Assessment of Copy Number Variation using the Illumina Infinium 1M SNP-array: A comparison of methodological approaches in the Spanish Bladder Cancer / EPICURO Study.

Gaëlle Marenne, Benjamín Rodríguez Santiago, Montserrat García-Closas, Luis Alberto Pérez Jurado, Nathaniel Rothman, Daniel Rico, Guillermo Pita, David G Pisano, Manolis Kogevinas, Debra T Silverman, et al.

Human Mutation

Assessment of Copy Number Variation using the Illumina Infinium 1M SNP-array: A comparison of methodological approaches in the Spanish Bladder Cancer / EPICURO Study.

Journal:	*Human Mutation*
Manuscript ID:	humu-2010-0239.R1
Wiley - Manuscript type:	Methods
Date Submitted by the Author:	30-Sep-2010
Complete List of Authors:	Marenne, Gaëlle; Spanish National Cancer Research Centre Rodríguez Santiago, Benjamín; Departament de Ciències Experimentals i de la Salut, Universitat Pompeu Fabra García-Closas, Montserrat; Division of Cancer Epidemiology and Genetics, National Cancer Institute, Department of Health and Human Services Pérez Jurado, Luis; Universitat Pompeu Fabra, Ciències Experimentals i de la Salut; Hospital Vall d'Hebron, Program in Molecular Medicine and Genetics Rothman, Nathaniel; Division of Cancer Epidemiology and Genetics, National Cancer Institute, Department of Health and Human Services Rico, Daniel; Spanish National Cancer Research Centre Pita, Guillermo; Spanish National Cancer Research Centre Pisano, David; Spanish National Cancer Research Centre Kogevinas, Manolis; Centre for Research in Environmental Epidemiology Silverman, Debra; Division of Cancer Epidemiology and Genetics, National Cancer Institute, Department of Health and Human Services Valencia, Alfonso; Spanish National Cancer Research Centre Real, Francisco; Spanish National Cancer Research Centre Chanock, Stephen; National Cancer Institute, Pediatric Oncology Branch Génin, Emmanuelle; Inserm UMR-S946, Univ. Paris Diderot, Institut Universitaire d'Hématologie, Malats, Núria; Spanish National Cancer Research Centre
Key Words:	Copy Number Variation, Genome Wide Association Study, Specificity, Sensitivity, Reliability, Accuracy, CNVpartition,

Title: Assessment of Copy Number Variation using the Illumina Infinium 1M SNP-array: A comparison of methodological approaches in the Spanish Bladder Cancer / EPICURO Study.

Authors: Gaëlle Marenne(1, 2), Benjamín Rodríguez-Santiago(3, 4), Montserrat García Closas(5), Luis Pérez-Jurado(3, 4, 6, 7), Nathaniel Rothman(5), Daniel Rico(1), Guillermo Pita(1), David G. Pisano(1), Manolis Kogevinas(8, 9), Debra T Silverman(5), Alfonso Valencia(1), Francisco X Real(1), Stephen Chanock*(5), Emmanuelle Génin*(2), Núria Malats*(1). (* co-senior authors)

Affiliations of authors: (1) Centro Nacional de Investigaciones Oncológicas (CNIO) Madrid, Spain; (2) Inserm UMR-S946, Univ. Paris Diderot, Institut Universitaire d'Hématologie, Paris, France; (3) Departament de Ciències Experimentals i de la Salut, Universitat Pompeu Fabra, Barcelona, Spain; (4) CIBER de Enfermedades Raras, CIBERER, E-08003 Barcelona, Spain; (5) Division of Cancer Epidemiology and Genetics, National Cancer Institute, Department of Health and Human Services, Bethesda, MD, USA; (6) Programa de Medicina Molecular i Genètica, Hospital Universitari Vall d'Hebron, E-08035 Barcelona, Spain; (7) Department of Genome Sciences, University of Washington, Seattle, WA 98195, United States; (8) Institut Municipal d'Investigació Mèdica (IMIM-Hospital del Mar), Barcelona, Spain; (9) Centre for Research in Environmental Epidemiology (CREAL), Barcelona, Spain.

Corresponding author: Núria Malats (nmalats@cnio.es).

Running Title: Accuracy study on CNV assessment

Conflict of interest statement: We declare we have no conflict of interest.

Deleted: 29/09/2010

ABSTRACT

High-throughput SNP-array technologies allow to investigate CNVs in genome-wide scans and specific calling algorithms have been developed to determine CNV location and copy number.

We report the results of a reliability analysis comparing data from 96 pairs of samples processed with CNVpartition, PennCNV and QuantiSNP for Infinium Illumina Human 1Million probe chip data. We also performed a validity assessment with multiplex ligation-dependent probe amplification (MLPA) as a reference standard.

The number of CNVs per individual varied according to the calling algorithm. Higher numbers of CNVs were detected in saliva than in blood DNA samples regardless of the algorithm used. All algorithms presented low agreement with mean Kappa Index (KI) <66. PennCNV was the most reliable algorithm (KI_w=98.96) when assessing the number of copies. The agreement observed in detecting CNV was higher in blood than in saliva samples. When comparing to MLPA, all algorithms identified poorly known copy aberrations (sensitivity=0.19-0.28). In contrast, specificity was very high (0.97-0.99). Once a CNV was detected, the number of copies was truly assessed (sensitivity>0.62).

Our results indicate that the current calling algorithms should be improved for high performance CNV analysis in genome-wide scans. Further refinement is required to assess CNVs as risk factors in complex diseases.

Key Words: Copy Number Variation, Genome Wide Association Study, Specificity, Sensitivity, Reliability, Accuracy, CNVpartition, PennCNV, QuantiSNP

Deleted: 29/09/2010

INTRODUCTION

Structural variations of the human genome emerge as novel major contributors to genetic diversity and disease susceptibility. Copy number variation (CNV) refers to deletions or duplications larger than 1kb (Feuk et al., 2006). It was estimated that 12% of the genome could be affected by such variants in comparison to 1-2% covered by single nucleotide polymorphisms (SNPs) (Redon et al., 2006); although a recent study provided a lower figure: 3.7% (Conrad et al., 2010). These large variations can overlap with genes and there is substantial evidence for correlation between CNVs and gene expression levels (Stranger et al., 2007). CNVs are also known to be involved both in mendelian disorders, such as Williams–Beuren Syndrome (deletion at chromosome region 7q11.23) or Charcot–Marie Tooth neuropathy Type 1A (duplications at chromosome region 17p11.2), and complex traits such as HIV infection and asthma, among others (Ionita-Laza et al., 2009).

Recently, efforts have been made to provide resources supporting studies of structural variation in human diseases such as the Database of Genomic Variation which annotates genomic coordinates along with estimated frequencies of the CNVs (Conrad et al., 2010; Iafrate et al., 2004; Redon et al., 2006). However, the cost and the complexity of CNV assessment have restricted CNV studies to a list of carefully selected candidate genes. The possibility to study CNVs at a genome-wide scale is now possible using high-throughput SNP-array technologies. The new-generation SNP-arrays, such as the Infinium Illumina Human 1Million probe chip and the Affymetrix 6.0 platform, allow a cost-effective detection of CNVs by interpreting allele intensities for each marker. These platforms also include monomorphic probes in regions of common CNVs that presented technical problems for SNP array design due to a lack of polymorphic probes or because of disruption from Mendelian inheritance and Hardy-Weinberg equilibrium. The Illumina 1 Million SNP-array works with Beadstudio software that provides the variables used to perform the CNV calling. Different

Deleted: 29/09/2010

algorithms can then be employed to locate CNVs by finding breakpoints and assessing the number of copies present per individual. The most frequently-used algorithms for Illumina data are CNVpartition – an Illumina developed plug-in –, PennCNV (Wang et al., 2007) and QuantiSNP (Colella et al., 2007).

Several studies have successfully assessed the role of CNVs in complex diseases such as asthma, autism, schizophrenia or cancer by applying high throughput analysis at genome-wide level (Bae et al., 2008; Bassett et al., 2008; Blauw et al., 2008; Cronin et al., 2008; Diskin et al., 2009; Friedman et al., 2006; Glessner et al., 2009; Greenway et al., 2009; InternationalSchizophreniaConsortium, 2008; Ionita-Laza et al., 2008; Kathiresan et al., 2009; Liu et al., 2009; Marshall et al., 2008; Matarin et al., 2008; Need et al., 2009; Sha et al., 2009; Simon-Sanchez et al., 2008; Stefansson et al., 2008; Walsh et al., 2008; Weiss et al., 2008; Xu et al., 2008; Yang et al., 2008). A review of these studies indicates that they have used a wide range of methodologies, thus raising the issue of comparability of discovery rates. The rapid development of technologies in this field has not been accompanied by a careful evaluation of the software tools to assess disease risk association. In contrast to the nearly 100% concordance observed for bi-allelic genotypes, a recent study reported very low agreement estimates when the performance of different algorithms assessing CNV was compared using HapMap data (Winchester et al., 2009).

Here, we report the results from reliability and validity analyses comparing three CNV calling algorithms for Illumina 1M probe-array data (CNVpartition, PennCNV and QuantiSNP) using multiplex ligation-dependent probe amplification (MLPA) as the gold-standard analysis. The study was conducted on 96 duplicate samples from the Spanish Bladder Cancer Study. We also assessed whether the source of DNA (blood or saliva) and the number and type of SNPs considered in the CNV definition influenced the performance of the SNP calling algorithms.

Deleted: 29/09/2010

MATERIALS AND METHODS

Samples and genotyping data

Study subjects were recruited to the Spanish Bladder Cancer Study (SBCS)/EPICURO, conducted between 1998-2000. Individuals were from 5 different regions in Spain (Barcelona, Vallès/Bages, Alicante, Tenerife and Asturias). Leukocyte and saliva DNA were obtained as described elsewhere (Garcia-Closas et al., 2005). Genotyping was performed at the Core Genotyping Facility, National Cancer Institute, USA, using the Infinium Illumina Human 1M probe BeadChip containing 1,072,820 markers, among which 206,665 are in reported CNVs regions. For quality control reasons, 141 individuals were genotyped two to four times providing genetic data for 178 pairs out of 299 assays (Supp. Table S1).

Log R Ratio (LRR) and the B Allele Frequency (BAF) were exported from the normalized Illumina data through the Beadstudio software to perform CNV calling. LRR is the ratio between the observed and the expected probe intensity. The expected intensity is an interpolation of the mean intensities of the surrounding genotype clusters. BAF represents the proportion of B alleles in the genotype. A region without evidence of CNV should show a LRR around zero and three clusters of BAF of 0, 0.5 and 1 corresponding to the three genotypes AA, AB and BB, respectively (Supp. Figure S1). Individuals not fitting at least one of the CNV specific quality control metric recommended by PennCNV (Wang et al., 2007) were excluded from the analysis: LRR-Standard Deviation>0.28, 0.45>BAF-median>0.55, BAF-drift>0.002, and -0.04>Wave Factor>0.04. After applying the abovementioned criteria, 92 individuals (90 duplicates and 2 triplicates) were suitable for this study, thus providing 96 pairs for comparison (90 from duplicate individuals and 6 from triplicate individuals) and 186 assays (90 individuals * 2 samples and 2 individuals * 3 samples). Among the duplicates there were 63 and 33 pairs from blood and saliva samples, respectively (Supp. Table S1).

Deleted: 29/09/2010

CNV calling

Three algorithms available for Illumina data were applied: CNVpartition, PennCNV (Wang et al., 2007) and QuantiSNP (Colella et al., 2007). CNVpartition was developed by Illumina and is available as a plug-in in the Beadstudio software. It is based on the assumption that the majority of CNV vary between 0 and 4 copies (i.e. AAAA, AAAB, AABB …) thus yielding five options (homozygous deletion, heterozygous deletion, dizygous (normal state), trizygous (one extra copy), and tetrazygous (two extra copies). CNVpartition model LRR and BAF as simple bivariate Gaussian distributions for each of the fourteen possible copy genotypes. A preliminary copy number estimate is computed for each assayed locus by comparing its observed LRR and BAF to values predicted from each of the fourteen genotypes. Specifically, the likelihood of observing a given LRR and BAF under each of the fourteen models is computed and the number of copies is estimated by maximizing the likelihood. Once each probe is assigned a number of copies, breakpoints are determined by a partitioning method identifying regions where the estimated number of copies of the probes inside and outside the region is different. A confidence value is also provided to allow the filtering of the CNV and limit the number of false positive callings.

PennCNV and QuantiSNP are algorithms developed by academic teams and freely available (Colella et al., 2007; Wang et al., 2007). They are both based on a Hidden Markov Model (HMM) in which the number of gene copies is the hidden state and the LRR and the BAF are the two observed states that are considered independent of each other given the number of copies. A first-order HMM is considered where the number of copies at one probe depends on the number of copies at the previous probe. However, the two algorithms differ in their transition and emission probabilities. While transition probabilities depend on the distance between adjacent probes for both approaches, the probabilities for PennCNV are also state-specific, accounting for the fact that some state transition events (e.g., from normal state to

Deleted: 29/09/2010

heterozygous deletion) are more likely than others (e.g., from heterozygous deletion to trizygous). Regarding the BAF emission probabilities, PennCNV uses a more sophisticated model than QuantiSNP. Both algorithms provide a confidence value to filter CNVs. For QuantiSNP, the confidence value is the Log Bayes Factor (LBF). All algorithms were used with their default options and CNV calls from QuantiSNP with a LBF lower than 10 were filtered out as recommended whereas no filter was applied on CNVpartition and PennCNV calls.

Each of the 1,029,591 probes of the Illumina 1M array corresponding to the autosomal chromosomes was assigned with an estimated number of copies if were included in a CNV and with two copies otherwise. This procedure was applied to each of the 186 experiments performed in this study and for each of the algorithms.

Reliability analysis

The calling agreement between duplicates was evaluated for each of the algorithms to determine presence of CNV and number of copies. First, we assessed the agreement in detecting the presence of an aberration by estimating the kappa index (KI) between duplicates. KI compared the observed agreement against the agreement expected by chance in all the probes (Cohen, 1960). For probes in which the algorithm was concordant in detecting an aberration, we computed the agreement in assessing the number of copies by estimating the weighted Kappa Index (KI_w). This was done by applying quadratic weights that decreased while increasing differences in copy numbers (Supp. Figure S2). A total of 96 KI and KI_w values were obtained for each algorithm. Summary statistics (mean, median, standard-deviation, and quartiles) were computed and differences between algorithms were tested using paired t-tests.

To further limit the number of false positive CNV callings from SNP-array platforms, Itsara et al proposed to filter the called CNVs according to the type of aberration and the number of

Deleted: 29/09/2010

genotyped SNPs included in the CNV (Itsara et al., 2009). The LRR intensities were transformed into standard normal measurements (Z-scores) and the B-deviation value for each probe was estimated. Putative CNVs were classified into two categories (small and large) according to a cut-off of 100 probes and 1 Mb length. Large CNVs were manually curated. Small CNVs were subject to automated filtering. Homozygous deletions were required to comply with: 1) ≥ 3 probes, median LRR Z-score ≤ —4, and mean B-deviation ≥ 0.1 or 2) ≥ 3 probes and median LRR Z-score ≤ -8. Heterozygous deletions were required to span ≥10 probes, have LRR Z-score ≤ -1.5, and less than 10% of probes called as heterozygous. To define duplications, the requirements were: ≥ 10 probes, LRR Z-score ≥ 1.5, and B-deviation among heterozygote probes ≥ 0.075. The reliability of applying the Itsara's filter was assessed, too.

We analyzed the calling agreement of paired samples depending on the DNA source by stratifying the data according to whether the DNA was from blood (N=63) or saliva (N=33). In addition, we assessed whether the number of SNPs included in each CNV influenced the agreement rate by comparing the CNV calling performance between replicates by filtering for the number of SNPs in the CNVs. The reliability results were plotted for the three algorithms and the number of CNVs called according to the number of SNPs.

Select commercial SNP genotyping platforms contain monomorphic probes in regions of known common CNVs to facilitate analysis, particularly when prior analyses in HapMap indicated a substantial problem of fitness with Hardy Weinberg proportions. The overall percentage of monomorphic probes in the 1M Illumina Infinium platform in autosomal chromosomes is 1.4% (14,716/1,029,591). To test the impact of the type of probe (monomorphic or polymorphic) on the reliability of the calling, we compared for these two types of probes the ratio of concordant vs. discordant probes included in CNVs. We excluded the regions with a concordant result for the absence of CNV because the density of the

Deleted: 29/09/2010

monomorphic probes in those regions was lower according to the design of the SNP-array, hence not being comparable.

Validity Study

Multiplex ligation-dependent probe amplification (MLPA) assay is a standard laboratory approach to assess differences in the number of alleles copies at a particular locus. It is based on hybridization, specific probe ligation, amplification and capillary migration, and it was used as the gold-standard method to assess the number of copies of a given sequence. Regions were selected for validation with MLPA if at least one algorithm detected a minimum of 8 individuals carrying a CNV to avoid performing experiments in regions where no CNV exist. Commercial probe mixes (kits P070 and P036 covering the selected regions (MRC-Holland Amsterdam, The Netherlands) and custom designed probes (Supp. Table S2) were used. MLPA reactions were carried out as described previously (Schouten et al., 2002) with slight modifications when custom probes were used (Rodriguez-Santiago et al., 2009). The relative peak height (RPH) method recommended by MRC-Holland was used to determine the copy number status. Theoretically, heterozygous deletions and duplications showed a relative peak height of approximately 0.5 and 1.5, respectively. Only blood samples were considered for this analysis.

Leukocyte DNA from 56 individuals was analyzed twice by MLPA, providing a concordance rate of 97.25%. Among the discordant assays, 10 showing a "non-calling" rate greater than 70% were re-analyzed. Since the results of four of them slightly improved after the 2nd MLPA run they were included in the validity study and data were updated.

To assess the validity of each algorithm, sensitivity, specificity, and positive and negative predictive values were computed by comparing CNV callings with MLPA data. Sensitivity (SE) indicates the proportion of CNV identified by the algorithm over the total number of existing CNV according to MLPA. Specificity (SP) is the proportion of the non-CNV by an

Deleted: 29/09/2010

algorithm over the true non-CNV number. Positive (PPV) and negative predictive values (NPV) indicate the proportion of the true CNV and the true non-CNV over all CNV and non-CNV regions each algorithm assigns, respectively. These estimates are given as proportions with a 95%CI for the overall aberration assessment and for each type of CNV. The validity analysis considered those probes and individuals that provided agreement in detecting CN event according to each algorithm.

Statistical analyses were performed in R version 2.9.0 (http://www.r-project.org) with the *epiR* package (Mark Stevenson, http://epicentre.massey.ac.nz). Significance was declared when the p-value was smaller than 0.05.

RESULTS

The number of CNVs detected per individual varied substantially according to the calling algorithm (Table 1). CNVpartition identified an average of 28.0 CNVs per individual whereas the two algorithms based on the HMM, PennCNV and QuantiSNP, identified a median CNV number of 58.5 and 56.0, respectively. The number of CNVs per individual detected in saliva DNA was higher than in leukocyte DNA, regardless of the algorithm used (Table 1).

Reliability analysis

The SNP calling provided by the genotyping platform showed a very high agreement with a mean Kappa Index (KI) of 99.99 (95%CI, 99.94 – 100) (Figure 1a). The distribution of this KI was similar for experiments using blood or saliva DNA. Regarding CNV assessment in duplicate samples, PennCNV, QuantiSNP, and CNVpartition presented a lower agreement with mean KI values of 65.10, 63.09, and 57.24, respectively. The KI distribution based on CNVpartition callings significantly differed from that based on PennCNV and QuantiSNP callings ($p=2.68x10^{-10}$ and $p=7.28x10^{-5}$, respectively) (Figure 1b). Once a region of CNV was detected, the algorithms also showed differences in the KI distribution when assessing the

Deleted: 29/09/2010

number of copies (Figure 1c). PennCNV appeared to be the most reliable algorithm with an average KI_w (weighted KI) = 98.96 for the 96 pairs of replicates, and regardless the type of CNV (gain or loss). However, QuantiSNP and CNVpartition performed differently and poorly (Supp. Figure S3). This figure was significantly higher than those of CNVpartition (KI_w=94.55, p= $5.18x10^{-5}$) and QuantiSNP (KI_w=92.88, p=$7.43x10^{-8}$). Applying the Itsara filtering method, we did not observe an improvement of the agreement neither at the CNV detection level nor at the level of copy number (Supp. Figure S4).

Regardless of the algorithm applied, the agreement observed in detecting CNV was always higher in blood than in saliva samples (Figure 2), although the difference of the mean KI was only significant for CNVpartition and PennCNV callings (p=$3.93x10^{-7}$ and p=$8.16x10^{-5}$, respectively). The distribution of KI_w when assessing the number of copies, according to the DNA source, was similar for all algorithms (data not shown).

The number of probes selected by each algorithm to identify CNVs varied widely: 1,742 for CNVpartition, 2,361 for PennCNV, and 4,591 for QuantiSNP (Table 2). The percentage of probes showing agreement for the presence of a CNV was significantly different for the three algorithms: 37.7%, 50.7%, and 55.5% for CNVpartition, PennCNV, and QuantiSNP, respectively, (p=$2.43\ x10^{-35}$). The ratio between discordant/concordant probes was higher for monomorphic than polymorphic probes: 2.17 vs. 1.61 for CNVpartition (p=0.09), 1.78 vs. 0.94 for PennCNV (p=$4.34x10^{-4}$), and 1.51 vs. 0.72 for QuantiSNP (p=$1.31x10^{-17}$).

The correlation between the calling agreement and the number of probes or the length of a given CNV region is shown in Figure 3. A direct relationship between agreement and the number of probes included in the CNVs was observed suggesting that reliability is greater for CNVs containing more probes. This effect was observed for all algorithms but it was higher for PennCNV. Our results also suggested that filtering CNVs by QuantiSNP for length, by

Deleted: 29/09/2010

PennCNV for length lower than 500 kb or by CNVpartition for length lower than 1Mb did not increase the reliability.

Validity analysis

Sensitivity (SE) and Specificity (SP) estimates for the presence and the type of CNV were estimated according to each algorithm (Figure 4). When considering the presence of CNVs (first line in Figure 4), we found that none of the algorithms used identified known CNV well ($0.19 \leq SE \leq 0.28$]). In contrast, SP was very high ($0.97 \leq SP \leq 0.99$]), indicating that algorithms rarely assigned a CNV in a region where it did not exist. QuantiSNP showed the best SE (0.28) with a SP of 0.97, similar to that of the other two algorithms. Nonetheless, the false positive (FP) calling rate for this algorithm (FP=34) was 2.8-fold higher compared to CNVpartition (FP=12), the latter showing the highest SP (0.99) and the lowest SE (0.19) (Supp. Table S3). PennCNV presented intermediate values of SE (0.23) and SP (0.98), yielding 22 false positive CNVs out of 1319 true "non-CNV".

We also aimed at assessing whether copy number was well estimated when a CNV was identified. Since MLPA is prone to misclassify copy number states >3, we classified CNVs in the following categories, instead: "duplications", "homozygous deletions", and "heterozygous deletions"; for specific purposes, we used the combined category "deletions" including both homozygous and heterozygous deletions. Once a CNV was identified, gene copy number was usually well estimated, the overall SEs for all types of CNVs being >0.62. As expected, SP estimates remained very high (SP>0.87). PennCNV and CNVpartition performed better than QuantiSNP, the latter showing the highest rates of FP and FN callings. QuantiSNP performed especially poorly when calling homozygous deletions (SE=0.68 and SP=0.92). When the Itsara filter was used, SE estimates were significantly decreased to values of 0.05, 0.07, and 0.08 for CNVpartition, PennCNV, and QuantiSNP, respectively; SP increased up to 0.997 for all algorithms (Supp. Table S3).

Deleted: 29/09/2010

DISCUSSION

In the past few years, the genomics community has began to annotate a CNV genome wide map that provides better information on the contribution of structural genomic variation to genetic diversity in humans. SNP-array based-methods have allowed their association with disease susceptibility. However, the tools to carry out this task are still relatively rudimentary and the approach applied until now has mainly been based on reporting and validating individual CNVs located in candidate genes rather than assessing disease risk using genome wide analyses. This is primarily because of issues related to the accuracy of the available CNV calling algorithms. Which is, then, the most suitable method to identify CNVs for association studies using data from SNP-arrays?

The early comparisons have focused on evaluations using simulations or data from a few HapMap or CEPH samples (Kidd et al., 2008; Korbel et al., 2007; Redon et al., 2006; Winchester et al., 2009). Here we provide, for the first time, a direct comparison of the accuracy (reliability and validity) of 3 CNV calling algorithms (PennCNV, QuantiSNP, and CNVpartition) using MLPA as a gold standard and therefore eliminating some of the concerns for the validity when using simulation or resequencing data. We also investigated a more stable platform, Illumina Infinium 1M array that may not suffer from the same clustering biases as the former ones.

The algorithms used displayed wide variation in the number of CNV events. Overall, we conclude that the reproducibility of the algorithms is less than optimal. Our results indicate that PennCNV and QuantiSNP are more reliable in detecting CNVs than CNVpartition. Yet, the agreement achieved with these algorithms was much lower (mean KI ranged 57-65) than that observed for SNP calling (KI=99.99). Winchester et *al,* reported a moderate overlap between PennCNV and QuantiSNP, ranging from 58-78% for the NA15510 CEPH sample

Deleted: 29/09/2010

(Winchester et al., 2009). One explanation for the unsatisfactory concordance in experimental replicates for CNV detection and breakpoint identification relates to the different signal to noise tolerance for SNP genotyping and CNV assessment. While the background signal of SNP-arrays does not significantly affect SNP genotyping, it may affect CNV assessment due to the need of different normalization approaches for the latter (Curtis et al., 2009; Winchester et al., 2009).

Importantly, the three tools used performed poorly regarding their sensitivity to detect CNVs when using MLPA experimental results as the gold standard, the percentage of missed CNV ranging from 72-81%. Therefore, improved sensitivity of algorithms is a must in order to use genome wide chip data for CNV detection and disease association studies. When the analysis was restricted to concordant CNVs according to the applied algorithms, these estimated adequately gene copy number. This result supports the notion of performing a two-stage calling to increase accuracy. That is, to assess first the identification of CNVs and second, to characterize those already detected.

Another important finding of our work relates to the source of DNA. Many studies have shown that buccal cell and blood DNA provide similar calling rates for SNP. By contrast, we found that leukocyte DNA is more reliable for CNV detection and that buccal cell DNA yields a higher CNV calling rate. These findings are compatible with the idea that the abundance of bacterial DNA in buccal samples can interfere with the performance of genotyping bi-alleles as well, notably demonstrated by the higher discordance rates and lower completion rates. Furthermore, while tissue-related differences in genome architecture leading to variation in the number of CNVs may be real, other technical explanations such as DNA quality should also be considered. In the Spanish Bladder Cancer/EPICURO Study, saliva was obtained after a buccal rinse with Listerine® as a fixative. Saliva was then frozen until DNA extraction. This simple and costless procedure yielded substantial amounts of DNA and

Deleted: 29/09/2010

allowed accurate SNP genotyping using TaqMan assays as well as Illumina technology. For the latter, the calling agreement for leukocyte and buccal DNA was 99.99%. In the absence of other studies providing similar information, caution is needed when analyzing buccal cell DNA and new methodological studies specifically addressing these issues are needed.

Select commercial SNP-array platforms have included monomorphic probes to improve coverage of CNV analyses. We have analyzed whether monomorphic and polymorphic probes performed differently in assessing CNV. Surprisingly, we observed that, regardless of the algorithm used, CNVs showing discordance between duplicates contained a higher proportion of monomorphic probes than CNVs that were concordant. The difference was greater for QuantiSNP. Hence, our findings indicate that polymorphic probes deliver more robust information than monomorphic probes, at least using the current CNV calling tools. Alternatively, it is possible that monomorphic probes may concentrate in a small number of large CNVs being difficult to call since they are not homogenously distributed across the genome and are placed in those regions suspected of harbouring CN changes (Iafrate et al., 2004; Redon et al., 2006). Nevertheless, there is no evidence that CNVs in these regions are larger that those elsewhere.

Despite the limitations described above, SNP-arrays offer important advantages over other techniques to assess CNV at a genome wide level, including the possibility of analyzing a large number of samples because of their relatively low cost and the small amount of DNA required. CNV detection largely depends on the coverage of the platform. The low reliability that we have observed may be partially due to the fact that the localization of the CNV breakpoints depends on the position of the markers. While the Illumina 1M platform is one of the densest arrays offering a genome wide coverage, the average distance between two probes is around 3kb, larger than the smallest CNVs which are defined as having 1kb length. We have found that the average distance between surrounding probes was greater for discordant

Deleted: 29/09/2010

than for concordant CN events. This effect was stronger for PennCNV and QuantiSNP than for CNVpartition (results not shown). Small CNVs containing a small number of probes were less reliable than large CNVs that are generally called based on more probes. Furthermore, because the algorithms discard CNVs containing <3 probes, there was also an inherited disadvantage to small CNVs as compared to larger ones. By applying the filter proposed by Itsara *et al* (Itsara et al., 2009), agreement did not improve while sensitivity decreased dramatically.

The relatively poor agreement between algorithms increases the heterogeneity in CNV detection, raising the chance of false positive results in association studies. Furthermore, current algorithms lack sensitivity for CNV identification, mainly when they are small. To partially overcome this limitation, some authors have proposed to use the normalized intensity obtained from the SNP-arrays, without performing the calling, and compare its distribution at the individual probe level between cases and controls (Ionita-Laza et al., 2009; McCarroll and Altshuler, 2007). Although this strategy has not been formally evaluated and power is probably limited because of lack of biological meaning, it constitutes an alternative exploratory approach to assess association of CNVs and phenotypes. Others have suggested performing the calling and the association test simultaneously to take into account the uncertainty of the calling in the test(Barnes et al., 2008; Gonzalez et al., 2009). However, these methods require a priori definition of CNVs.

We used MLPA as the gold standard technique to estimate sensitivity and specificity of the algorithms used. MLPA is reproducible, allows the detection of small differences in gene copy number, requires low amounts of DNA, can be applied for mid-throughput studies, and has a low cost. Among its limitations are the fact that it only detects CNVs in targeted/selected genes and the results are bound to be affected by sequence polymorphisms and by the occurrence of gene copy number changes in mosaicism Despite careful probe

Deleted: 29/09/2010

design, we cannot rule out that an incomplete overlapping between probes and CNVs may contribute to the low sensitivity for CNV detection found.

The algorithms used here are those that model both LRR and BAF to assess CNV, a practice that allows the correction for bias effects and minimizes noise in the intensity measures (Yau and Holmes, 2008). In addition, these algorithms are widely applied for CNV assessment using Illumina derived data. Other CNV calling softwares are also available, such as Circular Binary Segmentation (Olshen et al., 2004), GADA originally developed for array-CGH data and adapted for SNP-array (Pique-Regi et al., 2008), DchipSNP (Lin et al., 2004), Tri Typer (Franke et al., 2008) and SCIMM (Cooper et al., 2008). However, they do not jointly incorporate both LRR and BAF information, their strengths and weaknesses have been reviewed elsewhere (Winchester et al., 2009). Nevertheless, none of them has proven to be superior to the ones used here. Winchester et *al* (Winchester et al., 2009) reported that QuantiSNP yielded a higher number of events when measuring CNV in the NA15510 CEPH sample in our study, QuantiSNP and PennCNV provided a similar mean number of CN changes that was higher than that provided by CNVpartition. Recently, Dellinger et al reported a comparison of 7 algorithms, including QuantiSNP, CNVpartition and PennCNV on simulation studies on the basis of genotyped data by Affymetrix 6.0. The authors compared sensitivity and specificity of the algorithms with CNV described in external databases (DGV, HapMap Asian and HapMap confirmed) and concluded that QuantiSNP performed better that the other algorithms (Dellinger et al., 2010).

Nevertheless, the current CNV calling algorithms do not yet provide stable, high quality calls comparable to those in common usage for SNP calling algorithms. In particular, the sensitivity is extremely low. Small/common CNVs may be less detectable because the cumulative likelihood of CNV versus normal copy for a limited number of markers suffers from a low signal-to-noise ratio. In order to improve this sensitivity in regions of known

Deleted: 29/09/2010

CNVs, some authors have proposed to look at some specific markers located within these regions and use reported deletion and duplication frequencies as prior probabilities in the calling. Such models are implemented in two widely used approaches, namely Canary (Korn et al., 2008) and PennCNV-validation packages in which they have been shown to substantially increase the sensitivity of calling CNV in these known regions. Efforts are also made to improve technologies such as CGH-arrays and (Park et al., 2010) and next generation sequencing. Hopefully, these will improve the detection of rare or novel CNVs in the near future.

In conclusion, there is a need for better assays and tools to identify CNVs at the genome wide level and test for their association with disease in large samples of cases and controls. The main current limitations are the low reliability and sensitivity. Sensitivity showed differences according to the algorithm applied and the type of change. The use of leukocyte DNA, polymorphic probes, and a high number of probes per CNV should contribute to increase reliability and PennCNV algorithm yield higher concordance rates.

The annotation of large CNVs across the genome has opened a new scenario to explore genetic variation and its association with complex diseases and traits. While a few studies support a major contribution of CNV to disease, there is an urgent need to develop and refine better techniques and algorithms to assess CNVs at a genome wide level as disease-predisposing variants.

ACKOWLEDGEMENTS

We thank Juan Cruz Cigudosa, Ramón Díaz-Uriarte, Gonzalo Gómez, Kevin Jacobs, Kristel Van Steen, and Marc Zindel for scientific sound comments and for technical support.

Deleted: 29/09/2010

We also acknowledge the support provided by Adonina Tardón, Alfredo Carrato, Consol Serra, Reina García-Closas, Josep Lloreta, Montserrat Torà, Gemma Castaño, María Salas, and Francisco Fernández, physicians, field workers, and lab technicians during the study.

This work was partially supported by the Fondo de Investigación Sanitaria, Spain (G03/174, PI061614, FI09/00205), Asociación Española Contra el Cáncer (AECC), Fundació Marató de TV3, Red Temática de Investigación Cooperativa en Cáncer (RTICC), Spain; by the Intramural Research Program of the Division of Cancer Epidemiology and Genetics, National Cancer Institute, USA; and by Egide-PHRC Picasso travel grant.

REFERENCE

Bae JS, Cheong HS, Kim JO, Lee SO, Kim EM, Lee HW, Kim S, Kim JW, Cui T, Inoue I, Shin HD. 2008. Identification of SNP markers for common CNV regions and association analysis of risk of subarachnoid aneurysmal hemorrhage in Japanese population. Biochem Biophys Res Commun 373:593-6.

Barnes C, Plagnol V, Fitzgerald T, Redon R, Marchini J, Clayton D, Hurles ME. 2008. A robust statistical method for case-control association testing with copy number variation. Nat Genet 40:1245-52.

Bassett AS, Marshall CR, Lionel AC, Chow EW, Scherer SW. 2008. Copy number variations and risk for schizophrenia in 22q11.2 deletion syndrome. Hum Mol Genet 17:4045-53.

Blauw HM, Veldink JH, van Es MA, van Vught PW, Saris CG, van der Zwaag B, Franke L, Burbach JP, Wokke JH, Ophoff RA, van den Berg LH. 2008. Copy-number variation in sporadic amyotrophic lateral sclerosis: a genome-wide screen. Lancet Neurol 7:319-26.

Cohen J. 1960. A coefficient of agreement for nominal scales. Educational and psychological measurement 20:37-46.

Colella S, Yau C, Taylor JM, Mirza G, Butler H, Clouston P, Bassett AS, Seller A, Holmes CC, 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 35:2013-25.

Conrad DF, Pinto D, Redon R, Feuk L, Gokcumen O, Zhang Y, Aerts J, Andrews TD, Barnes C, Campbell P, Fitzgerald T, Hu M, Ihm CH, Kristiansson K, Macarthur DG, Macdonald JR,

Deleted: 29/09/2010

Onyiah I, Pang AW, Robson S, Stirrups K, Valsesia A, Walter K, Wei J, Tyler-Smith C, Carter NP, Lee C, Scherer SW, Hurles ME. 2010. Origins and functional impact of copy number variation in the human genome. Nature 464:704-12.

Cooper GM, Zerr T, Kidd JM, Eichler EE, Nickerson DA. 2008. Systematic assessment of copy number variant detection via genome-wide SNP genotyping. Nat Genet 40:1199-203.

Cronin S, Blauw HM, Veldink JH, van Es MA, Ophoff RA, Bradley DG, van den Berg LH, Hardiman O. 2008. Analysis of genome-wide copy number variation in Irish and Dutch ALS populations. Hum Mol Genet 17:3392-8.

Curtis C, Lynch AG, Dunning MJ, Spiteri I, Marioni JC, Hadfield J, Chin SF, Brenton JD, Tavare S, Caldas C. 2009. The pitfalls of platform comparison: DNA copy number array technologies assessed. BMC Genomics 10:588.

Dellinger AE, Saw SM, Goh LK, Seielstad M, Young TL, Li YJ. 2010. Comparative analyses of seven algorithms for copy number variant identification from single nucleotide polymorphism arrays. Nucleic Acids Res 38:e105.

Diskin SJ, Hou C, Glessner JT, Attiyeh EF, Laudenslager M, Bosse K, Cole K, Mosse YP, Wood A, Lynch JE, Pecor K, Diamond M, Winter C, Wang K, Kim C, Geiger EA, McGrady PW, Blakemore AI, London WB, Shaikh TH, Bradfield J, Grant SF, Li H, Devoto M, Rappaport ER, Hakonarson H, Maris JM. 2009. Copy number variation at 1q21.1 associated with neuroblastoma. Nature 459:987-91.

Feuk L, Carson AR, Scherer SW. 2006. Structural variation in the human genome. Nat Rev Genet 7:85-97.

Franke L, de Kovel CG, Aulchenko YS, Trynka G, Zhernakova A, Hunt KA, Blauw HM, van den Berg LH, Ophoff R, Deloukas P, van Heel DA, Wijmenga C. 2008. Detection, imputation, and association analysis of small deletions and null alleles on oligonucleotide arrays. Am J Hum Genet 82:1316-33.

Friedman JM, Baross A, Delaney AD, Ally A, Arbour L, Armstrong L, Asano J, Bailey DK, Barber S, Birch P, Brown-John M, Cao M, Chan S, Charest DL, Farnoud N, Fernandes N, Flibotte S, Go A, Gibson WT, Holt RA, Jones SJ, Kennedy GC, Krzywinski M, Langlois S, Li HI, McGillivray BC, Nayar T, Pugh TJ, Rajcan-Separovic E, Schein JE, Schnerch A, Siddiqui A, Van Allen MI, Wilson G, Yong SL, Zahir F, Eydoux P, Marra MA. 2006.

Deleted: 29/09/2010

Oligonucleotide microarray analysis of genomic imbalance in children with mental retardation. Am J Hum Genet 79:500-13.

Garcia-Closas M, Malats N, Silverman D, Dosemeci M, Kogevinas M, Hein DW, Tardon A, Serra C, Carrato A, Garciia-Closas R, Lloreta J, Castano-Vinyals G, Yeager M, Welch R, Chanock S, Chatterjee N, Wacholder S, Samanic C, Tora M, Fernandez F, Real FX, Rothman N. 2005. NAT2 slow acetylation, GSTM1 null genotype, and risk of bladder cancer: Results from the Spanish Bladder Cancer Study and meta-analyses. Lancet 366:649-659.

Glessner JT, Wang K, Cai G, Korvatska O, Kim CE, Wood S, Zhang H, Estes A, Brune CW, Bradfield JP, Imielinski M, Frackelton EC, Reichert J, Crawford EL, Munson J, Sleiman PM, Chiavacci R, Annaiah K, Thomas K, Hou C, Glaberson W, Flory J, Otieno F, Garris M, Soorya L, Klei L, Piven J, Meyer KJ, Anagnostou E, Sakurai T, Game RM, Rudd DS, Zurawiecki D, McDougle CJ, Davis LK, Miller J, Posey DJ, Michaels S, Kolevzon A, Silverman JM, Bernier R, Levy SE, Schultz RT, Dawson G, Owley T, McMahon WM, Wassink TH, Sweeney JA, Nurnberger JI, Coon H, Sutcliffe JS, Minshew NJ, Grant SF, Bucan M, Cook EH, Buxbaum JD, Devlin B, Schellenberg GD, Hakonarson H. 2009. Autism genome-wide copy number variation reveals ubiquitin and neuronal genes. Nature 459:569-73.

Gonzalez JR, Subirana I, Escaramis G, Peraza S, Caceres A, Estivill X, Armengol L. 2009. Accounting for uncertainty when assessing association between copy number and disease: a latent class model. BMC Bioinformatics 10:172.

Greenway SC, Pereira AC, Lin JC, DePalma SR, Israel SJ, Mesquita SM, Ergul E, Conta JH, Korn JM, McCarroll SA, Gorham JM, Gabriel S, Altshuler DM, Quintanilla-Dieck Mde L, Artunduaga MA, Eavey RD, Plenge RM, Shadick NA, Weinblatt ME, De Jager PL, Hafler DA, Breitbart RE, Seidman JG, Seidman CE. 2009. De novo copy number variants identify new genes and loci in isolated sporadic tetralogy of Fallot. Nat Genet 41:931-5.

Iafrate AJ, Feuk L, Rivera MN, Listewnik ML, Donahoe PK, Qi Y, Scherer SW, Lee C. 2004. Detection of large-scale variation in the human genome. Nat Genet 36:949-51.

InternationalSchizophreniaConsortium. 2008. Rare chromosomal deletions and duplications increase risk of schizophrenia. Nature 455:237-41.

Ionita-Laza I, Perry GH, Raby BA, Klanderman B, Lee C, Laird NM, Weiss ST, Lange C. 2008. On the analysis of copy-number variations in genome-wide association studies: a translation of the family-based association test. Genet Epidemiol 32:273-84.

Deleted: 29/09/2010

Ionita-Laza I, Rogers AJ, Lange C, Raby BA, Lee C. 2009. Genetic association analysis of copy-number variation (CNV) in human disease pathogenesis. Genomics 93:22-6.

Itsara A, Cooper GM, Baker C, Girirajan S, Li J, Absher D, Krauss RM, Myers RM, Ridker PM, Chasman DI, Mefford H, Ying P, Nickerson DA, Eichler EE. 2009. Population analysis of large copy number variants and hotspots of human genetic disease. Am J Hum Genet 84:148-61.

Kathiresan S, Voight BF, Purcell S, Musunuru K, Ardissino D, Mannucci PM, Anand S, Engert JC, Samani NJ, Schunkert H, Erdmann J, Reilly MP, Rader DJ, Morgan T, Spertus JA, Stoll M, Girelli D, McKeown PP, Patterson CC, Siscovick DS, O'Donnell CJ, Elosua R, Peltonen L, Salomaa V, Schwartz SM, Melander O, Altshuler D, Ardissino D, Merlini PA, Berzuini C, Bernardinelli L, Peyvandi F, Tubaro M, Celli P, Ferrario M, Fetiveau R, Marziliano N, Casari G, Galli M, Ribichini F, Rossi M, Bernardi F, Zonzin P, Piazza A, Mannucci PM, Schwartz SM, Siscovick DS, Yee J, Friedlander Y, Elosua R, Marrugat J, Lucas G, Subirana I, Sala J, Ramos R, Kathiresan S, Meigs JB, Williams G, Nathan DM, MacRae CA, O'Donnell CJ, Salomaa V, Havulinna AS, Peltonen L, Melander O, Berglund G, Voight BF, Kathiresan S, Hirschhorn JN, Asselta R, Duga S, Spreafico M, Musunuru K, Daly MJ, Purcell S, Voight BF, Purcell S, Nemesh J, Korn JM, McCarroll SA, Schwartz SM, Yee J, Kathiresan S, Lucas G, Subirana I, Elosua R, Surti A, Guiducci C, Gianniny L, Mirel D, Parkin M, Burtt N, Gabriel SB, Samani NJ, Thompson JR, Braund PS, Wright BJ, Balmforth AJ, Ball SG, Hall AS, Schunkert H, Erdmann J, Linsel-Nitschke P, Lieb W, Ziegler A, Konig I, Hengstenberg C, Fischer M, Stark K, Grosshennig A, Preuss M, Wichmann HE, Schreiber S, Schunkert H, Samani NJ, Erdmann J, Ouwehand W, Hengstenberg C, Deloukas P, Scholz M, Cambien F, Reilly MP, Li M, Chen Z, Wilensky R, Matthai W, Qasim A, Hakonarson HH, Devaney J, Burnett MS, Pichard AD, Kent KM, Satler L, Lindsay JM, Waksman R, Epstein SE, Rader DJ, Scheffold T, Berger K, Stoll M, Huge A, Girelli D, Martinelli N, Olivieri O, Corrocher R, Morgan T, Spertus JA, McKeown P, Patterson CC, Schunkert H, Erdmann E, Linsel-Nitschke P, Lieb W, Ziegler A, Konig IR, Hengstenberg C, Fischer M, Stark K, Grosshennig A, Preuss M, Wichmann HE, Schreiber S, Holm H, Thorleifsson G, Thorsteinsdottir U, Stefansson K, Engert JC, Do R, Xie C, Anand S, Kathiresan S, Ardissino D, Mannucci PM, Siscovick D, O'Donnell CJ, Samani NJ, Melander O, Elosua R, Peltonen L, Salomaa V, Schwartz SM, Altshuler D. 2009. Genome-wide association of early-onset myocardial infarction with single nucleotide polymorphisms and copy number variants. Nat Genet 41:334-41.

Deleted: 29/09/2010

Kidd JM, Cooper GM, Donahue WF, Hayden HS, Sampas N, Graves T, Hansen N, Teague B, Alkan C, Antonacci F, Haugen E, Zerr T, Yamada NA, Tsang P, Newman TL, Tuzun E, Cheng Z, Ebling HM, Tusneem N, David R, Gillett W, Phelps KA, Weaver M, Saranga D, Brand A, Tao W, Gustafson E, McKernan K, Chen L, Malig M, Smith JD, Korn JM, McCarroll SA, Altshuler DA, Peiffer DA, Dorschner M, Stamatoyannopoulos J, Schwartz D, Nickerson DA, Mullikin JC, Wilson RK, Bruhn L, Olson MV, Kaul R, Smith DR, Eichler EE. 2008. Mapping and sequencing of structural variation from eight human genomes. Nature 453:56-64.

Korbel JO, Urban AE, Grubert F, Du J, Royce TE, Starr P, Zhong G, Emanuel BS, Weissman SM, Snyder M, Gerstein MB. 2007. Systematic prediction and validation of breakpoints associated with copy-number variants in the human genome. Proc Natl Acad Sci U S A 104:10110-5.

Korn JM, Kuruvilla FG, McCarroll SA, Wysoker A, Nemesh J, Cawley S, Hubbell E, Veitch J, Collins PJ, Darvishi K, Lee C, Nizzari MM, Gabriel SB, Purcell S, Daly MJ, Altshuler D. 2008. Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. Nat Genet 40:1253-60.

Lin M, Wei LJ, Sellers WR, Lieberfarb M, Wong WH, Li C. 2004. dChipSNP: significance curve and clustering of SNP-array-based loss-of-heterozygosity data. Bioinformatics 20:1233-40.

Liu W, Sun J, Li G, Zhu Y, Zhang S, Kim ST, Sun J, Wiklund F, Wiley K, Isaacs SD, Stattin P, Xu J, Duggan D, Carpten JD, Isaacs WB, Gronberg H, Zheng SL, Chang BL. 2009. Association of a germ-line copy number variation at 2p24.3 and risk for aggressive prostate cancer. Cancer Res 69:2176-9.

Marshall CR, Noor A, Vincent JB, Lionel AC, Feuk L, Skaug J, Shago M, Moessner R, Pinto D, Ren Y, Thiruvahindrapduram B, Fiebig A, Schreiber S, Friedman J, Ketelaars CE, Vos YJ, Ficicioglu C, Kirkpatrick S, Nicolson R, Sloman L, Summers A, Gibbons CA, Teebi A, Chitayat D, Weksberg R, Thompson A, Vardy C, Crosbie V, Luscombe S, Baatjes R, Zwaigenbaum L, Roberts W, Fernandez B, Szatmari P, Scherer SW. 2008. Structural variation of chromosomes in autism spectrum disorder. Am J Hum Genet 82:477-88.

Matarin M, Simon-Sanchez J, Fung HC, Scholz S, Gibbs JR, Hernandez DG, Crews C, Britton A, Wavrant De Vrieze F, Brott TG, Brown RD, Jr., Worrall BB, Silliman S, Case LD,

Deleted: 29/09/2010

Hardy JA, Rich SS, Meschia JF, Singleton AB. 2008. Structural genomic variation in ischemic stroke. Neurogenetics 9:101-8.

McCarroll SA, Altshuler DM. 2007. Copy-number variation and association studies of human disease. Nat Genet 39:S37-42.

Need AC, Ge D, Weale ME, Maia J, Feng S, Heinzen EL, Shianna KV, Yoon W, Kasperaviciute D, Gennarelli M, Strittmatter WJ, Bonvicini C, Rossi G, Jayathilake K, Cola PA, McEvoy JP, Keefe RS, Fisher EM, St Jean PL, Giegling I, Hartmann AM, Moller HJ, Ruppert A, Fraser G, Crombie C, Middleton LT, St Clair D, Roses AD, Muglia P, Francks C, Rujescu D, Meltzer HY, Goldstein DB. 2009. A genome-wide investigation of SNPs and CNVs in schizophrenia. PLoS Genet 5:e1000373.

Olshen AB, Venkatraman ES, Lucito R, Wigler M. 2004. Circular binary segmentation for the analysis of array-based DNA copy number data. Biostatistics 5:557-72.

Park H, Kim JI, Ju YS, Gokcumen O, Mills RE, Kim S, Lee S, Suh D, Hong D, Kang HP, Yoo YJ, Shin JY, Kim HJ, Yavartanoo M, Chang YW, Ha JS, Chong W, Hwang GR, Darvishi K, Kim H, Yang SJ, Yang KS, Kim H, Hurles ME, Scherer SW, Carter NP, Tyler-Smith C, Lee C, Seo JS. 2010. Discovery of common Asian copy number variants using integrated high-resolution array CGH and massively parallel DNA sequencing. Nat Genet 42:400-5.

Pique-Regi R, Monso-Varona J, Ortega A, Seeger RC, Triche TJ, Asgharzadeh S. 2008. Sparse representation and Bayesian detection of genome copy number alterations from microarray data. Bioinformatics 24:309-18.

Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, Fiegler H, Shapero MH, Carson AR, Chen W, Cho EK, Dallaire S, Freeman JL, Gonzalez JR, Gratacos M, Huang J, Kalaitzopoulos D, Komura D, MacDonald JR, Marshall CR, Mei R, Montgomery L, Nishimura K, Okamura K, Shen F, Somerville MJ, Tchinda J, Valsesia A, Woodwark C, Yang F, Zhang J, Zerjal T, Zhang J, Armengol L, Conrad DF, Estivill X, Tyler-Smith C, Carter NP, Aburatani H, Lee C, Jones KW, Scherer SW, Hurles ME. 2006. Global variation in copy number in the human genome. Nature 444:444-54.

Rodriguez-Santiago B, Brunet A, Sobrino B, Serra-Juhe C, Flores R, Armengol L, Vilella E, Gabau E, Guitart M, Guillamat R, Martorell L, Valero J, Gutierrez-Zotes A, Labad A, Carracedo A, Estivill X, Perez-Jurado LA. 2009. Association of common copy number

Deleted: 29/09/2010

variants at the glutathione S-transferase genes and rare novel genomic changes with schizophrenia. Mol Psychiatry.

Schouten JP, McElgunn CJ, Waaijer R, Zwijnenburg D, Diepvens F, Pals G. 2002. Relative quantification of 40 nucleic acid sequences by multiplex ligation-dependent probe amplification. Nucleic Acids Res 30:e57.

Sha BY, Yang TL, Zhao LJ, Chen XD, Guo Y, Chen Y, Pan F, Zhang ZX, Dong SS, Xu XH, Deng HW. 2009. Genome-wide association study suggested copy number variation may be associated with body mass index in the Chinese population. J Hum Genet 54:199-202.

Simon-Sanchez J, Scholz S, Matarin Mdel M, Fung HC, Hernandez D, Gibbs JR, Britton A, Hardy J, Singleton A. 2008. Genomewide SNP assay reveals mutations underlying Parkinson disease. Hum Mutat 29:315-22.

Stefansson H, Rujescu D, Cichon S, Pietilainen OP, Ingason A, Steinberg S, Fossdal R, Sigurdsson E, Sigmundsson T, Buizer-Voskamp JE, Hansen T, Jakobsen KD, Muglia P, Francks C, Matthews PM, Gylfason A, Halldorsson BV, Gudbjartsson D, Thorgeirsson TE, Sigurdsson A, Jonasdottir A, Jonasdottir A, Bjornsson A, Mattiasdottir S, Blondal T, Haraldsson M, Magnusdottir BB, Giegling I, Moller HJ, Hartmann A, Shianna KV, Ge D, Need AC, Crombie C, Fraser G, Walker N, Lonnqvist J, Suvisaari J, Tuulio-Henriksson A, Paunio T, Toulopoulou T, Bramon E, Di Forti M, Murray R, Ruggeri M, Vassos E, Tosato S, Walshe M, Li T, Vasilescu C, Muhleisen TW, Wang AG, Ullum H, Djurovic S, Melle I, Olesen J, Kiemeney LA, Franke B, Sabatti C, Freimer NB, Gulcher JR, Thorsteinsdottir U, Kong A, Andreassen OA, Ophoff RA, Georgi A, Rietschel M, Werge T, Petursson H, Goldstein DB, Nothen MM, Peltonen L, Collier DA, St Clair D, Stefansson K, Kahn RS, Linszen DH, van Os J, Wiersma D, Bruggeman R, Cahn W, de Haan L, Krabbendam L, Myin-Germeys I. 2008. Large recurrent microdeletions associated with schizophrenia. Nature 455:232-6.

Stranger BE, Forrest MS, Dunning M, Ingle CE, Beazley C, Thorne N, Redon R, Bird CP, de Grassi A, Lee C, Tyler-Smith C, Carter N, Scherer SW, Tavare S, Deloukas P, Hurles ME, Dermitzakis ET. 2007. Relative impact of nucleotide and copy number variation on gene expression phenotypes. Science 315:848-53.

Walsh T, McClellan JM, McCarthy SE, Addington AM, Pierce SB, Cooper GM, Nord AS, Kusenda M, Malhotra D, Bhandari A, Stray SM, Rippey CF, Roccanova P, Makarov V, Lakshmi B, Findling RL, Sikich L, Stromberg T, Merriman B, Gogtay N, Butler P, Eckstrand

Deleted: 29/09/2010

K, Noory L, Gochman P, Long R, Chen Z, Davis S, Baker C, Eichler EE, Meltzer PS, Nelson SF, Singleton AB, Lee MK, Rapoport JL, King MC, Sebat J. 2008. Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia. Science 320:539-43.

Wang K, Li M, Hadley D, Liu R, Glessner J, Grant SF, Hakonarson H, 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 17:1665-74.

Weiss LA, Shen Y, Korn JM, Arking DE, Miller DT, Fossdal R, Saemundsen E, Stefansson H, Ferreira MA, Green T, Platt OS, Ruderfer DM, Walsh CA, Altshuler D, Chakravarti A, Tanzi RE, Stefansson K, Santangelo SL, Gusella JF, Sklar P, Wu BL, Daly MJ. 2008. Association between microdeletion and microduplication at 16p11.2 and autism. N Engl J Med 358:667-75.

Winchester L, Yau C, Ragoussis J. 2009. Comparing CNV detection methods for SNP arrays. Brief Funct Genomic Proteomic 8:353-66.

Xu B, Roos JL, Levy S, van Rensburg EJ, Gogos JA, Karayiorgou M. 2008. Strong association of de novo copy number mutations with sporadic schizophrenia. Nat Genet 40:880-5.

Yang TL, Chen XD, Guo Y, Lei SF, Wang JT, Zhou Q, Pan F, Chen Y, Zhang ZX, Dong SS, Xu XH, Yan H, Liu X, Qiu C, Zhu XZ, Chen T, Li M, Zhang H, Zhang L, Drees BM, Hamilton JJ, Papasian CJ, Recker RR, Song XP, Cheng J, Deng HW. 2008. Genome-wide copy-number-variation study identified a susceptibility gene, UGT2B17, for osteoporosis. Am J Hum Genet 83:663-74.

Yau C, Holmes CC. 2008. CNV discovery using SNP genotyping arrays. Cytogenet Genome Res 123:307-12.

FIGURE LEGENDS

Figure 1. Box plots of the distribution of kappa index estimates comparing duplicated pairs for A) the SNP callings, B) the detection of CNVs according to the different algorithms, and C) the number of copies assigned by the different algorithms in the regions where a CNV was detected.

Deleted: 29/09/2010

Figure 2. Box plots of the distribution of kappa indexes comparing the callings on duplicated samples by the different algorithms depending on the source of DNA.

Figure 3. Average Kappa Index for the agreement in detecting CNVs (first row) and median number of CNVs across the 92 individuals (second row) for each algorithm while filtering the called CNVs according the number of probes in the CNV (first column) and the length of the CNV (second column).

Figure 4. Sensitivity (SE) and Specificity (SP) estimates for the presence and for the type-specific CNV according to each algorithm.

Deleted: 29/09/2010

Table 1: Median number of CNVs detected in the 92 individuals included in this study. The results are displayed according to the algorithm applied and the source of DNA. One of the replicates was randomly selected to obtain these estimates.

Algorithm	Source of DNA	Number of Copies				Total
		0	1	3	4	
CNVpartition	**All**	10	10	8	1	**28**
	Blood	8	10	6	1	**25**
	Saliva	14	12	13	2	**51**
PennCNV	**All**	5	31.5	23	2	**58.5**
	Blood	5	28	19	1	**53**
	Saliva	6	40	32	2	**101**
QuantiSNP	**All**	18.5	24	9	2	**56**
	Blood	18	22	8	1	**51**
	Saliva	20	30	12	4	**90**

Table 2: Distribution of probes in the two agreement categories (disagree and agree on calling CNV) for each of the algorithms. Results are displayed for all (All), monomorphic (Mono) and polymorphic (Poly) probes.

	CNVpartition			PennCNV			QuantiSNP		
	All	**Mono**	**Poly**	**All**	**Mono**	**Poly**	**All**	**Mono**	**Poly**
Disagree	1085	113	972	1165	89	1076	2044	385	1659
	100 %	10.43%	89.57%	1	7.63%	92.37%	100 %	18.83 %	81.17 %
Agree in calling CNV	657	52	605	1196	50	1146	2547	255	2292
	100 %	7.97%	92.03%	1	4.16%	95.84%	100 %	10.00 %	90.00 %
ratio Disagree/Agree		2.17	1.61		1.78	0.94		1.51	0.72

Supplementary Table S1. Number of Individuals, assays, and pairs analyzed (before CNV criteria) and considered in the accuracy study (after CNV criteria) and according to DNA source.

Overall			Blood			Saliva			Blood / Saliva		
Individuals	**Assays**	**Pairs**	**Individuals**	**Assays**	**Pairs**	**Individuals**	**Assays**	**Pairs**	**Individuals**	**Assays**	**Pairs**
Before CNV criteria											
141	299	178	71	142	71	66	146	97	4	11	10
127 dup			71 dup			55 dup			1 dup	5 Blood	1 B/B
11 trip						8 trip			3 trip	6 Saliva	2 S/S
3 quadrip						3 quadrip					7 B/S
After CNV criteria											
92	186	96	63	126	63	29	60	33	-	-	-
90 dup			63 dup			27 dup					
2 trip						2 trip					

Assays are count by summing all duplicate, triplicate and quadruplicate samples
Pairs refer to the by-two comparisons provided by duplicate (2), triplicate (3) and quadruplicate (6) samples.

Supplementary Table S2. MLPA probes considered in the MLPA analysis.

Probe	Chromosome	Band	Start	End
SKI	1	1p36.33	2,150,969	2,151,029
IL1B	2	2q13	113,306,801	113,306,852
A_14_P103008	2	2q37.3	242,228,984	242,229,042
PLCD1	3	3p22.3	38,026,650	38,026,709
Chr3_46771035	3	3p21.31	46,781,196	46,781,253
Chr4_69231671	4	4q13.2	69,109,638	69,109,698
PCDHA9	5	5q31.1	140,208,267	140,208,335
DOM3Z	6	6p21.32	32,047,183	32,047,228
HLA-DRB5	6	6p21.32	32,593,310	32,593,379
FZD9	7	7q11.23	72,294,840	72,294,901
Chr8_39356595	8	8p11.23	39,401,744	39,401,802
RXRa	9	9q34.2	136,453,357	136,453,414
NOTCH1	9	9q34.3	138,523,724	138,523,783
PPYR1	10	10q11.22	46,507,740	46,507,809
ADAM8	10	10q26.3	134,933,411	134,933,468
HRAS	11	11p15.5	523,758	523,813
A_14_P114204	11	11q13.1	66,952,984	66,953,039
OR4K2	14	14q11.2	19,414,387	19,414,452
Chr16_32481309	16	16p11.2	32,516,918	32,516,977
chr17_415_A	17	17q21.31	41,539,152	41,539,211
chr17_42061812_42110026_B	17	17q21.31	41,889,427	41,889,486
NSF	17	17q21.32	42,166,492	42,166,551
STK11	19	19p13.3	1,171,375	1,171,442
ENm007_1	19	19q13.42	59,427,206	59,427,263
ENm007_2	19	19q13.42	59,968,534	59,968,593
A_14_P105195	20	20q11.21	30,111,471	30,111,530
GSTT1	22	22q11.23	22,706,190	22,706,250
Chr22_22690592	22	22q11.23	22,709,442	22,709,496
Chr22_Pop_1	22	22q13.1	37,684,655	37,684,714

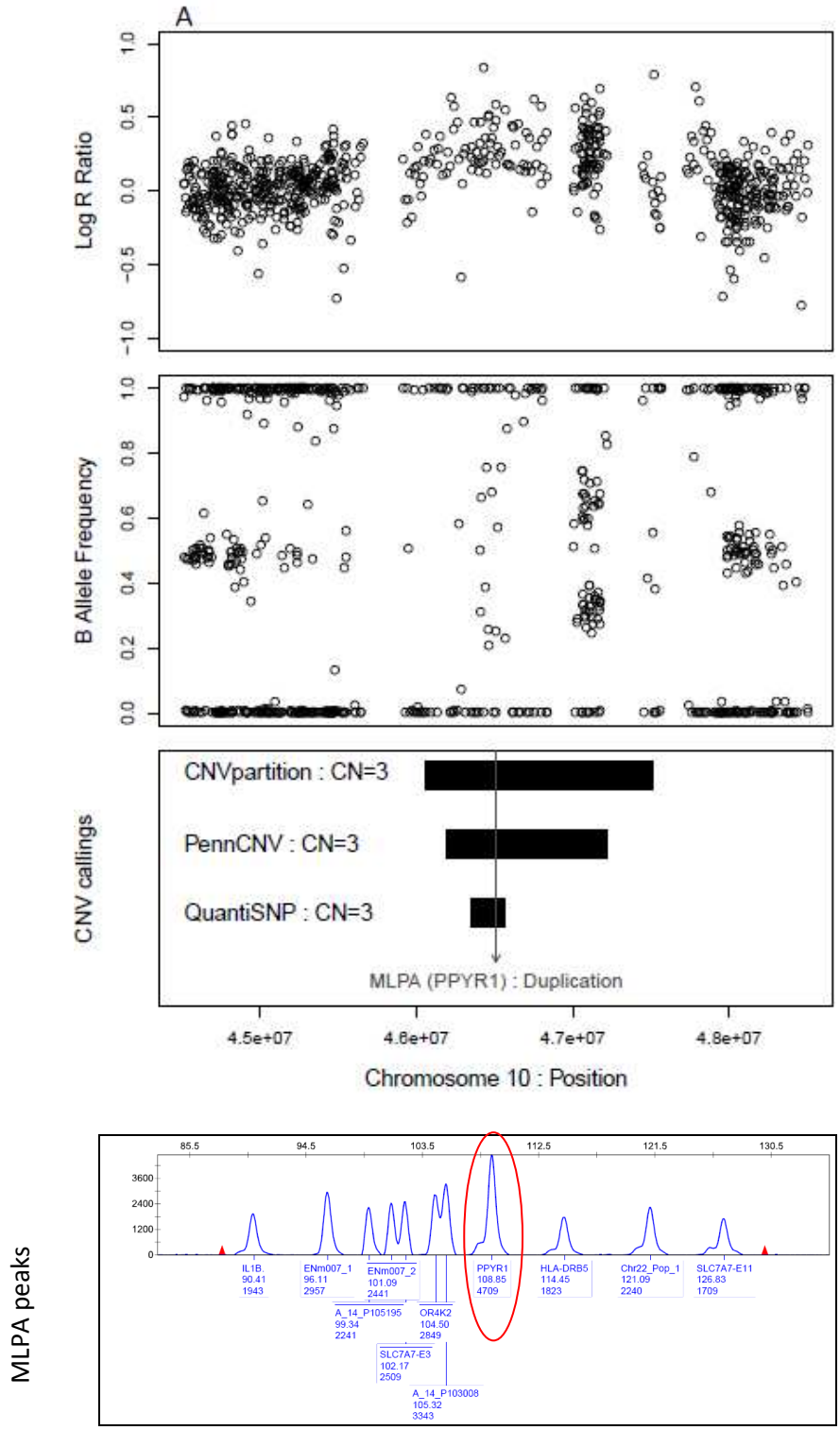
Supplementary Table S3. Validity estimates for blood samples comparing the calling results with those obtained using MLPA as a reference. The estimates and their 95% confidence intervals (CI) for sensitivity (SE), specificity (SP), positive predictive value (VPP) and negative predictive value (VPN) are displayed according to the algorithms and the different types of aberrations with and without filtering using the Itsara *et al.* criteria.

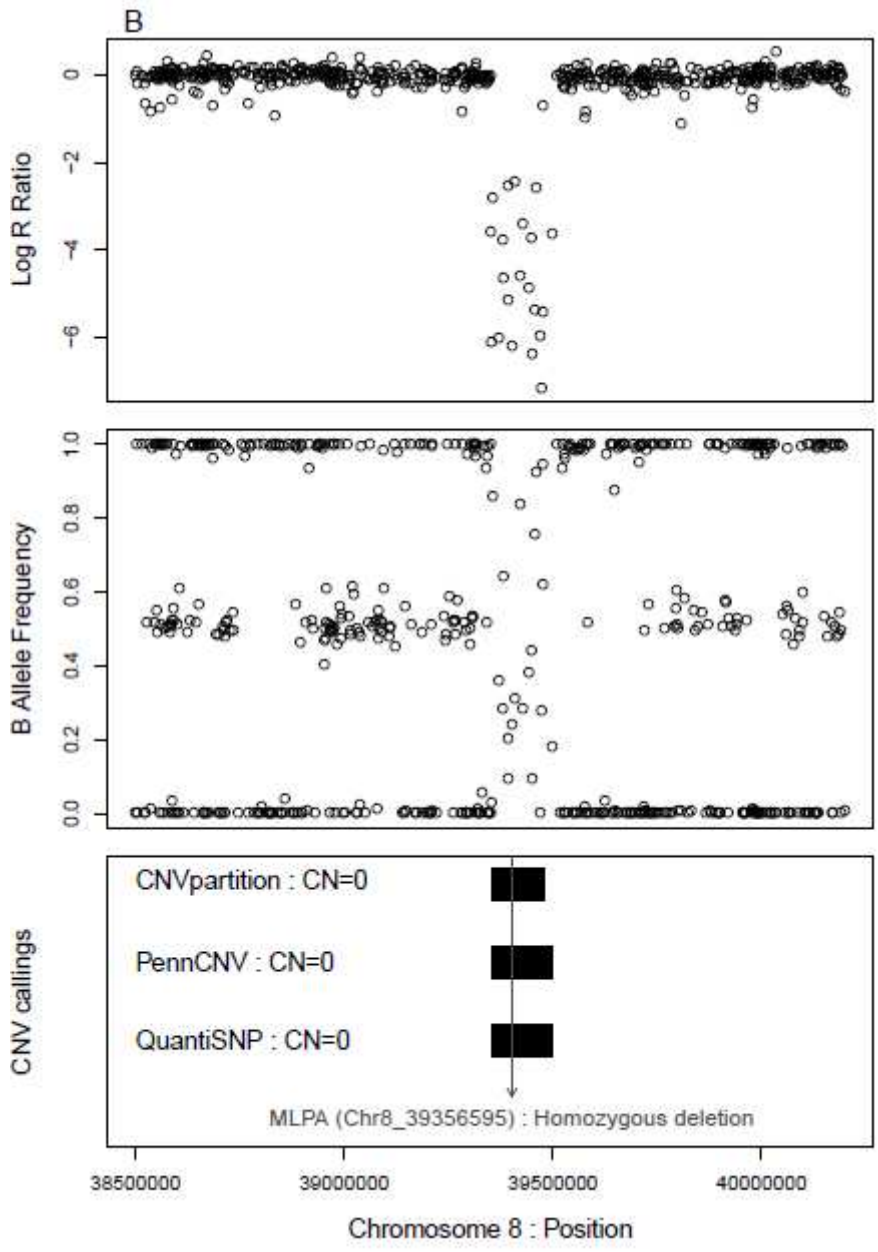
			CNVpartition				PennCNV				QuantiSNP			
			No filter		Itsara et al. filter		No filter		Itsara et al. filter		No filter		Itsara et al. filter	
Steps	**CNV type**		**Est.**	**95% CI**	**Est.**	**95% CI**	**Est.**	**95% CI**	**Est.**	**95% CI**	**Est.**	**95% CI**	**Est.**	**95% CI**
1	**CNV**	**SE**	0.19	[0.14 - 0.23]	0.05	[0.03 - 0.08]	0.23	[0.18 - 0.28]	0.07	[0.05 - 0.11]	0.28	[0.23 - 0.33]	0.08	[0.05 - 0.11]
		SP	0.99	[0.98 - 1.00]	1.00	[1.00 - 1.00]	0.98	[0.97 - 0.99]	1.00	[0.99 - 1.00]	0.97	[0.96 - 0.98]	1.00	[0.99 - 1.00]
		VPP	0.83	[0.73 - 0.91]	0.95	[0.75 - 1.00]	0.76	[0.66 - 0.85]	0.86	[0.68 - 0.96]	0.71	[0.62 - 0.79]	0.90	[0.73 - 0.98]
		VPN	0.83	[0.81 - 0.85]	0.80	[0.78 - 0.82]	0.84	[0.83 - 0.86]	0.81	[0.79 - 0.83]	0.86	[0.84 - 0.87]	0.81	[0.79 - 0.83]
2a	**Deletion***	**SE**	0.97	[0.86 - 1.00]	1.00	[0.73 - 1.00]	0.95	[0.84 - 0.99]	1.00	[0.79 - 1.00]	0.98	[0.91 - 1.00]	1.00	[0.81 - 1.00]
		SP	1.00	[0.79 - 1.00]	1.00	[0.09 - 1.00]	1.00	[0.83 - 1.00]	1.00	[0.09 - 1.00]	0.92	[0.73 - 0.99]	1.00	[0.01 - 1.00]
		VPP	1.00	[0.86 - 1.00]	1.00	[0.73 - 1.00]	1.00	[0.87 - 1.00]	1.00	[0.79 - 1.00]	0.97	[0.88 - 1.00]	1.00	[0.81 - 1.00]
		VPN	0.96	[0.79 - 1.00]	1.00	[0.09 - 1.00]	0.94	[0.79 - 0.99]	1.00	[0.09 - 1.00]	0.96	[0.78 - 1.00]	1.00	[0.01 - 1.00]
2b	**Homozygous deletion***	**SE**	1.00	[0.83 - 1.00]	1.00	[0.66 - 1.00]	0.86	[0.57 - 0.98]	1.00	[0.64 - 1.00]	0.68	[0.45 - 0.86]	1.00	[0.66 - 1.00]
		SP	0.94	[0.79 -	1.00	[0.42 -	1.00	[0.91 -	1.00	[0.66 -	0.92	[0.82 -	1.00	[0.68 -

				0.99]		1.00]		1.00]		1.00]		0.97]		1.00]
		VPP	0.94	[0.79 - 0.99]	1.00	[0.66 - 1.00]	1.00	[0.64 - 1.00]	1.00	[0.64 - 1.00]	0.75	[0.51 - 0.91]	1.00	[0.66 - 1.00]
		VPN	1.00	[0.83 - 1.00]	1.00	[0.42 - 1.00]	0.97	[0.88 - 1.00]	1.00	[0.66 - 1.00]	0.89	[0.78 - 0.95]	1.00	[0.68 - 1.00]
2c	**Heterozygous deletion***	**SE**	0.63	[0.24 - 0.91]	1.00	[0.28 - 1.00]	0.93	[0.76 - 0.99]	1.00	[0.62 - 1.00]	0.92	[0.78 - 0.98]	1.00	[0.66 - 1.00]
		SP	1.00	[0.9 - 1.00]	1.00	[0.7 - 1.00]	0.95	[0.84 - 0.99]	1.00	[0.68 - 1.00]	0.87	[0.74 - 0.95]	1.00	[0.68 - 1.00]
		VPP	1.00	[0.36 - 1.00]	1.00	[0.28 - 1.00]	0.93	[0.76 - 0.99]	1.00	[0.62 - 1.00]	0.85	[0.7 - 0.94]	1.00	[0.66 - 1.00]
		VPN	0.95	[0.85 - 0.99]	1.00	[0.7 - 1.00]	0.95	[0.84 - 0.99]	1.00	[0.68 - 1.00]	0.93	[0.81 - 0.99]	1.00	[0.68 - 1.00]
2d	**Duplication***	**SE**	1.00	[0.79 - 1.00]	1.00	[0.09 - 1.00]	1.00	[0.83 - 1.00]	1.00	[0.09 - 1.00]	0.92	[0.73 - 0.99]	1.00	[0.01 - 1.00]
		SP	0.97	[0.86 - 1.00]	1.00	[0.73 - 1.00]	0.95	[0.84 - 0.99]	1.00	[0.79 - 1.00]	0.98	[0.91 - 1.00]	1.00	[0.81 - 1.00]
		VPP	0.96	[0.79 - 1.00]	1.00	[0.09 - 1.00]	0.94	[0.79 - 0.99]	1.00	[0.09 - 1.00]	0.96	[0.78 - 1.00]	1.00	[0.01 - 1.00]
		VPN	1.00	[0.86 - 1.00]	1.00	[0.73 - 1.00]	1.00	[0.87 - 1.00]	1.00	[0.79 - 1.00]	0.97	[0.88 - 1.00]	1.00	[0.81 - 1.00]

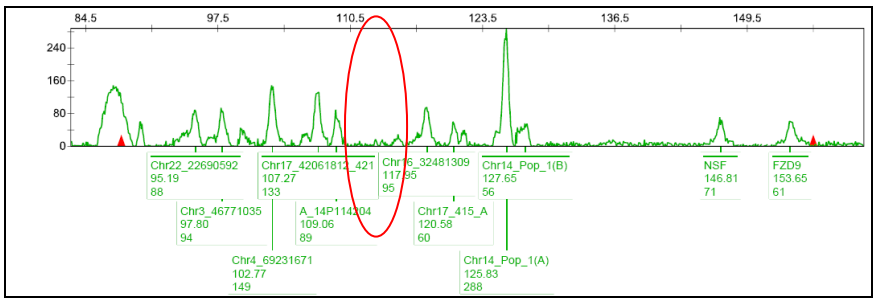
*Estimates for each CNV type were calculated only for these true positive CNVs identified in step 1.

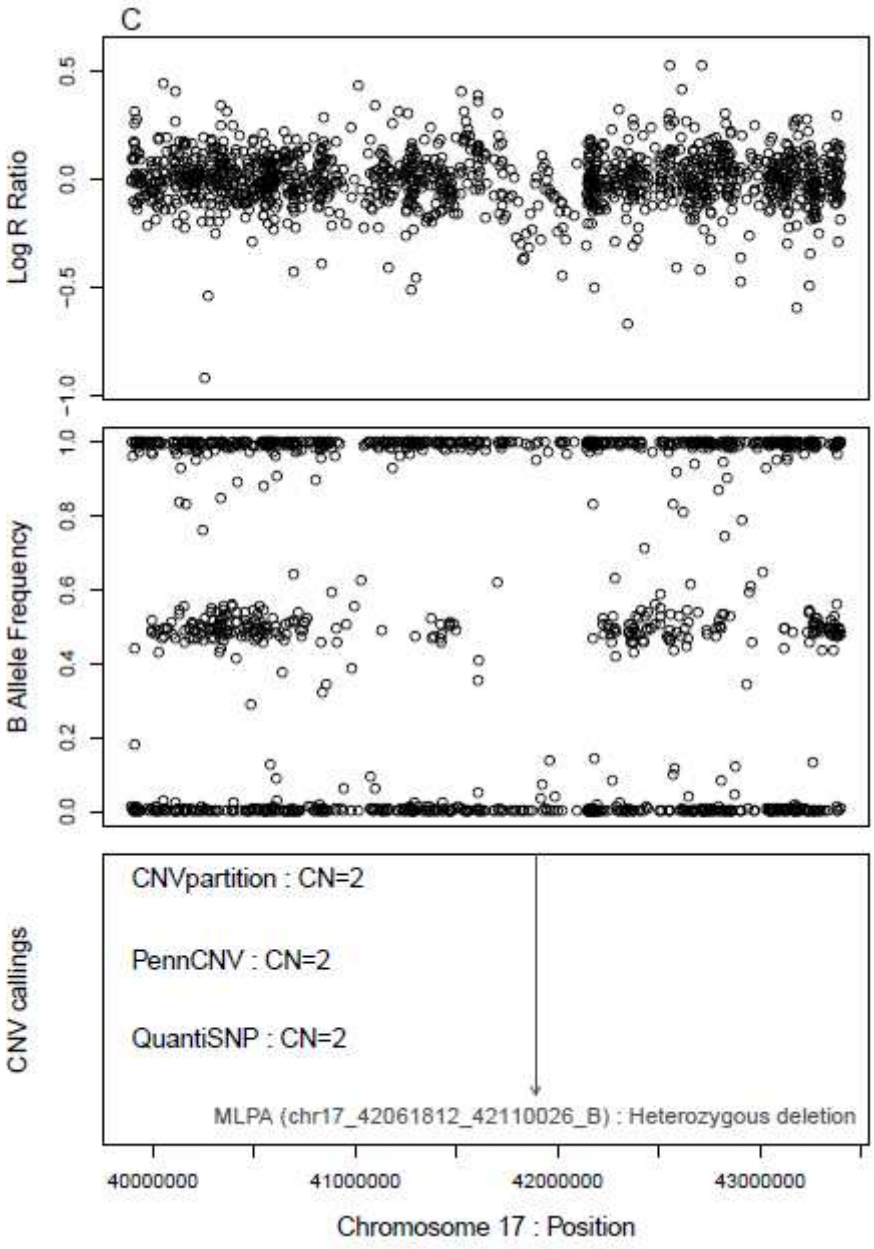
Supplementary Figure S1: Log R Ratio (LRR), B Allele Frequency (BAF), algorithm and MLPA callings and MLPA peaks for A) a true positive duplication, B) a true positive homozygous deletion, C) a false negative heterozygous deletion and D) a false positive duplication. MLPA peaks are shown for the considering individual and for various probes used for validation.



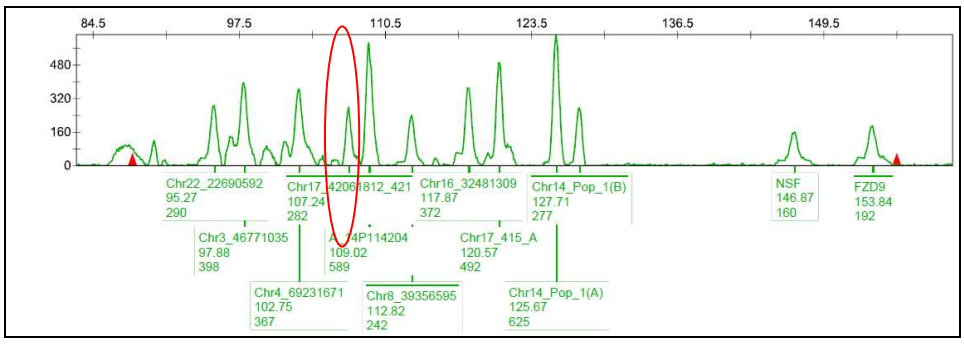

B
Log R Ratio
0
-2
-4
-6
B Allele Frequency
0.0
0.2
0.4
0.6
0.8
1.0
CNV callings
CNVpartition : CN=0
PennCNV : CN=0
QuantiSNP : CN=0
MLPA (Chr8_39356595) : Homozygous deletion
38500000
39000000
39500000
40000000
Chromosome 8 : Position

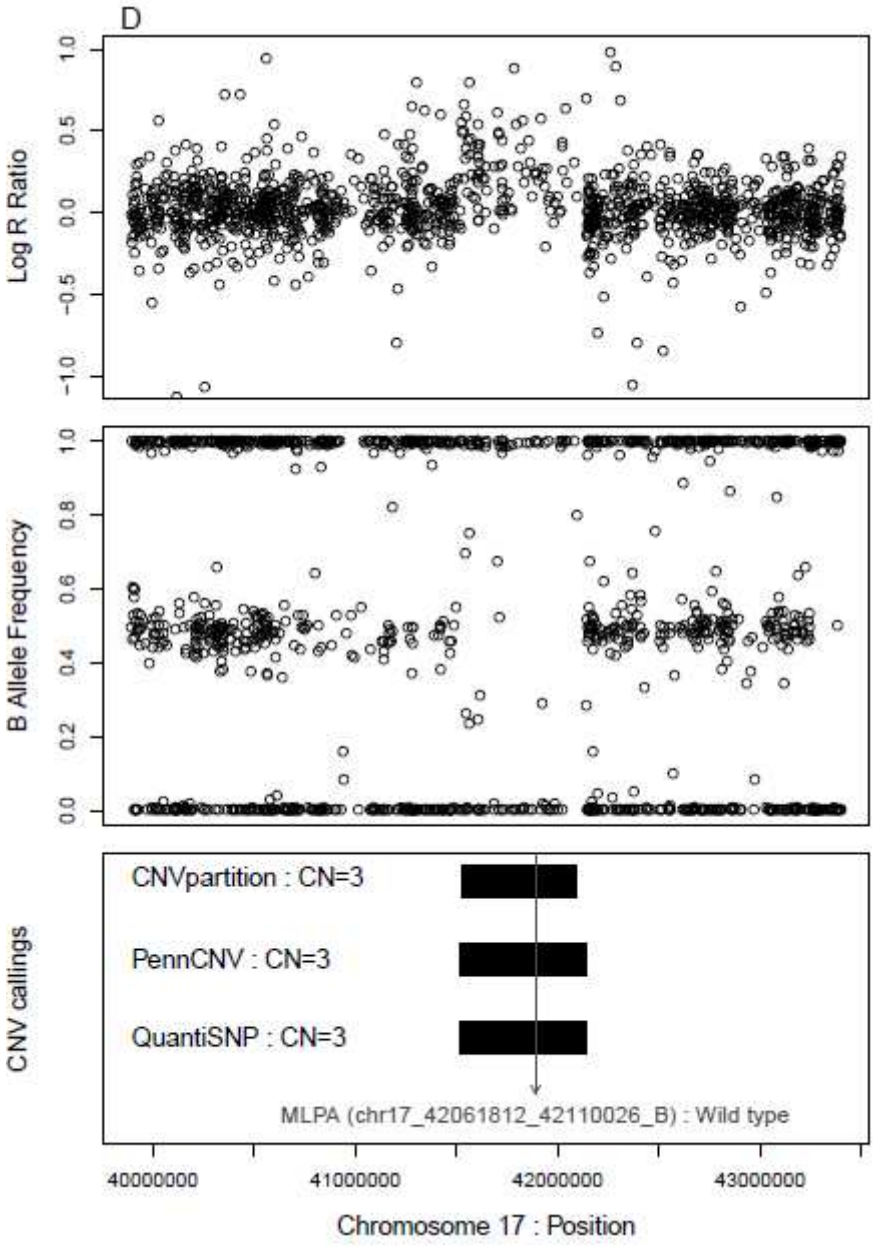
MLPA peaks


84.5
97.5
110.5
123.5
136.5
149.5
240
160
80
0
Chr22_22690592
95.19
88
Chr17_42061812_421
107.27
133
Chr16_32481309
117.95
95
Chr14_Pop_1(B)
127.65
56
NSF
146.81
71
FZD9
153.65
61
Chr3_46771035
97.80
94
A_14P114204
109.06
89
Chr17_415_A
120.58
60
Chr4_69231671
102.77
149
Chr14_Pop_1(A)
125.83
288

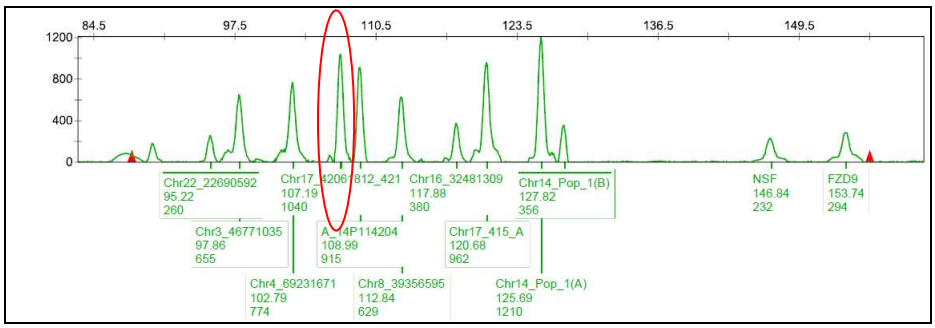


MLPA peaks

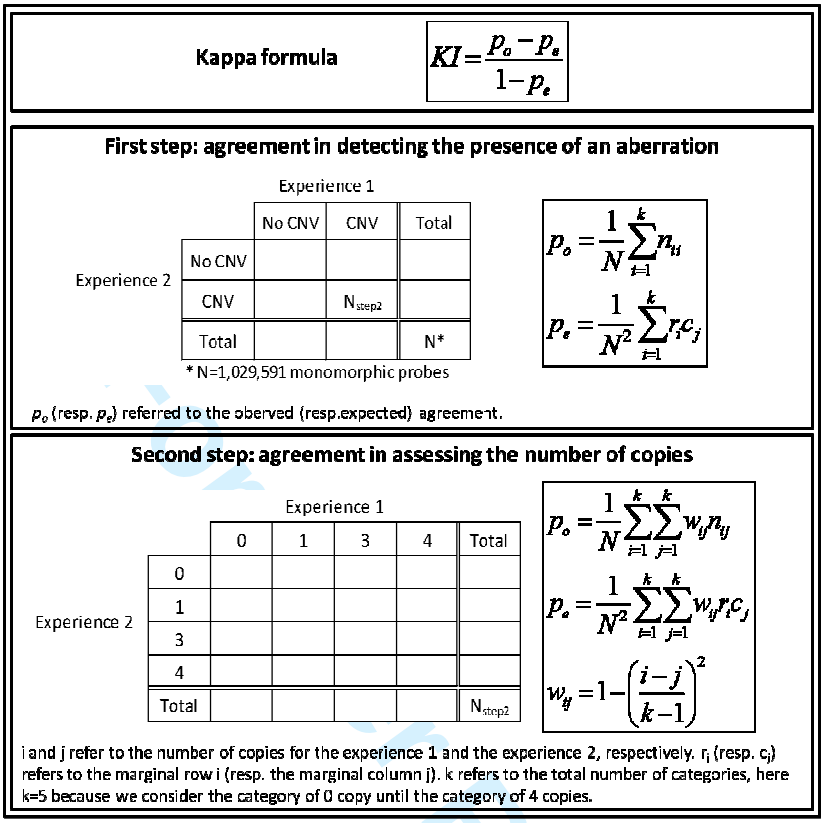



D
Log R Ratio
1.0
0.5
0.0
−0.5
−1.0
B Allele Frequency
1.0
0.8
0.6
0.4
0.2
0.0
CNV callings
CNVpartition : CN=3
PennCNV : CN=3
QuantiSNP : CN=3
MLPA (chr17_42061812_42110026_B) : Wild type
40000000
41000000
42000000
43000000
Chromosome 17 : Position

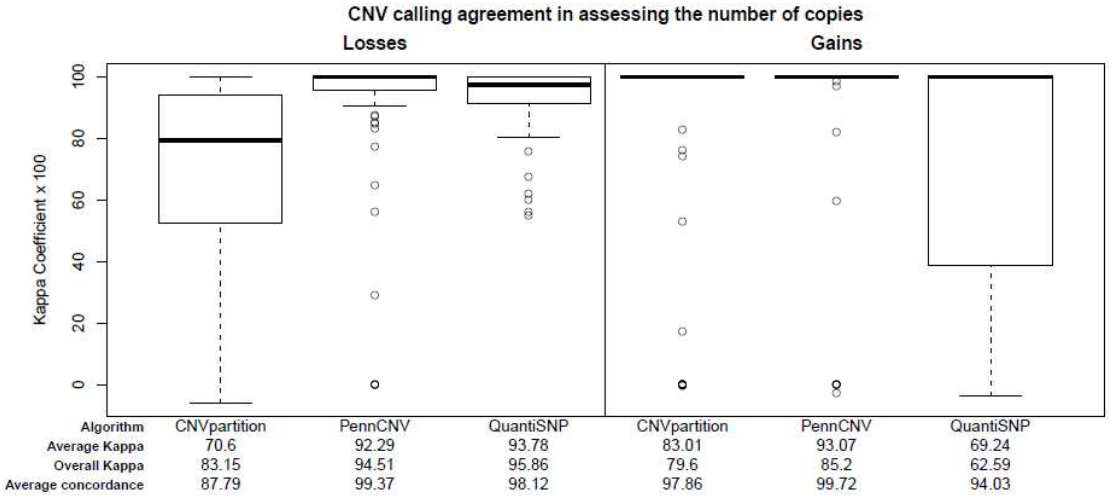
MLPA peaks


84.5
97.5
110.5
123.5
136.5
149.5
1200
800
400
0
Chr22_22690592
95.22
260
Chr17_42061812_421
107.19
1040
Chr16_32481309
117.88
380
Chr14_Pop_1(B)
127.82
356
NSF
146.84
232
FZD9
153.74
294
Chr3_46771035
97.86
655
A_14P114204
108.99
915
Chr17_415_A
120.68
962
Chr4_69231671
102.79
774
Chr8_39356595
112.84
629
Chr14_Pop_1(A)
125.69
1210

Supplementary Figure S2: Detail of the kappa calculation for the two-step agreement on calling CNVs.



Supplementary Figure S3: Agreement on assessing the number of copies once the type of CNV (loss or gain) was concordant for both replicates. For each type of CNV and each algorithm, we computed 1) the Kappa coefficient for each pair of duplicate and we provided the average Kappa across the 96 pairs, 2) a overall Kappa coefficient computed over all the 96 pairs of replicates and concordant probes, and 3) the classic concordance rate for each pair of duplicate and we provided the average concordance across the 96 pairs.



Algorithm	CNVpartition	PennCNV	QuantiSNP	CNVpartition	PennCNV	QuantiSNP
Average Kappa	70.6	92.29	93.78	83.01	93.07	69.24
Overall Kappa	83.15	94.51	95.86	79.6	85.2	62.59
Average concordance	87.79	99.37	98.12	97.86	99.72	94.03

Supplementary Figure S4: Impact of the filtering on PennCNV calling agreement. Box plots before and after filtering for the distribution of A) Kappa Index estimates for CNV detection on duplicated samples, and B) weighted Kappa Index estimates for copy-number assessment when a CNV was detected.

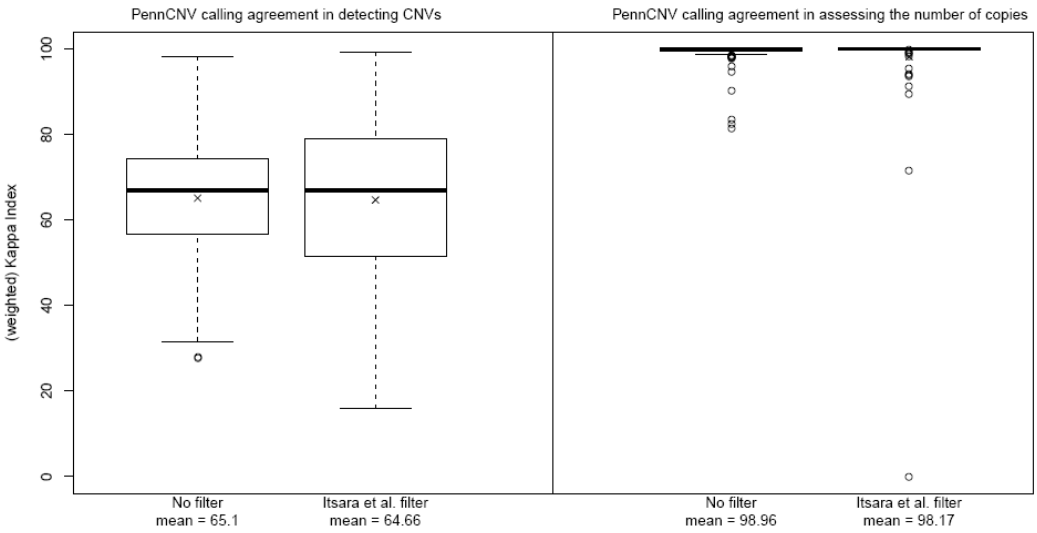
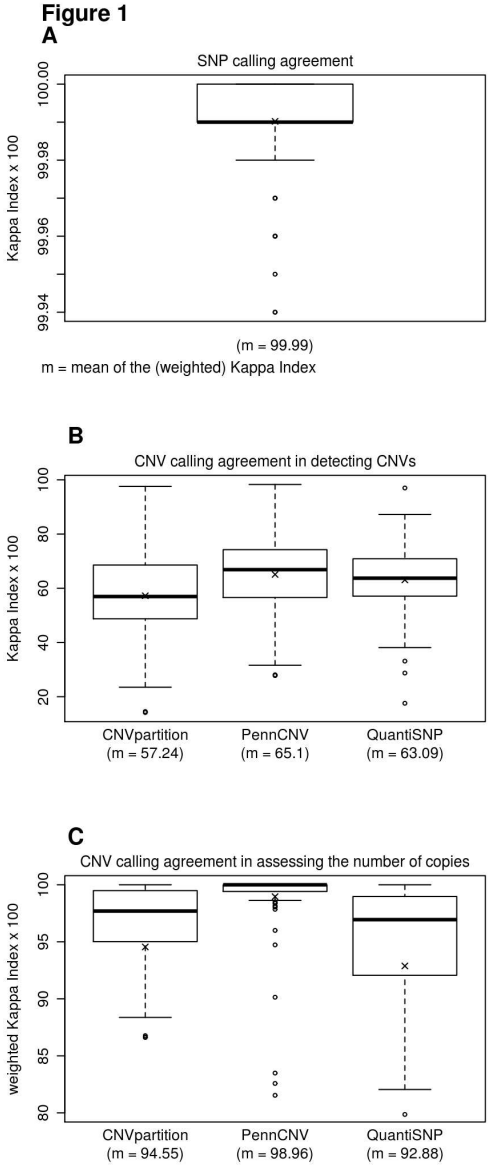
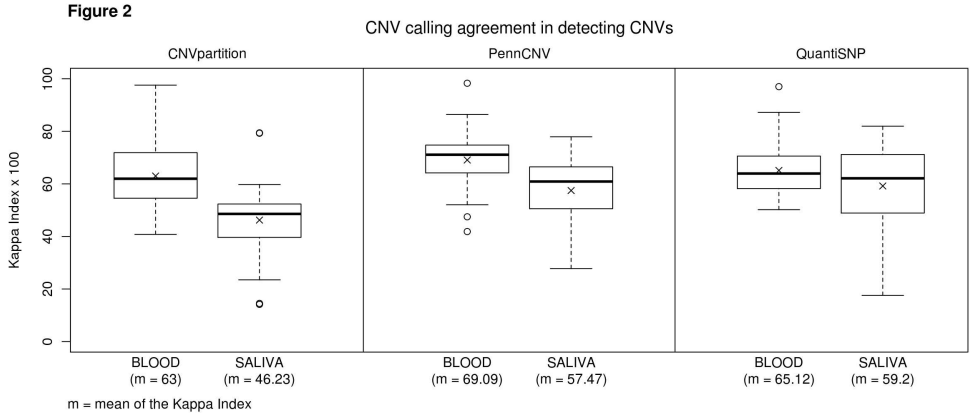


Figure 1



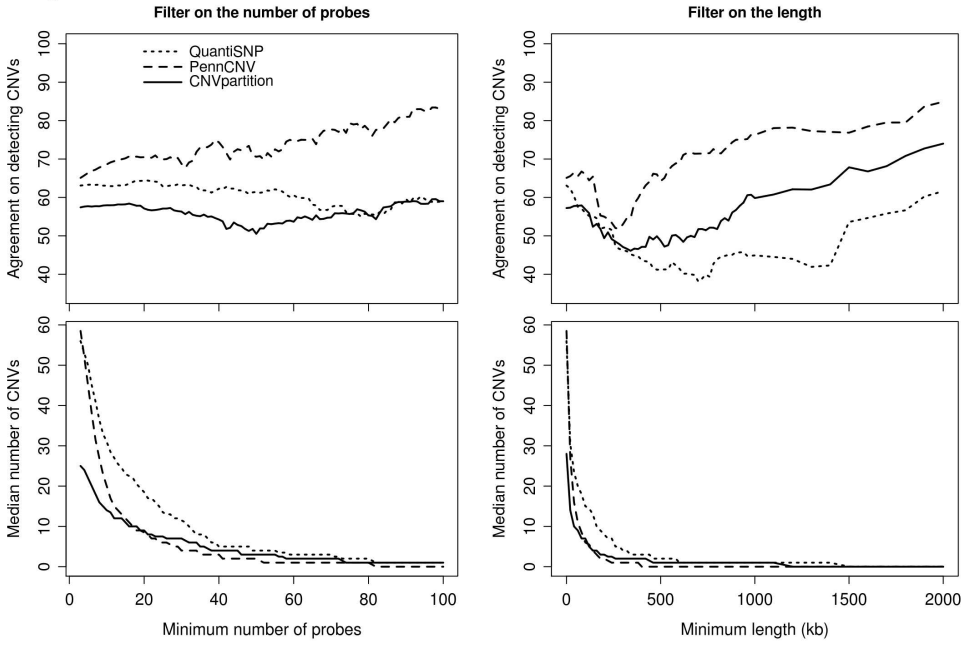
Box plots of the distribution of kappa index estimates comparing duplicated pairs for A) the SNP callings, B) the detection of CNVs according to the different algorithms, and C) the number of copies assigned by the different algorithms in the regions where a CNV was detected.
114x266mm (200 x 200 DPI)

Figure 2



Box plots of the distribution of kappa indexes comparing the callings on duplicated samples by the different algorithms depending on the source of DNA.
304x133mm (200 x 200 DPI)

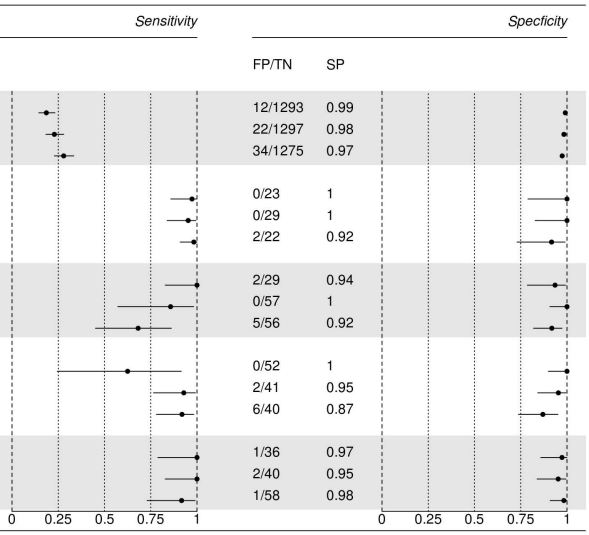
Figure 3



Average Kappa Index for the agreement in detecting CNVs (first row) and median number of CNVs across the 92 individuals (second row) for each algorithm while filtering the called CNVs according the number of probes in the CNV (first column) and the length of the CNV (second column).
279x190mm (200 x 200 DPI)

Figure 4

Steps	CNV type	Algorithm	TP/FN	SE	FP/TN	SP
1.	All CNVs	CNVpartition	60/263	0.19	12/1293	0.99
		PennCNV	71/239	0.23	22/1297	0.98
		QuantiSNP	83/214	0.28	34/1275	0.97
2a.	Deletion*	CNVpartition	36/1	0.97	0/23	1
		PennCNV	40/2	0.95	0/29	1
		QuantiSNP	58/1	0.98	2/22	0.92
2b.	Homozygous deletion*	CNVpartition	29/0	1	2/29	0.94
		PennCNV	12/2	0.86	0/57	1
		QuantiSNP	15/7	0.68	5/56	0.92
2c.	Heterozygous deletion*	CNVpartition	5/3	0.62	0/52	1
		PennCNV	26/2	0.93	2/41	0.95
		QuantiSNP	34/3	0.92	6/40	0.87
2d.	Duplication*	CNVpartition	23/0	1	1/36	0.97
		PennCNV	29/0	1	2/40	0.95
		QuantiSNP	22/2	0.92	1/58	0.98



*Estimates for each CNV type were calculated only for these true positive CNVs identified in step 1.

Sensitivity (SE) and Specificity (SP) estimates for the presence and for the type-specific CNV according to each algorithm.
304x190mm (200 x 200 DPI)