean LRR across the called CNV. For each sample, we calculate a mean LRR figure for each chromosome. To ensure that this measure is based on normal two-copy intensities, we only use common SNPs with a high call rate (>99%) where the sample appears heterozygous (B allele frequency [BAF] around 0.5). Then we calculate the difference between this mean sample LRR and the mean of the LRR observed for that sample cross the probes in the called CNV. From the CNVs that match well-imputed CNVs we observe that greater than 99% of the sample shift scores lie within the range of −0.2 to −0.8 for deletions and 0.1–0.4 for duplications. Thus we exclude any called CNVs outside these ranges.
2.6 | BAF score

The BAF can be used alongside the LRR to identify CNVs. For samples with normal two-copy DNA at a probe, this score should be near zero for homozygotes for the reference allele (AA), 0.5 for heterozygotes (AB), and one for homozygotes for the alternate allele (BB). For deletions, the BAF has some value as a negative predictor; any values not close to zero (A−) or one (B−) suggest that it is not a true deletion. Thus, for each potential deletion (as an exclusion score) we calculate the percentage of probes with BAF greater than 0.3 and less than 0.7. We exclude deletions in which ≥25% of probes fail this check.

For three copy duplications, scores should be around zero (AAA), one (BBB), 0.33 (AAB) or 0.66 (ABB), so we can record a positive predictor score as the percentage of probes with BAF greater than 0.1 and less than 0.4 or greater than 0.6 and less than 0.9. However, for short CNVs the value of the BAF is limited as we do not necessarily expect the CNV to cover probes that might be polymorphic, particularly on arrays targeting rare variants.

3 | TEST DATA

3.1 | Breast Cancer Association Consortium samples

To develop the pipeline we used data from 57,796 breast cancer cases and 48,209 controls of European ancestry from 54 studies (Table S2) in the Breast Cancer Association Consortium (BCAC) dataset. These were genotyped on the Illumina OncoArray chip (533,631 probes) at five genotyping centres (Amos et al., 2017). In addition, we included 1661 pairs of duplicate samples. Although the majority of duplicates were genotyped intentionally by studies for QC, 160 of the pairs were submitted by different studies and later identified as the same person from their genotypes and date of birth. The PCA described above was performed separately on each study. For eight studies the number of samples was too large to adjust as a single batch and these studies were divided into two or three batches.

3.2 | 1000 Genomes Project controls

As part of the OncoArray genotyping QC, samples from the 1000 Genomes Project were added to each plate at the genotyping centres. In addition, 240 samples were genotyped by Illumina. After sample QC we used OncoArray data for 233 individuals comprising 2226 replicates. A total of 169 individuals were genotyped once and 64 were genotyped more than once (between 2 and 349 times). We downloaded the structural variant calls from Phase 3 of the 1000 Genomes Project (Sudmant et al., 2015), extracted deletions and duplications for the 233 comparison individuals and calculated the start and end probes for any variant covered by OncoArray probes. There are some complex variants (e.g., one individual has a series of 60 deletions at the end of chromosome 18). To simplify the tests we excluded sets of CNVs where a sample had five or more CNVs starting in the same megabase and also used an arbitrary cut-off of 50 probes to exclude very long variants.

3.3 | Sample QC

For the BCAC dataset we included only DNA samples extracted from whole blood to minimise variation in sample quality. 1000 Genomes DNA samples were from lymphoblastic cell lines. Samples with genotype call rates below 99% or excess heterozygosity or sex chromosome anomalies were excluded (Michailidou et al., 2017). We calculate a derivative log ratio spread (DLRS) figure for each sample as the SD of the differences between LRR for probes ordered by genomic position, divided by the square root of two. This measures the variance from each probe to the next averaged over the whole genome and so it is insensitive to fluctuations in intensities between different chromosomes for individual samples (Cooper et al., 2015). We excluded samples with DLRS 3.5 SDs above the DLRS mean for the study. This exclusion was applied both before and after the PCA. We used the remaining samples to generate the z-scores and call CNVs.

After CNV calling, to exclude samples that may generate multiple false calls with our method, we counted the number of short segments (between 3 and 200 probes) for each sample. We observed that the distribution of segment counts is skewed to the right with many noisy samples with an excessive number of segments (Table S3). We calculated an exclusion cut-off for the maximum number of segments using the following formula where x is the segment count for each sample (based on the rationale that the true number of segments should be approximately Poisson):

\[ y = 2 \times \sqrt{x} \]
\[ \text{cut-off} = \text{median}(y) + 3.5 \]

3.4 | Imputed CNVs

BCAC genotypes were phased using ShapeIT (Delaneau et al., 2013) and imputed to the 1000 Genomes Phase 3
reference panel using Impute2 (Howie et al., 2012) as previously described (Michailidou et al., 2017). We compared called CNVs for samples with imputed dosages of one, that is, those imputed to have one or three copies at the CNV locus.

3.5 | PennCNV calling

As a comparison, we also called all samples using PennCNV with the LRR adjusted for principal components as the input. PennCNV was run with the following options:

detect_cnv.pl -test -hmm lib/hhall.hmm -pfb/pfb_files/pfb_onco.txt -conf -gcmold onco.gcmodel

GC model files were calculated for the OncoArray giving the percentage of G or C bases for 500 Kb either side of each probe. Population-specific (PFB) files were calculated to give the frequency of the B allele for each SNP. After calling, the PennCNV script clean_cnv.pl was used to merge neighbouring CNVs where the gap between them was less than 20% of the total length of the combined CNVs. The PennCNV script filter_cnv.pl was used to generate QC statistics for each sample. Samples with LRR SD greater than 0.2 or an excessive number (greater than 3.5 SD) of CNVs were excluded.

3.6 | Comparison of CNV calls

In the comparisons below, CNV calls were counted as an exact match if they covered the maximum number of probes on OncoArray for the reference CNV and were no longer. We categorised calls as partial matches if at least 50% of the expected probes were covered and/or where the called CNV extended further, excluding those covering greater than five times the number of probes in the reference CNV.

3.7 | Precision of the z-score cut-offs

z-Score cut-offs are determined based on the mean and SD of values in the sample. Thus the value corresponding to a z-score cut-off of $K$ is given by

$$y = \hat{\mu} + K\hat{\sigma},$$

where $\hat{\mu}$ and $\hat{\sigma}$ are the estimated mean and variance. In a finite sample size:

$$\text{var}(\hat{y}) \approx \text{var}(\hat{\mu}) + K^2\text{var}(\hat{\sigma}),$$

where $N$ is the sample size. Thus on the normalised scale, the variance of the estimated cut-off is approximately

$$\text{var}(\hat{K}) = \frac{1}{N} \left(1 + \frac{K^2}{2}\right).$$

4 | RESULTS

4.1 | Distribution of mean z-scores for imputed CNVs

To evaluate the distribution of mean z-scores for segments matching imputed CNVs, we compared the segments called by the pipeline (Figure 1) that matched a set of well-imputed ($r^2 > 0.8$) rare CNVs in the BCAC data. We identified 26 distinct deletions and 11 duplications, carried by 7189 and 4874 samples, respectively. We plotted histograms of the mean z-scores of the segments that perfectly matched the imputed CNVs ($n = 4881$ deletions and $n = 1051$ duplications) as shown in Figure 2.

4.2 | z-Score cut-offs and estimated FDR

To select mean z-score cut-offs to control the FDR, we compared called CNVs in the 1661 pairs of BCAC replicates. We assume a CNV seen in both replicates is likely to be genuine, whereas a CNV seen in only one sample is likely to be an artefact. We observed that the mean z-scores of the unmatched calls from the first of the replicate pairs (Figure 3) are concentrated towards the normal end of the distribution. This suggests that they are mostly false-positive calls at loci with two-copy DNA, although some may be genuine CNVs at probes with a large degree of overlap with the normal distribution (hence harder to call). The relationship between the mean z-score of the segment and the proportion that is unmatched is shown in Figure 3. To provide reasonable control of the FDR, although not impacting the sensitivity too severely, we chose cut-offs at $-3.7$ for deletions and $+2$ for duplications; this results in an overall FDR of $\sim 10\%$ for the remaining deletions and $\sim 30\%$ for duplications (Figure 4). The FDR does not reduce materially for duplications with more stringent cut-offs. At the other end of the distribution, we observed a small number of extreme values for the mean z-score. These may reflect homozygous deletions or duplications with more than
three copies but could also be due to probes failing to bind at all. We, therefore, excluded calls with $z < -14$ for deletions and $z > +10$ for duplications. Although this could exclude some true homozygotes, the main focus is on calling rare CNVs and hence this possibility can be ignored here.

Using these cut-offs, the results from the matching exercise are shown in Table 1. Based on exact or partial matches, the overall FDR was 10% for deletions and 28% for duplications. The FDR decreases with increasing numbers of probes as shown in Table 2, particularly for deletions. We observed no significant difference between the unmatched rate for samples from the same study and those collected by different studies ($n = 160$ pairs) where the DNA has been extracted by different laboratories at different ages and one might expect more variability from batch effects.

4.3 | Concordance of imputed and called CNVs

The overall match rate for imputed data using our chosen cut-offs is summarised in Table 3. The match rates for these CNVs will be affected by the uneven distribution of probe coverage (Table 4). The majority of these deletions only have three probes while most of the duplications have more than six.

4.4 | Summary of called CNVs

Using the above cut-offs we called a mean of 2.9 deletions ($SD = 1.6$) and 2.5 duplications ($SD = 2.0$) per sample in the BCAC dataset. The sample QC checks excluded 14% of the original samples (Table S4), with the proportion ranging from 2% to 46% for individual studies. Using the interview date as a proxy to estimate how long DNA has been stored before genotyping, we observed that a longer storage time is a strong predictor ($p < 2e-16$ adjusting for study) that a sample will fail QC (Table S5). The proportion of CNVs excluded by the LRR shift and BAF score checks is shown in Table S6. These are likely to exclude a small proportion of genuine CNVs but they appear to be effective at targeting unreliable calls from the higher exclusion rate for unmatched CNVs from the replicate comparisons and among samples with an excessive number of segments.
FIGURE 3 Mean z-scores of matching and mismatching deletions (a) and duplications (b) in replicate samples. CNV, copy number variant.

FIGURE 4 Estimated false discovery rate (FDR) by mean z-score of deletions (a) and duplications (b) in replicate samples.
After excluding long CNVs covering more than 50 probes, the mean length was 7.3 probes (SD = 7.1)/34 kb (SD = 71 kb) for deletions, and 10.0 probes (SD = 9.6)/70 kb (SD = 109 kb) for duplications. A total of 14% of the probes on the chip were covered by five or more deletions and 22% of probes by more than five duplications. The probes with most CNV calls overlap with previously published CNV regions. For example, 58% of probes in deletions and 39% of probes in duplications with a frequency greater than 0.5% overlap entries in the Database of Genomic Variants (Zarrei et al., 2015; Table S7).

### Table 1: Match rate for CNV calls from 1661 pairs of replicate samples in BCAC data

| Category     | Deletions       | Duplications  |
|--------------|-----------------|---------------|
| Perfect match| 3650 (74%)      | 1359 (34%)    |
| Partial match| 789 (16%)       | 1533 (38%)    |
| Unmatched    | 495 (10%)       | 1135 (28%)    |
| Total        | 4934            | 4027          |

Abbreviations: BCAC, Breast Cancer Association Consortium; CNV, copy number variant.

### Table 2: False discovery rate (FDR) by probe length from the comparison of replicate sample calls

| Probes | Deletions | Duplications |
|--------|-----------|--------------|
|        | Count     | FDR          | Count     | FDR          |
| 3      | 1541      | 23.40%       | 489       | 62.40%       |
| 4      | 852       | 8.90%        | 520       | 43.10%       |
| 5      | 479       | 5.80%        | 382       | 43.70%       |
| 6      | 288       | 2.40%        | 282       | 36.50%       |
| 7      | 231       | 2.20%        | 259       | 40.50%       |
| 8      | 172       | 1.70%        | 223       | 19.30%       |
| 9      | 170       | 0.60%        | 193       | 23.30%       |
| 10+    | 1201      | 1.20%        | 1679      | 8.50%        |

### Table 3: Match rate for called CNVs against BCAC imputed CNVs

|                | Separate CNVs | Imputed calls to compare | Matched rate | Perfect matches | Matches excluded by QC |
|----------------|---------------|--------------------------|--------------|-----------------|------------------------|
| Deletions      | 26            | 7189                     | 77.90%       | 65.20%          | 4.80%                  |
| Duplications   | 11            | 4874                     | 61.10%       | 20.70%          | 5.90%                  |

Abbreviations: BCAC, Breast Cancer Association Consortium; CNV, copy number variant; QC, quality control.

### 4.5 Detection of known CNVs in 1000 Genomes samples

We compared our pipeline calls for 1000 Genomes samples genotyped on the OncoArray with the CNVs published by the 1000 Genomes Project. We identified 523 rare (less than 1% frequency) deletions and 308 rare duplications that cover at least three probes on OncoArray. Because some CNVs were carried by more than one of our comparison samples and some samples were replicated multiple times in the OncoArray dataset (n = 3532 comparison CNV calls in total), we chose a random replicate to perform the comparisons. The results are shown in Table 5 and the detected CNVs in Table S8. A total of 80% of deletions and 60% of duplications were detected with either a perfect or partial match. CNVs that were not detected at all tended to be short (60% cover only three or four probes) and tended to have a smaller proportion covered by the chip probes. Those detected with 100% precision have a mean 65% (SD = 22%) of base pairs covered compared to a mean of 53% (SD = 25%) for CNVs with no matches. Around 6% of CNVs were detected but were excluded as they were outside the mean z-score cut-offs or failed other QC metrics. Where CNVs were carried by more than three replicates, deletions called in one sample tended to be called in the majority of samples, with a poorer replication rate for duplications. A total of 70% of these detected deletions (n = 63) and 40% of duplications (n = 35) had a greater than 90% replication rate (Figure S5).

### 4.6 Comparison with PennCNV calling

We also called the 1000 Genomes samples and BCAC replicates using PennCNV. For the 1000 Genomes comparison, we categorised the calling of the 525 deletions and 308 duplications as either “perfect,” if the CNV call matched exactly with the known genotype in more than 90% of replicates, and “imperfect” if at least some calls partially matched or the replicate was below 90%. Figure 5 and Table S9 show the comparison between
PennCNV and CamCNV calling. Overall CamCNV called 20% more deletions and 10% more duplications than PennCNV. The set of CNVs detected by CamCNV but not by PennCNV (Figure S6) tend to cover fewer probes: the deletions cover a mean of 3.9 probes (SD = 1.5) although those matched by both methods cover a mean of 8.5 probes (SD = 7.4). For duplications, the figures are 9.1 probes (SD = 5.5) and 14.2 probes (SD = 10.7).

We also estimated the FDR for PennCNV using the BCAC replicate samples. The CamCNV FDR was lower than PennCNV for deletions (Table 6) but higher for short duplications. In the replicate samples, CamCNV called around 1.5 times more deletions and ~1.4 times more duplications than PennCNV, with nearly all the additional deletions having fewer than six probes.

## DISCUSSION

There are a small number of analyses of CNVs called from large genotyping datasets. These have highlighted the importance of detecting CNVs from a small number of probes (Szatkiewicz et al., 2013) and removing artefactual calls (Cooper et al., 2015). Assessing common CNVs with multiple overlapping intensity distributions at each probe requires robust statistical methods (Barnes et al., 2008). The calling problem is simplified for rare CNVs as nearly all samples will belong to the two-copy distribution with a small number of outliers that are potentially one-copy deletions or three-copy duplications. The two-copy intensity distributions vary between probes but the PCA and conversion of raw intensities to z-scores adjusts for this variance. Thus our method incorporates information from all the samples at each probe; this may explain the improvement in sensitivity compared to the PennCNV method, which looks at each sample individually. From known CNVs and replicate samples we can derive the distribution of z-scores for one-copy deletions and three-copy duplications and calibrate the mean z-score cut-offs for segments to provide an acceptable FDR. From the replicate comparison, we observe that calls with more extreme mean z-scores are unlikely to be false positives, so the mean z-score provides an additional metric for the confidence of the calling alongside the number of probes.

A limitation is that the calling for duplications is less reliable than for deletions, as the duplication intensity distribution has a greater degree of overlap with normal copy samples. Also, there remains the possibility that technical artefacts will cause false calls to cluster at particular probes. For rare variant association tests a proportion of false CNV calls scattered across an array is not as detrimental as a cluster of false calls at particular probes, so chip-specific probe QC is an important element of this pipeline. Strict sample QC is also required. The proportion of samples with unreliable calls may vary greatly by study and this method may not work with samples not genotyped from the blood. We have tested the method on another Illumina array (iCOGs), and the mean z-score cut-offs used for OncoArray appeared appropriate for that data as well. Although we use a large number of samples, smaller sample sizes are likely to produce z-score cut-offs that are reasonably stable as they depend only on the precision of the mean and variance. For example, with a sample size of ~800 and a deletion cut-off of −3.7, the SE of the cut-off would be around ±0.1.

## CONCLUSION

We present a CNV calling pipeline tailored to detect rare CNVs from Illumina genotyping data. For deletions, it demonstrates a high degree of sensitivity while controlling the false call rate and is an improvement on the most commonly used method (PennCNV) particularly in

---

**TABLE 4** Probe coverage for CNVs in comparison with imputed data

| Probes covered | Deletions | Duplications |
|---------------|-----------|--------------|
| 3             | 14        | 1            |
| 4             | 3         | 2            |
| 5             | 3         | 1            |
| 6–10          | 3         | 5            |
| >10           | 3         | 3            |

Abbreviation: CNV, copy number variant.

**TABLE 5** Comparison of called CNVs for 1000 Genomes samples with CNVs from sequencing

| Category                                      | Deletions | Duplications |
|----------------------------------------------|-----------|--------------|
| Perfect match                                | 253 (48.2%) | 43 (14%)     |
| Partial match                                | 168 (32%)  | 140 (45.5%)  |
| Matched but outside mean z-score cut-offs    | 27 (5.1%)  | 11 (3.6%)    |
| Matched but excluded by QC metrics           | 6 (1.1%)   | 3 (1%)       |
| Called CNV covers >5 times as many probes as 1000 Genomes CNV | 6 (1.1%) | 7 (2.3%) |
| Uncalled                                    | 65 (12.4%) | 104 (33.8%)  |

Abbreviations: CNV, copy number variant; QC, quality control.
detecting deletions covering only a few probes. We believe that this method will be valuable in testing for associations with rare CNVs in existing genotyping datasets.

**ACKNOWLEDGMENTS**

Genotyping of the OncoArray was principally funded from three sources: the PERSPECTIVE project, funded from the Government of Canada through Genome Canada and the Canadian Institutes of Health Research (Grant No. GPH-129344), the Ministère de l’Économie, de la Science et de l’Innovation du Québec through Genome Québec and the PSRSIIIRI-701 grant, and the Quebec Breast Cancer Foundation; the NCI Genetic Associations and Mechanisms in Oncology (GAME-ON) initiative and Discovery, Biology and Risk of Inherited Variants in Breast Cancer project (NIH Grants U19 CA148065 and X01HG007492); and Cancer Research UK (C1287/A10118 and C1287/A16563). The authors would like to thank all BCAC investigators for providing access to the BCAC dataset. The BCAC studies are listed in Table S2. The authors would thank all the individuals who took part in these studies and all the researchers, clinicians, technicians, and administrative staff who have enabled this study to be carried out.

**CONFLICT OF INTERESTS**

The authors declare that there are no conflict of interests.
AUTHOR CONTRIBUTIONS
Joe Dennis and Douglas F. Easton designed the project. Joe Dennis, Douglas F. Easton, Logan Walker, and Jonathan Tyrer developed the pipeline. Kyriaki Michailidou performed the imputation. All authors read and approved the final manuscript.

DATA AVAILABILITY STATEMENT
Availability of data and materials: Example code and a test dataset with 1000 Genomes OncoArray genotyping is available at the project home page (https://github.com/jgd29/CamCNV). The BCAC OncoArray dataset supporting the conclusions of this article is available in the dbGap repository, Study ID: phs001265.v1.p1 https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs001265.v1.p1. The 1000 Genomes Phase 3 structural calls are available from http://ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/phase3/integrated_sv_map/ALL.wgs.integrated_sv_map_v2.20130502.svs.genotypes.vcf.gz.

ORCID
Joe Dennis https://orcid.org/0000-0003-4591-1214
Logan Walker https://orcid.org/0000-0003-0018-3719
Kyriaki Michailidou https://orcid.org/0000-0001-7065-1237

REFERENCES
Amos, C. I., Dennis, J., Wang, Z., Byun, J., Schumacher, F. R., Gayther, S. A., & Gruber, S. B. (2017). The OncoArray Consortium: A network for understanding the genetic architecture of common cancers. Cancer Epidemiology, Biomarkers and Prevention, 26, 126–135. https://doi.org/10.1186/1055-9965.epi-16-0106
Barnes, C., Plagnol, V., Fitzgerald, T., Redon, R., Marchini, J., Clayton, D., & Hurles, M. E. (2008). A robust statistical method for case-control association testing with copy number variation. Nature Genetics, 40(10), 1245–1252. https://doi.org/10.1038/ng.206
Cain, L. J. M., Asher, J. E., Walters, R. G., El-Sayed Moustafa, J. S., de Smith, A. J., Sladek, R., & Blakemore, A. I. F. (2010). cnvHap: an integrative population and haplotype-based multipplatform model of SNPs and CNVs. Nature Methods, 7, 541–546. https://doi.org/10.1038/nmeth.1466
Cooper, N. J. BigPCA package. https://cran.r-project.org/web/packages/bigpca/index.html
Cooper, N. J., Shtir, C. J., Smyth, D. J., Guo, H., Swafford, A. D., Zanda, M., & Todd, J. A. (2015). Detection and correction of artefacts in estimation of rare copy number variants and analysis of rare deletions in type 1 diabetes. Human Molecular Genetics, 24(6), 1774–1790. https://doi.org/10.1093/hmg/ddu581
Delaneau, O., Zagury, J.-F., & Marchini, J. (2013). Improved whole-chromosome phasing for disease and population genetic studies. Nature Methods, 10(1), 5–6. https://doi.org/10.1038/nmeth.2307
Diskin, S. J., Li, M., Hou, C., Yang, S., Glessner, J., Hakonarson, H., & Wang, K. (2008). Adjustment of genomic waves in signal intensities from whole-genome SNP genotyping platforms. Nucleic Acids Research, 36(19), e126. https://doi.org/10.1093/nar/gkn556
Howie, B., Fuchsberger, C., Stephens, M., Marchini, J., & Abecasis, G. R. (2012). Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. Nature Genetics, 44(8), 955–959. https://doi.org/10.1038/ng.2354
Jacquemont, S., Reymond, A., Zufferey, F., Harewood, L., Walters, R. G., Kutalik, Z., & Froguel, P. (2011). Mirror extreme BMI phenotypes associated with gene dosage at the chromosome 16p11.2 locus. Nature, 478(7367), 97–102. http://www.nature.com/nature/journal/v478/n7367/abs/nature10406.html#supplementary-information
Kirov, G., Pocklington, A. J., Holmans, P., Ivanov, D., Ikeda, M., Ruderfer, D., & Owen, M. J. (2011). De novo CNV analysis implicates specific abnormalities of postsynaptic signalling complexes in the pathogenesis of schizophrenia. Molecular Psychiatry, 17, 142–153. https://doi.org/10.1038/mp.2011.154. https://www.nature.com/articles/mp2011154#supplementary-information
McCarthy, S. E., Makarov, V., Kirov, G., Addington, A. M., McClellan, J., Yoon, S., & Sebat, J. (2009). Microduplications of 16p11.2 are associated with schizophrenia. Nature Genetics, 41(11), 1223–1227. https://doi.org/10.1038/ng.474
Michailidou, K., Lindström, S., Dennis, J., Beesley, J., Hui, S., Kar, S., & Easton, D. F. (2017). Association analysis identifies 65 new breast cancer risk loci. Nature, 551, 92–94. https://doi.org/10.1038/nature24284
Nik-Zainal, S., Wedge, D. C., Alexandrov, L. B., Petljak, M., Butler, A. P., Boil, N., & Stratton, M. R. (2014). Association of a germline copy number polymorphism of APOBEC3A and APOBEC3B with burden of putative APOBEC-dependent mutations in breast cancer. Nature Genetics, 46(5), 487–491. https://doi.org/10.1038/ng.2955
Olschen, A. B., Venkatraman, E. S., Lucito, R., & Wigler, M. (2004). Circular binary segmentation for the analysis of array-based DNA copy number data. Biostatistics, 5(4), 557–572. https://doi.org/10.1093/biostatistics/ksh008
Rees, E., Walters, J. T., Georgieva, L., Isles, A. R., Chambert, K. D., Richards, A. L., & Kirov, G. (2014). Analysis of copy number variations at 15 schizophrenia-associated loci. British Journal of Psychiatry, 204(2), 108–114. https://doi.org/10.1192/bjp.bp.113.131052
Rigau, M., Juan, D., Valencia, A., & Rico, D. (2019). Intronic CNVs and gene expression variation in human populations. PLoS Genetics, 15(1), e1007902. https://doi.org/10.1371/journal.pgen.1007902
Sudmant, P. H., Rausch, T., Gardiner, E. J., Handsaker, R. E., Abyzov, A., & Huddleston, J. (2015). An integrated map of structural variation in 2,504 human genomes. Nature, 526, 75–81. https://doi.org/10.1038/nature15394
Szatkiewicz, J. P., Neale, B. M., O’Dushlaine, C., Fromer, M., Goldstein, J. I., Moran, J. L., & Sullivan, P. F. (2013). Detecting large copy number variants using exome genotyping arrays in a large Swedish schizophrenia sample. Molecular Psychiatry, 18(11), 1178–1184. https://doi.org/10.1038/mp.2013.98
Truty, R., Paul, J., Kennemer, M., Lincoln, S. E., Olivares, E., Nussbaum, R. L., & Aradhya, S. (2018). Prevalence and properties of intragenic copy-number variation in Mendelian disease genes. *Genetics in Medicine, 21*, 114–123. https://doi.org/10.1038/s41436-018-0033-5

Venkatraman, E. S., & Olshen, A. B. (2007). A faster circular binary segmentation algorithm for the analysis of array CGH data. *Bioinformatics, 23*(6), 657–663. https://doi.org/10.1093/bioinformatics/btl646

Wang, K., Li, M., Hadley, D., Liu, R., Glessner, J., & Grant, S. F. (2007). PennCNV: An integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Research, 17*, 1665–1674. https://doi.org/10.1101/gr.6861907

Wyszynski, A., Hong, C. C., Lam, K., Michailidou, K., Lytle, C., Yao, S., & Cole, M. D. (2016). An intergenic risk locus containing an enhancer deletion in 2q35 modulates breast cancer risk by deregulating IGFBP5 expression. *Human Molecular Genetics, 25*, 3863–3876. https://doi.org/10.1093/hmg/ddw223

Zarrei, M., MacDonald, J. R., Merico, D., & Scherer, S. W. (2015). A copy number variation map of the human genome. *Nature Reviews Genetics, 16*(3), 172–183. https://doi.org/10.1038/nrg3871

**SUPPORTING INFORMATION**

Additional Supporting Information may be found online in the supporting information tab for this article.

---

**How to cite this article:** Dennis J, Walker L, Tyrer J, Michailidou K, Easton DF. Detecting rare copy number variants from Illumina genotyping arrays with the CamCNV pipeline: Segmentation of z-scores improves detection and reliability. *Genetic Epidemiology*. 2021;45:237–248. https://doi.org/10.1002/gepi.22367