Published in final edited form as:

Genet Epidemiol. 2020 March ; 44(2): 148–158. doi:10.1002/gepi.22269.

Critical Evaluation of Copy Number Variant Calling Methods Using DNA Methylation

Varun Kilaru[1], **Anna K Knight**[1], **Seyma Katrinli**[1], **Dawayland Cobb**[1], **Adriana Lori**[2], **Charles F Gillespie**[2], **Adam X Maihofer**[3], **Caroline M Nievergelt**[3,4,5], **Anne L Dunlop**[6,7], **Karen N Conneely**[8], **Alicia K Smith**[1,2]

[1]Department of Gynecology and Obstetrics, Emory University School of Medicine, Atlanta, USA

[2]Department of Psychiatry and Behavioral Sciences, Emory University School of Medicine, Atlanta, USA

[3]Department of Psychiatry, University of California San Diego

[4]Veterans Affairs San Diego Healthcare System

[5]Veterans Affairs Center of Excellence for Stress and Mental Health

[6]Nell Hodgson Woodruff School of Nursing, Emory University School of Medicine, Atlanta, USA was supported in part by the National Institutes of Health

[7]Department of Family and Preventive Medicine, Emory University School of Medicine, Atlanta, Georgia

[8]Department of Human Genetics, Emory University School of Medicine, Atlanta, USA

Abstract

Recent technological and methodological developments have enabled the use of array-based DNA methylation data to call copy number variants (CNVs). ChAMP, Conumee and cnAnalysis450k are popular methods currently used to call CNVs using methylation data. However, so far, no studies have analyzed the reliability of these methods using real samples. Data from a cohort of individuals with genotype and DNA methylation data generated using the HumanMethylation450 and MethylationEPIC BeadChips was used to assess the consistency between the CNV calls generated by methylation and genotype data. We also took advantage of repeated measures of methylation data collected from the same individuals to compare the reliability of CNVs called by ChAMP, Conumee, and cnAnalysis450k for both the methylation arrays. ChAMP identified more CNVs than Conumee and cnAnalysis450k for both the arrays and, as a consequence, had a higher overlap (~62%) with the calls from the genotype data. However, all methods had relatively low reliability. For the MethylationEPIC array, Conumee had the highest reliability (57.6%), whereas for the HumanMethylation450 array, cnAnalysis450k had the highest reliability (43.0%). Overall, the MethylationEPIC array provided significant gains in reliability for CNV calling over the HumanMethylation450 array but not for overlap with CNVs called using genotype data.

Contact: alicia.smith@emory.edu.

Data Availability Statement

Data available on request due to privacy/ethical restrictions

Keywords

CNV; ChAMP; Conumee; cnAnalysis450k; DNA methylation

Introduction

The human genome has extensive copy number variation (CNV), which is defined as a segment of DNA larger than 50 base pairs that differs in copy number from a reference genome (McCarroll et al., 2006; Pinto, Marshall, Feuk, & Scherer, 2007; Redon et al., 2006; Zarrei, MacDonald, Merico, & Scherer, 2015). Such variants range in size from 50 base pairs to >500 kilobases (kb), and most healthy people carry numerous CNVs (Feuk, Carson, & Scherer, 2006; Pinto et al., 2007; Zarrei et al., 2015). Despite the presence of CNVs in healthy individuals, they have also been associated with changes in gene expression, alterations in gene dosage, and a wide range of other outcomes, including the onset of disease (Redon et al., 2006; Zarrei et al., 2015). For example, pathogenic CNVs have been associated with congenital malformations (Di Gregorio et al., 2015), schizophrenia (Cnv, Schizophrenia Working Groups of the Psychiatric Genomics, & Psychosis Endophenotypes International, 2017), and environmental exposures (Du et al., 2017; Martinez et al., 2010), among others (Cuccaro, De Marco, Cittadella, & Cavallaro, 2017; Poniah et al., 2017; Quintela et al., 2017). Thus, the ability to identify and characterize CNVs as pathogenic or benign may provide novel insight into a variety of adverse health outcomes.

The first CNVs were identified through low-resolution karyotypes, which gradually improved with the advent of chromosome banding and were further refined by fluorescence in situ hybridization (FISH). These techniques could identify large CNVs, but microarray based technology was required to further improve resolution of CNVs that could be detected (Feuk et al., 2006). The gold standard of CNV detection quickly became array-based comparative genome hybridization (array-CGH). Array-CGH uses fluorescently tagged reference and test samples to detect regions with copy number gains or losses (Lockwood, Chari, Chi, & Lam, 2006). As the field has advanced, several other methods of CNV detection have emerged, including using signals from genotyping and DNA methylation arrays and whole genome sequencing.

CNV calling algorithms for genotyping generally rely on two measures derived from probe intensities, log R Ratio (LRR) and allele frequency (BAF), to call CNVs. The most commonly used programs for CNV-calling using genotype data are Birdsuite (Malhotra et al., 2011), iPattern (International Schizophrenia, 2008) and PennCNV (Wang et al., 2007). Birdsuite is mainly used for Affymetrix assays. It assigns copy numbers based on the summarized intensity for common pre-identified copy number polymorphisms. A Hidden Markov Model (HMM) based algorithm (Colella et al., 2007) is used to assign rare CNVs based on the probe-specific mean and variance estimated for each SNP locus, which is expected to have only 2 copies. iPattern uses a pattern recognition approach to analyze the probe intensities of a group of samples and then uses a sliding window approach to identify consecutive outliers. For the defined region, samples with probe intensities that are significantly lower or higher than the average probe intensity of all the samples are then identified as gains or losses. PennCNV also employs a HMM to detect CNVs from both Illumina and Affymetrix arrays. The algorithm uses not only the total signal intensity and allele intensity ratio like other algorithms but also incorporates distance between neighboring SNPs and the allele frequency of SNPs to call CNVs. To account for the relative strengths of the various programs, it is common to require CNVs to be called by more than one program and hence, pipelines for calling CNVs using genotype data involve CNV calling using multiple programs.

There are currently three commonly used packages to call CNVs from DNA methylation data: ChAMP (Feber et al., 2014), Conumee (Hovestadt & Zapatka), and cnAnalysis450k (Knoll, Debus, & Abdollahi, 2017). All three methods operate on the premise that the total methylation signal (Unmethylated Signal + Methylated Signal) is directly reflective of the copy number state. ChAMP has the first implementation that works on this premise and calls CNVs by normalizing probes based on GC content followed by circular binary segmentation to define regions with variations in copy number. Conumee identifies CNVs by normalizing experimental samples to reference samples using multiple linear regression, and then taking the log2-ratio of probe intensities of experimental and reference samples, followed by combining probes into bins and then dividing the genome into segments with the same copy numbers (Hovestadt & Zapatka). Finally, cnAnalysis450k identifies segments by screening for changes in variance of intensity between the samples and reference samples.

Genotyping arrays cannot detect translocations and inversions, while conventional DNA methylation arrays are biased towards detecting CNVs within genes compared to intergenic regions (Feber et al., 2014; Feuk et al., 2006). However, there are several advantages to using DNA methylation arrays exclusively to call CNVs. Feber and colleagues describe several of these advantages including reduction in cost and sample consumed compared to running both genotyping and DNA methylation arrays, as well as advantages in detecting and characterizing tumor heterogeneity in cancer research (Feber et al., 2014). A challenge of using DNA methylation arrays to call CNVs is variation in the density of methylation probes across different genomic regions. However, great strides have been made towards more completely evaluating methylation across the genome. The HumanMethylation27 BeadChip did not have sufficient probe density to call CNVs, but the HumanMethylation450 BeadChip (450k array) saw the development of CNV calling algorithms for DNA methylation datasets. The MethylationEPIC BeadChip (EPIC array) arrays provide further density and much more comprehensive coverage of the genome, potentially resulting in improved reliability for CNV calling.

Despite the wide implementation of these methods, there has been little research on the reliability of the algorithms in identifying the CNVs. Analyzing the reliability of these methods has become increasingly crucial with the advent of the EPIC array, which queries many more CpG sites than the 450k array. Hence, this paper evaluates the reliability of the popular CNV calling algorithms by: i) testing the ability of the algorithm to call the same CNVs as the ones called using genotype data within the same individuals and ii) testing the ability of the algorithm to call the same CNVs at multiple time points of a longitudinal dataset and comparing repeated measures within individuals.

Methods

Study cohorts

Grady Trauma Project (GTP) cohort—Samples from the Grady Trauma Project (GTP) were used to assess the ability of each algorithm to call CNVs from a GWAS pipeline. The participants are all adult, primarily African American, female, and have been previously described (Binder et al., 2008; Gillespie et al., 2009; Ressler et al., 2011). Samples were collected from the general medical clinics of Grady Memorial Hospital, Atlanta, GA, USA. Participants who were waiting for appointments in the primary care and obstetrics and gynecology clinics were approached by a member of the research team to conduct screening interviews and collect saliva samples. After completion of initial interviews, study participants were invited to participate in a secondary phase of the study in which blood samples were collected. DNA was extracted from saliva by Oragene collection vials (DNA Genotek, Ottawa, ON, Canada) using the DNAdvance kit (Beckman Coulter Genomics, Danvers, MA, USA) for genotyping. DNA was extracted from blood samples using the E.Z.N.A. Mag-Bind Blood DNA Kit (Omerga Bio-Tek, Nocross, GA, USA) or ArchivePure DNA Blood Kit (5 Prime, Gaithersburg, MD, USA) and assessed for DNA methylation. This study was approved by the Emory Institutional Review Board.

Emory University African American Microbiome in Pregnancy (AAMP) cohort —Samples from AAMP were used to assess reliability for each algorithm to call CNVs in multiple samples from the same individual. AAMP is a longitudinal cohort comprised of 142 pregnant African American women, the recruitment and demographics of which have been previously described (Corwin et al., 2017; Knight et al., 2017). Briefly, women were recruited from prenatal care clinics affiliated with two Atlanta metro area hospitals, Emory University Midtown Hospital and Grady Memorial Hospital. Study participants contributed two blood samples each over the course of their pregnancy. The first sample was collected between 6–15 weeks gestation and the second sample was collected at 22–33 weeks gestation. Peripheral blood mononuclear cells (PBMCs) were isolated from whole blood using a Ficoll density gradient and were stored in AllProtect Buffer (Qiagen) at −80 °C until a simultaneous DNA and RNA extraction using the AllPrep RNA/DNA Mini Kit (Qiagen) was performed according to manufacturer’s instructions. DNA quantification and quality was assessed using the Quant-it Pico Green kit (Invitrogen).This study was approved by the Emory Institutional Review Board.

Statistical analyses

All analyses were performed in R version 3.4.1.

CNV calling from genotype data—Genome-wide SNP genotyping was done using the Illumina Omni1-Quad BeadChip, which interrogates 1,011,219 individual SNPs. For each GTP subject, GenomeStudio exported files were used to call CNVs using the iPattern (International Schizophrenia, 2008) and PennCNV (Wang et al., 2007) algorithms following the methodology laid out by Marshall et al (Marshall et al., 2017). The intersection of PennCNV and iPattern calls was taken, such that calls were only retained if they were made by both methods and the predicted copy number state matched. Quality control metrics, including Log R Ratio SD, B Allele Frequency SD, and GC waviness were estimated from PennCNV. Samples were checked to ensure that the total region spanned by CNV or any of the other 3 QC measures was lesser than the overall sample median + 3 standard deviations. Large CNVs appearing to be split were annealed. Samples were checked for possible aneuploidy based on >10% of any chromosome being spanned by CNVs. CNVs spanning centromeres or overlapping telomeres (defined as 100kb from the chromosome end) were excluded, as were CNVs with >50% overlap with known segmental duplication, immunglobulin, or T-cell receptor loci. CNVs spanning <20kb, or containing <10 probes were excluded from further analyses.

DNA methylation data—DNA methylation was assessed on either the Illumina HumanMethylation450 BeadChip or the Illumina MethylationEPIC BeadChip, which assess >450,000 and >850,000 CpG sites across the genome, respectively. Briefly, 1 μg of DNA was processed and hybridized to the BeadChip, according to manufacturer’s instructions (Illumina, San Diego, CA). The raw data was processed using default parameters recommended by the developers so as to allow for a realistic comparison between the methods. All the samples included in the analyses were filtered through the same QC protocol to ensure that they have a reliable signal for more than 95% of the CpG sites. DNA methylation data from the GTP cohort can be accessed through NCBI’s Gene Expression Omnibus: GSE72680 (450k) and GSE132203 (EPIC). DNA methylation data from the AAMP cohort can be accessed through NCBI’s Gene Expression Omnibus: GSE107459 (450k) and GSE122408 (EPIC).

Calling CNVs using methylation data—CNVs were called using the R packages ChAMP and Conumee, with default parameters. cnAnalysis450k allows the user to choose from a range of normalization methods and several workflow options (Workflow A: Raw normalized data followed by Dasen normalization, Workflow B: Raw normalized data followed by Dasen normalization and z-transformation, Workflow C: Raw normalized data is used to identify segments with the Conumee package)(Knoll et al., 2017). We followed the recommendation of the package’s authors in the original manuscript and implemented “Workflow B”, which uses a z-transformation following Illumina normalization(Knoll et al., 2017). All the methods work on the assumption that the total intensity of the reference samples represents the baseline state of the genome and the observed differences between the total intensity of the reference samples and the individual samples represent a copy number variant. Since it is important to ensure that CNV detection is not affected by cell type differences or any other technical artifacts, samples from the same experiment were chosen as the reference samples. To ensure a realistic comparison, the same samples were used as the reference samples to call CNVs for all three methods. CNV calls were performed separately for the two cohorts and the two arrays. CNV location and size were annotated based on the NCBI build 37. DNA methylation data was used to estimate proportions of cell types for each sample using the method described by Houseman et al(Houseman et al., 2012).

Overlap evaluation between CNVs generated by methylation and genotype data—In the GTP dataset, 90 samples had methylation data from both the

HumanMethylation450 and Human MethylationEPIC arrays, and genotype data from the Illumina Omni1-Quad BeadChip. CNVs called from each platform were independently compared to the CNVs generated from the genotype data. For the purposes of this comparison, an overlap with any region of the CNVs called by the genotype data was considered to be an overlap as long as they are the same copy number type. The final overlap % was calculated for each method and array type using the formula:

$$\frac{\#\ \ CNVs\ detected\ by\ both\ genotype\ data\ and\ methylation\ data}{\#\ \ CNVs\ detected\ by\ the\ genotype\ data}$$

For comparison across methylation methods, CNVs are considered to be overlapping as long as they are identified within the same sample, have a positional overlap of greater than 50% and reflect the same copy number. Once the counts of overlapping CNVs between any two methods and all the three methods were determined, the R library VennDiagram (Chen & Boutros, 2011) was used to generate the Venn diagram.

Reliability evaluation—To evaluate reliability in the same person, CNVs were called for each participant in the AAMP study at both the baseline and the follow-up visits. For evaluations of reliability, we focused on CNVs >10 kb in size as smaller CNVs may be less reliably detected on these platforms. Only CNVs that were identified at both time points were considered to be reliable. Since the boundaries of a particular CNV call can vary between samples, CNVs were considered to be reliable if the CNVs identified at the two time points within an individual have an overlap of greater than 50% and are the same copy number variation.

Reliability was calculated and reported for each time point within each method/array and was calculated as follows:

$$\frac{Total\ \ \#\ \ reliable\ CNVs\ across\ all\ individuals}{Total\ \ \#\ \ CNVs\ detected\ across\ all\ individuals\ within\ a\ time\ point}$$

Reliability for each individual within each method/array was calculated as follows:

$$\frac{2*\ \#\ \ reliable\ CNVs\ detected}{\#\ CNVs\ detected\ in\ time\ point\ 1+\ \#\ CNVs\ detected\ in\ time\ point\ 2}$$

Results

Overlap of CNV calls from genotyping data

Genotyping data was available for 90 participants in the GTP study, all of which also had DNA methylation data assessed on both the HumanMethylation450 BeadChip and the MethylationEPIC BeadChip. The GTP cohort was used to assess the overlap between the genotype calls and the methylation calls, and, the overlap between the methylation methods. The number of CNVs called (Supplementary Table 1) using the genotype data was much lower than the number of CNVs called using the methylation data. Within the methods used to call CNVs using methylation data, ChAMP called the highest number of CNVs of the three methods across all subjects and within subject (Table 1) for both the 450k and EPIC arrays. Conumee called the fewest CNVs within the 450k array and cnAnalysis450k called the fewest CNVs with the EPIC array. The distribution of CNVs detected differed by methods and arrays is shown in Figure 1. The average size of CNVs called using the genotype data was much lower than the average size of CNVs called using the methylation data. Within the methylation methods, cnAnalysis450k detected CNVs with the largest size within both the 450k and EPIC arrays. The CNVs called by the three methods using methylation data were compared to the CNVs called using genotype data and the overlap percentages are reported in Table 2 for each individual array and method. ChAMP detected 62% of the CNVs detected by the genotype data for both the 450k and EPIC arrays. The overlap percentages for Conumee and cnAnalysis450k were substantially lower regardless of the array used. Interestingly, most of the CNVs detected using methylation data are not detected when using genotype data irrespective of the array/ method used.

Evaluation of CNV overlap across the DNA methylation methods

A higher degree of overlap was observed between the CNVs detected by cnAnalysis450k and those detected by Champ for both the EPIC array and 450k array. Figure 2 represents this comparison for both the arrays. There were 17 CNVs (Supplementary Figure 1) that were detected across all three methods within the 450K array and 71 CNVs that were detected across all three methods within the EPIC array.

Evaluation of reliability across CNVs called by methylation methods

DNA methylation data were available for 142 women in the AAMP study, with samples at both study timepoints over pregnancy. Of these 142 women, methylation was assayed via the HumanMethylation450 BeadChip for 53 (37%) participants and via the MethylationEPIC BeadChip for 89 (63%) participants. This cohort was used to calculate and analyze reliability measures. To compare reliability of detection across methods, we computed the percent of baseline calls that were reliable, and the percent of follow-up calls that were reliable. As detailed in Table 3, all three methods detected CNVs more reliably using the EPIC array. Conumee, which identified CNVs with a low reliability when using the 450k array, identified CNVs with the highest reliability of all three methods when the EPIC array was used (57.6%). cnAnalysis450k had the best reliability of the three methods when detecting CNVs using the 450k array (43.0%). However, there was a wide range of reliability metrics across subjects (Table 4; Supplementary Figure 2).

Evaluation of the impact of cell composition and array position on reliability

To evaluate whether differences in cell composition across blood samples were responsible for the variance in reliability at the sample level, the difference in the cellular composition (operationalized as percent CD8+ T cells, CD4+ T cells, natural killer cells, monocytes, B cells) between the matched samples was regressed on the individual reliability measures (Supplementary Table 2). CD8+ T cell composition was associated with individual reliability measures for Conumee within the 450K array; there were no other significant associations between the differences in cell types and the individual reliability measures. Also, considering that the number of CNVs detected varied across samples, the number of CNVs called per sample was evaluated for association with cell composition, batch, and position (Tables 5 and 6). The number of CNVs called per sample was highly associated with the chip the samples were run on within the EPIC array for all the methods and was associated with the chip within the 450k array for cnAnalysis450k (Supplementary Figure 3). As the number of CNVs detected is associated with the chip the samples were run on, the individual-specific reliability measures were tested for association with the chip they were run on and array position. Supplementary Table 3 shows that individual reliability of CNV calls is associated with chip for ChAMP within the EPIC array (F=3.0, p=5E-03) and for cnAnalysis450k (F=2.2, p=0.04) and ChAMP within the 450K array (F=1.7, p=0.05).

Discussion

Genotyping data is widely used to call CNVs and we compared calls generated from this gold-standard to those made by three methylation-based methods (ChAMP, cnAnalysis450k and Conumee). ChAMP calls 62% of the CNVs called by the genotype data but the other methods call a substantially lower number of CNVs called by the genotype data and this may partially be explained by the fact that ChAMP calls substantially more CNVs than Conumee and cnAnalysis450k. However, almost 95% of the CNVs called using methylation data are not called by genotype data irrespective of the array or method used. This can be attributed to the fact that the number of CNVs detected from the genotype data is much lower than the number of CNVs detected by any of the methylation-based methods. This is to be expected considering that, for the genotype data, only the CNVs that were commonly detected across two methods were used for this comparison. The methylation methods not only call substantially more CNVs but also substantially larger CNVs. In some cases, the packages call almost whole chromosomes as CNVs in subjects that are extremely unlikely to be cases of aneuploidy. Given the stark differences in the CNVs called by the methylation-based arrays compared to the ones called using the genotype arrays, we further examined the reliability of CNVs called using methylation-based methods.

Each of the three methods called more CNVs using the EPIC array compared to the 450K array, due to the increased CpG density within the EPIC array. ChAMP detected the highest number of CNVs regardless of the array but also had a tendency to detect very small CNVs. Conumee detected more CNVs than cnAnalysis450k within the EPIC array but detected fewer CNVs than cnAnalysis450k within the 450k array. However, none of the methods substantially agreed with each other, as the overlap between the CNVs detected by any two methods for the baseline samples did not exceed 36%. Of the three methods, cnAnalysis450k called CNVs that have a high degree of overlap with the ones called by ChAMP (36% within the 450k array and 30% within the EPIC array).

Calling CNVs using an EPIC array improved the reliability for all pipelines as the higher density of CpGs within the EPIC array allowed the algorithms to call the regions with a higher confidence than when calling within the 450k array. Across the pipelines and arrays, Conumee called CNVs with the highest reliability within an individual using the EPIC array. Conumee calls CNVs based on predefined regions unlike the other methods, which is the most likely reason for its high reliability. However, this pattern does not hold within the 450k array where cnAnalysis450k calls CNVs with the highest reliability. To evaluate the sensitivity of the pipelines to the reference samples used, all the pipelines were rerun for the

EPIC array using technical replicates as the reference instead of other samples from the same experiment. This change does not appear to substantially affect ChAMP and Conumee as ChAMP's reliability increased only slightly from 39.5% to 41.2%, and Conumee's reliability decreased slightly from 57.6% to 53.2%. However, cnAnalysis450k seems to be sensitive to the reference samples used as the reliability decreased from 47.3% to 32.9% when technical replicates were used instead of the other samples.

When the reliabilities were compared at the sample level, they varied dramatically. For example, even though Conumee called CNVs at a reliability of 57.6% on average, the individual-level reliability ranged from 0% to 82.5%. Data generated by Infinium methylation arrays is prone to batch effects but that is unlikely to affect the reliabilities greatly as both the samples from within an individual were run on the same chip. Hypothesizing that the variance in reliability was caused by the differing cell composition within the samples, the sample level reliabilities were regressed on the difference between the cell compositions. Although within the 450k array, change in CD8+ cells was associated with individual-specific reliability for Conumee, the same pattern was not detected within the EPIC array. The variance in the reliabilities was also due, in part, to the differing number of CNVs detected per sample. The number of CNVs detected per sample was associated with the chip the sample was run on and consequently, the reliability was mildly associated with chip for Champ within EPIC and for ChAMP and cnAnalysis450k within 450k. The cell types and array position do not seem to have a consistent effect on the number of CNVs called per sample across the methods and arrays.

From the results, it is clear that the EPIC array detects CNVs with a substantially higher reliability than the 450k array regardless of the method chosen, with the best achievable reliability being 57.6% using Conumee with the EPIC array. Also, considering that the number of CNVs detected per sample is highly associated with array designation, it is evident that running the pipelines with default parameters does not handle the array specific biases in the data and so, it may be prudent to call CNVs after adjusting for the effect of the array designation using methods such as ComBAT (Johnson, Li, & Rabinovic, 2007). Considering CNVs detected using array based methylation data is not highly reproducible and do not overlap with the ones detected using the genotype data, CNVs detected using the EPIC arrays are only useful as a first pass analysis. Hence, we recommend that researchers validate CNVs called using array based DNA methylation data with techniques such as FISH or quantitative PCR before analyzing, interpreting and reporting their biological implications.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Funding:

This study was supported in part by the National Institutes of Health, National Institute of Nursing Research [R01NR014800], National Institute on Minority Health and Health Disparities [R01MD009064], and the Office of the Director [UH3OD023318].

References

Binder EB, Bradley RG, Liu W, Epstein MP, Deveau TC, Mercer KB, … Ressler KJ (2008). Association of FKBP5 polymorphisms and childhood abuse with risk of posttraumatic stress disorder symptoms in adults. JAMA. 299(11), 1291–1305. doi: 10.1001/jama.299.11.1291 [PubMed: 18349090]

Chen H, & Boutros PC (2011). VennDiagram: a package for the generation of highly-customizable Venn and Euler diagrams in R. BMC Bioinformatics. 12, 35. doi: 10.1186/1471-2105-12-35 [PubMed: 21269502]

Cnv, Schizophrenia Working Groups of the Psychiatric Genomics, Consortium, & Psychosis Endophenotypes International, Consortium. (2017). Contribution of copy number variants to schizophrenia from a genome-wide study of 41,321 subjects. Nat Genet. 49(1), 27–35. doi: 10.1038/ng.3725 [PubMed: 27869829]

Colella S, Yau C, Taylor JM, Mirza G, Butler H, Clouston P, … Ragoussis J (2007). QuantiSNP: an Objective Bayes Hidden-Markov Model to detect and accurately map copy number variation using SNP genotyping data. Nucleic Acids Res. 35(6), 2013–2025. doi: 10.1093/nar/gkm076 [PubMed: 17341461]

Corwin EJ, Hogue CJ, Pearce B, Hill CC, Read TD, Mulle J, & Dunlop AL (2017). Protocol for the Emory University African American Vaginal, Oral, and Gut Microbiome in Pregnancy Cohort Study. BMC Pregnancy Childbirth. 17(1), 161. doi: 10.1186/s12884-017-1357-x [PubMed: 28571577]

Cuccaro D, De Marco EV, Cittadella R, & Cavallaro S (2017). Copy Number Variants in Alzheimer's Disease. J Alzheimers Dis. 55(1), 37–52. doi: 10.3233/JAD-160469 [PubMed: 27662298]

Di Gregorio E, Gai G, Botta G, Calcia A, Pappi P, Talarico F, … Brussino A (2015). Array-Comparative Genomic Hybridization Analysis in Fetuses with Major Congenital Malformations Reveals that 24% of Cases Have Pathogenic Deletions/Duplications. Cytogenet Genome Res. 147(1), 10–16. doi: 10.1159/000442308 [PubMed: 26658296]

Du L, Sun W, Li XM, Li XY, Liu W, & Chen D (2017). DNA methylation and copy number variation analyses of human embryonic stem cell-derived neuroprogenitors after low-dose decabromodiphenyl ether and/or bisphenol A exposure. Hum Exp Toxicol, 960327117710535. doi: 10.1177/0960327117710535

Feber A, Guilhamon P, Lechner M, Fenton T, Wilson GA, Thirlwell C, … Beck S (2014). Using high-density DNA methylation arrays to profile copy number alterations. Genome Biol. 15(2), R30. doi: 10.1186/gb-2014-15-2-r30 [PubMed: 24490765]

Feuk L, Carson AR, & Scherer SW (2006). Structural variation in the human genome. Nat Rev Genet. 7(2), 85–97. doi: 10.1038/nrg1767 [PubMed: 16418744]

Gillespie CF, Bradley B, Mercer K, Smith AK, Conneely K, Gapen M, … Ressler KJ (2009). Trauma exposure and stress-related disorders in inner city primary care patients. Gen Hosp Psychiatry. 31(6), 505–514. doi: 10.1016/j.genhosppsych.2009.05.003 [PubMed: 19892208]

Houseman EA, Accomando WP, Koestler DC, Christensen BC, Marsit CJ, Nelson HH, … Kelsey KT (2012). DNA methylation arrays as surrogate measures of cell mixture distribution. BMC Bioinformatics. 13, 86. doi: 10.1186/1471-2105-13-86 [PubMed: 22568884]

Hovestadt V, & Zapatka M conumee: Enhanced copy-number variation analysis using Illumina DNA methylation arrays (Version 1.9.0). Retrieved from http://bioconductor.org/packages/conumee/

International Schizophrenia, Consortium. (2008). Rare chromosomal deletions and duplications increase risk of schizophrenia. Nature. 455(7210), 237–241. doi: 10.1038/nature07239 [PubMed: 18668038]

Johnson WE, Li C, & Rabinovic A (2007). Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics. 8(1), 118–127. doi: 10.1093/biostatistics/kxj037 [PubMed: 16632515]

Knight AK, Conneely KN, Kilaru V, Cobb D, Payne JL, Meilman S, … Smith AK. (2017). SLC9B1 methylation predicts fetal intolerance of labor. Epigenetics, 1–22. doi: 10.1080/15592294.2017.1411444 [PubMed: 27830979]

Knoll M, Debus J, & Abdollahi A (2017). cnAnalysis450k: an R package for comparative analysis of 450k/EPIC Illumina methylation array derived copy number data. Bioinformatics. 33(15), 2266–2272. doi: 10.1093/bioinformatics/btx156 [PubMed: 28379302]

Lockwood WW, Chari R, Chi B, & Lam WL (2006). Recent advances in array comparative genomic hybridization technologies and their applications in human genetics. Eur J Hum Genet. 14(2), 139–148. doi: 10.1038/sj.ejhg.5201531 [PubMed: 16288307]

Malhotra D, McCarthy S, Michaelson JJ, Vacic V, Burdick KE, Yoon S, … Sebat J (2011). High frequencies of de novo CNVs in bipolar disorder and schizophrenia. Neuron. 72(6), 951–963. doi: 10.1016/j.neuron.2011.11.007 [PubMed: 22196331]

Marshall CR, Marshall CR, Howrigan DP, Merico D, Thiruvahindrapuram B, Wu WT, … Endophenotypes, Psychosis. (2017). Contribution of copy number variants to schizophrenia from a genome-wide study of 41,321 subjects. Nature Genetics. 49(1), 27–35. doi: 10.1038/ng.3725 [PubMed: 27869829]

Martinez VD, Buys TP, Adonis M, Benitez H, Gallegos I, Lam S, … Gil L (2010). Arsenic-related DNA copy-number alterations in lung squamous cell carcinomas. Br J Cancer. 103(8), 1277–1283. doi: 10.1038/sj.bjc.6605879 [PubMed: 20842114]

McCarroll SA, Hadnott TN, Perry GH, Sabeti PC, Zody MC, Barrett JC, … International HapMap, Consortium. (2006). Common deletion polymorphisms in the human genome. Nat Genet. 38(1), 86–92. [PubMed: 16468122]

Pinto D, Marshall C, Feuk L, & Scherer SW (2007). Copy-number variation in control population cohorts. Hum Mol Genet. 16 Spec No. 2, R168–173. doi: 10.1093/hmg/ddm241 [PubMed: 17911159]

Poniah P, Mohd Zain S, Abdul Razack AH, Kuppusamy S, Karuppayah S, Sian Eng H, & Mohamed Z (2017). Genome-wide copy number analysis reveals candidate gene loci that confer susceptibility to high-grade prostate cancer. Urol Oncol. doi: 10.1016/j.urolonc.2017.04.017

Quintela I, Eiris J, Gomez-Lado C, Perez-Gay L, Dacruz D, Cruz R, … Barros F (2017). Copy number variation analysis of patients with intellectual disability from North-West Spain. Gene. doi: 10.1016/j.gene.2017.05.032

Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, … Hurles ME (2006). Global variation in copy number in the human genome. Nature. 444(7118), 444–454. doi: 10.1038/nature05329 [PubMed: 17122850]

Ressler KJ, Mercer KB, Bradley B, Jovanovic T, Mahan A, Kerley K, … May V (2011). Post-traumatic stress disorder is associated with PACAP and the PAC1 receptor. Nature. 470(7335), 492–497. doi: 10.1038/nature09856 [PubMed: 21350482]

Wang K, Li M, Hadley D, Liu R, Glessner J, Grant SF, … Bucan M (2007). PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. Genome Res. 17(11), 1665–1674. doi: 10.1101/gr.6861907 [PubMed: 17921354]

Zarrei M, MacDonald JR, Merico D, & Scherer SW (2015). A copy number variation map of the human genome. Nat Rev Genet. 16(3), 172–183. doi: 10.1038/nrg3871 [PubMed: 25645873]

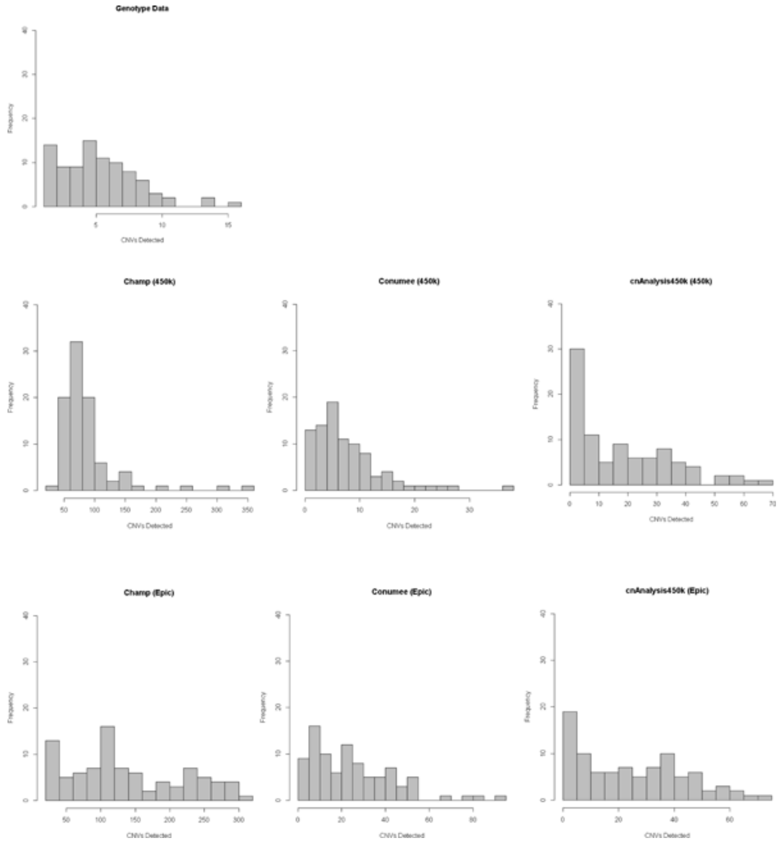

Figure 1.

Histograms of the number of CNVs detected per sample for the genotype data and the methylation data separated by method/array. The x-axis represents the number of CNVs detected per sample and the y-axis represents the frequency.

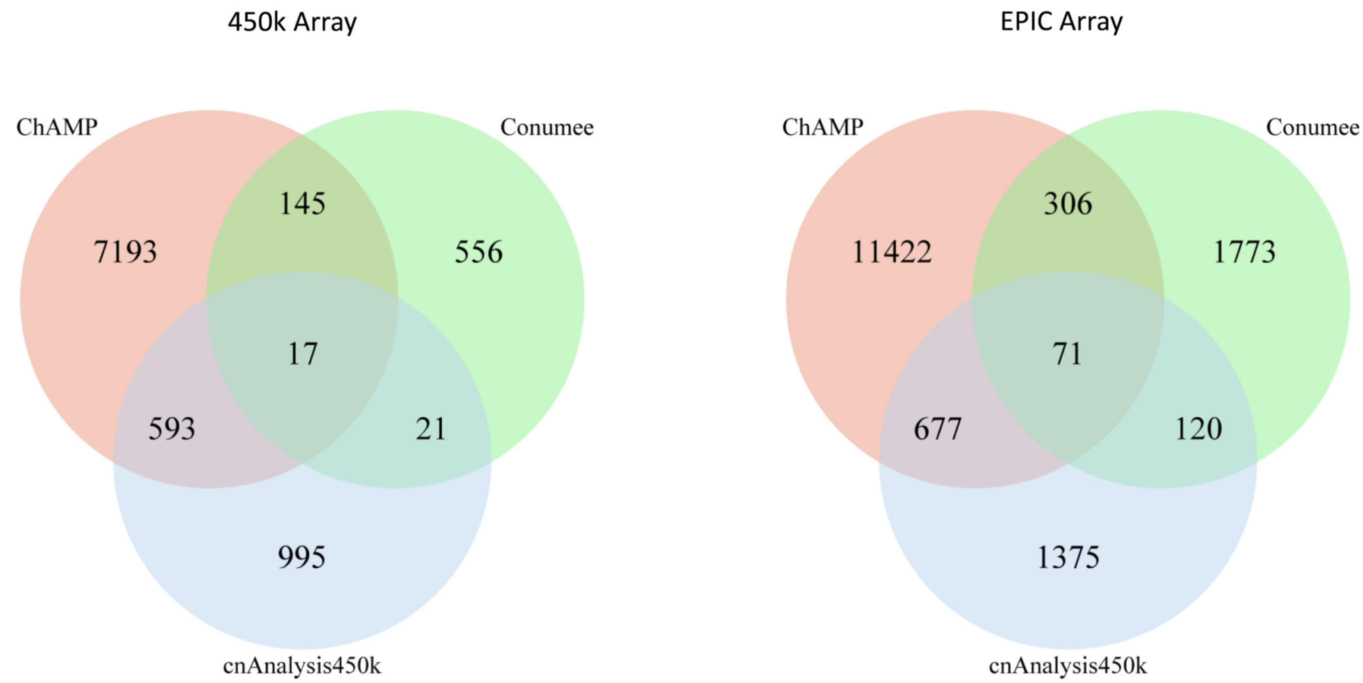

Figure 2.

The Venn diagrams below represents the total number of CNVs detected for each method, the number of CNVs that overlap between the two methods, and the number of CNVs that overlap between all three methods for both the arrays.

Table 1.

Average number and size of the CNVs called by genotype and methylation data for each array and method within the GTP cohort for the same subjects.

	450k, N=90		**EPIC, N= 90**	
	Average #CNVs	**Size (in kb)**	**Average #CNVs**	**Size (in kb)**
Genotype	5	103	5	103
CHAMP	88	31,540	138	20,124
Conumee	8	38,465	25	28,281
cnAnalysis450k	21	48,832	29	49,218

Table 2.

Overlap between the CNVs called by genotype data and the methylation data for all the three methods within the GTP cohort.

	450k Array		EPIC Array	
	% of genotype CNVs called by methylation data	**% of methylation CNVs called by genotype data**	**% of genotype CNVs called by methylation data**	**% of methylation CNVs called by genotype data**
CHAMP	62.3%	4.0%	62.1%	2.5%
Conumee	7.8%	5.4%	14.3%	3.2%
cnAnalysis450k	12.7%	4.0%	14.7%	3.3%

Table 3.

Reliability of CNVs called by method for the 450k and EPIC arrays. Data is presented at the baseline visit, notated as V1, and the follow-up visit, notated as V2. The values in the table represent the total number of the CNVs detected across all the samples for each method/time point.

450k Array, N=106 (53 pairs)						
	V1 & V2	V1	Reliability V1 (%)	V2	Reliability V2 (%)	Overall Reliability (V1+V2)/2
CHAMP	1,854	6,506	28.5%	6,560	28.3%	28.4%
Conumee	290	1,271	22.8%	1,379	21.0%	21.9%
cnAnalysis450k	1,365	3,240	42.1%	3,107	43.9%	43.0%
EPIC Array, N=178 (89 pairs)						
	V1 & V2	V1	Reliability V1 (%)	V2	Reliability V2 (%)	Overall Reliability (V1+V2)/2
CHAMP	39,550	97,321	40.6%	103,032	38.4%	39.5%
Conumee	6,780	11,412	59.4%	12,120	55.9%	57.6%
cnAnalysis450k	1,891	4,036	47.1%	3,992	47.4%	47.3%

Table 4.

Summary statistics of the individual reliabilities for each method/array.

	450k Array, N=106 (53 pairs)			**EPIC Array, N=178 (89 pairs)**		
	Mean	**SD**	**Range**	**Mean**	**SD**	**Range**
CHAMP	30.8%	13.7%	3.4%−63.6%	35.4%	19.4%	1.2%−78.2%
Conumee	19.2%	18.2%	0.0%−85.7%	51.0%	25.4%	0.0%−82.5%
cnAnalysis450k	39.8%	18.7%	0.0%−73.0%	40.9%	24.0%	0.0%−82.3%

Table 5.

Association between the number of CNVs detected and cellular composition.

	450k Array, N=106 (53 pairs)						**EPIC Array, N= 178 (89 pairs)**					
	ChAMP		**Conumee**		**cnAnalysis 450k**		**ChAMP**		**Conumee**		**cnAnalysis 450k**	
	t	**p**	**t**	**p**	**t**	**p**	**t**	**p**	**t**	**p**	**t**	**p**
CD8	−0.1	0.96	−0.4	0.71	0.6	0.56	**2.6**	**1E-02**	1.5	0.12	1.2	0.25
CD4	0.5	0.64	−1.5	0.15	0.9	0.39	−0.9	0.34	−1.2	0.23	0.8	0.45
NK	−1.4	0.15	−1	0.3	0.1	0.97	**−2.4**	**1E-02**	−0.8	0.42	−1.5	0.13
BCell	−0.2	0.86	−0.1	0.93	−0.3	0.74	0.5	0.60	0.4	0.72	1.9	0.06
Mono	−0.5	0.6	0.9	0.37	−1	0.31	−0.1	0.93	−0.7	0.47	−1.5	0.13

Table 6.

Association between the number of CNVs and array designation (chip) or position on the array (row).

	450k, N=106 (53 pairs)						**EPIC N= 178 (89 pairs)**					
	ChAMP		**Conumee**		**cnAnalysis450k**		**ChAMP**		**Conumee**		**cnAnalysis450k**	
	F	**p**	**F**	**p**	**F**	**p**	**F**	**p**	**F**	**p**	**F**	**P**
Chip	1.8	0.06	1.1	0.41	**3**	**2E-03**	**7.2**	**4.3E-16**	**5.8**	**5.6E-13**	**2.9**	**4.8E-05**
Row	1.5	0.13	1.7	0.09	0.7	0.77	0.8	0.6	1	0.45	2	0.06