Bioinformatics, 2016, 1–8
doi: 10.1093/bioinformatics/btw477
Advance Access Publication Date: 10 July 2016
Original Paper

OXFORD

Genetics and population analysis

New quality measure for SNP array based CNV detection

A. Macé[1,2,3,*], M.A. Tuke[4], J.S. Beckmann[3], L. Lin[5], S. Jacquemont[6], M.N. Weedon[4], A. Reymond[7] and Z. Kutalik[1,3,*]

[1]Institute of Social and Preventive Medicine, University Hospital of Lausanne, Lausanne, Switzerland, [2]Department of Computational Biology, University of Lausanne, Lausanne, Switzerland, [3]Swiss Institute of Bioinformatics, Lausanne, Switzerland, [4]Genetics of Complex Traits, University of Exeter Medical School, University of Exeter, Exeter, UK, [5]Division of Cardiology, Geneva University Hospital, Geneva, Switzerland, [6]Service de Génétique Médicale, Centre Universitaire Hospitalier Vaudois, Lausanne, Switzerland and [7]Center for Integrative Genomics, University for Lausanne, Lausanne, Switzerland

*To whom correspondence should be addressed.
Associate Editor: Oliver Stegle

Received on April 4, 2016; revised on June 30, 2016; accepted on July 3, 2016

Abstract

Motivation: Only a few large systematic studies have evaluated the impact of copy number variants (CNVs) on common diseases. Several million individuals have been genotyped on single nucleotide variation arrays, which could be used for genome-wide CNVs association studies. However, CNV calls remain prone to false positives and only empirical filtering strategies exist in the literature. To overcome this issue, we defined a new quality score (QS) estimating the probability of a CNV called by PennCNV to be confirmed by other software.

Results: Out-of-sample comparison showed that the correlation between the consensus CNV status and the QS is twice as high as it is for any previously proposed CNV filters. ROC curves displayed an AUC higher than 0.8 and simulations showed an increase up to 20% in statistical power when using QS in comparison to other filtering strategies. Superior performance was confirmed also for alternative consensus CNV definition and through improving known CNV-trait associations.

Availability and Implementation: http://goo.gl/T6yuFM

Contact: zoltan.kutalik@unil.ch or aurelien@mace@unil.ch

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

1 Introduction

Genetic variations range from single nucleotide variations (SNPs) to large chromosomal rearrangement (aneuploidy). Within this spectrum deleted, inserted and duplicated stretches of nucleotides longer than 1 kb (sometimes 500 bp with the use of deep coverage Next-Generation Sequencing) are referred to as copy number variants (CNVs) (Feuk *et al.*, 2006; Valsesia *et al.*, 2013). CNVs have been found genome-wide in both disease and healthy populations (Craddock *et al.*, 2010; Grozeva *et al.*, 2010; Iafrate *et al.*, 2004; Itsara *et al.*, 2009; Jacquemont *et al.*, 2011; Mannik *et al.*, 2015; Redon *et al.*, 2006; Walters *et al.*, 2010; Zhang *et al.*, 2009). Their function has been widely studied, in the context of the rare diseases—rare variant paradigm (Jacquemont *et al.*, 2011; Walters *et al.*, 2010; Zarrei *et al.*, 2015; Zufferey *et al.*, 2012) or as motor of evolution (Feuk *et al.*, 2006; Nguyen *et al.*, 2008; Sudmant *et al.*, 2015). Association with complex traits such as BMI has been shown in the case of the 16p11.2 rearrangement (Jacquemont *et al.*, 2011; Walters *et al.*, 2010; Zufferey et al., 2012), but also with cognitive functions in several general population cohorts [Decode (Stefansson *et al.*, 2014) and Estonia (Mannik *et al.*, 2015)]. Most successful CNV association studies are based on candidate gene approaches (Brasch-Andersen *et al.*, 2004; Fanciulli *et al.*, 2007; Gonzalez *et al.*, 2005; McKinney *et al.*, 2008; Yang *et al.*, 2007) and—unlike SNPs—their genome-wide impact has not yet been fully elucidated.

Consequently, large genome-wide CNV studies would be instrumental to decipher the impact of genetic rearrangements underlying complex traits and disease susceptibility.

Initially, extremely large copy number alterations (>3–5 megabases) were detected using karyotyping (Dowjat and Wlodarska, 1981; Nister *et al.*, 1987; Pepler *et al.*, 1968). Then, fluorescence *in situ* hybridization increased the resolution enabling the detection of sub-microscopic events (Feuk *et al.*, 2005, 2006; Ravnan *et al.*, 2006). Nowadays, CNVs can be detected using different molecular technologies and methods. NGS technologies allow sequencing millions of reads in parallel and new methods for structural variants analysis have been developed, including paired-end mapping, read-depth analysis, split-read strategies and sequence assembly comparisons (Tan *et al.*, 2014; Valsesia *et al.*, 2013; Zhao *et al.*, 2013). The other main methods to detect CNVs are based on micro-array technologies, either Comparative Genome Hybridization (CGH) arrays (Carter, 2007; Kallioniemi *et al.*, 1992; Redon *et al.*, 2009; Ylstra *et al.*, 2006) or SNP genotyping arrays (Conrad *et al.*, 2006; McCarroll *et al.*, 2006).

We focused on CNV detection based on SNP genotyping arrays, as millions of individuals have already been genotyped on SNP arrays and many of these samples were used in meta-analysis of Genome-Wide Association Studies (GWAS) (Locke *et al.*, 2015; Wood *et al.*, 2014). Thus, a wealth of unexploited information remains available in these cohorts and could be re-analyzed for genome-wide CNVs association. The flip side of using genotyping arrays is the lower reliability of the CNV detection, as these platforms were not initially designed to detect such genomic events. However, more recent ones can contain CNV probes.

Despite this limitation, several algorithms have been developed for CNV detection. At each probe, a relative copy number ratio can be obtained by combining the intensities of the two alleles and normalizing this quantity with respect to a reference (log R ratio—LRR). Deviation from the copy number ratio baseline will correspond to either a loss or gain. Several publicly available software (Valsesia *et al.*, 2013) improve copy number calls by exploiting the ratio of allelic intensities (normalized measure of the signal intensity ratio of the B and A alleles—B Allele Frequency—BAF). In this article, we focus on the PennCNV software (Wang *et al.*, 2007), currently the most widely used software for Illumina chips (PennCNV: 955 citations—QuantiSNP: 432 citations). This software, based on a Hidden Markov Model (HMM), is extensively used for CNV detection on SNP. Its speed is a key advantage as number of available samples increase at an unprecedented scale.

Nevertheless, CNV detection remains prone to false positives, which adds noise when such calls are used for trait association. There is a need to develop a new quality score (QS) for CNVs detected by PennCNV, prior to performing CNV-based association studies. Our contribution can be viewed as a post-processing step of PennCNV calls, whereby various CNV metrics are combined to estimate the probability of a called CNV to be a likely consensus call. This probability could then be used as copy number dosage for trait associations. We chose to improve CNV detection in a way that is directly applicable for large meta-analytic GWAS, where analysts preferably want to run only a single and fast CNV calling pipeline and provide association summary statistics from various platforms.

2 Methods

2.1 Cohorts

HYPERGENES is a case/control cohort of 4206 individuals, where controls were selected based on the absence of hypertension while cases were hypertensive (Salvi *et al.*, 2012). Genotyping was done on Illumina 1M-Duo BeadChips capturing 1 199 187 SNPs and was performed in two different centers, one in Geneva and one in Milano; leading to two sub cohorts: HYPERGENES Geneva with 1995 individuals and HYPERGENES Milano with 2,211 individuals. The GenomeStudio software produced final reports with LRR and BAF values.

The Swiss Hepatitis C Cohort Study (SCCS) is a prospective multicentre study carried out in Switzerland and recruiting HCV-positive patients (Prasad *et al.*, 2007). A total of 1152 patients were genotyped on the same Illumina platform as HYPERGENES individuals. The GenomeStudio software produced final reports with LRR and BAF values.

The Swiss Kidney Project on Genes Hypertension (SKIPOGH) is a population-based cross-sectional family study that examines the genetic determinants of blood pressure. The study population included 1128 participants from 271 nuclear families. In our project we used genetic data from 169 trios genotyped on Illumina 2.5. (Ponte *et al.*, 2014; Pruijm *et al.*, 2013).

The UK BioBank is a study of 500 000 individuals from the UK aged between 37 and 73 years and genotyped on Affymetrix Axiom (http://www.ukbiobank.ac.uk/). Data from 119 873 individuals, with genetics and BMI information, were used in the scope of this project.

2.2 CNV calling

CNVs were called using three different software: PennCNV (Wang *et al.*, 2007), QuantiSNP (Colella *et al.*, 2007) and CNVpartition (http://www.illumina.com/documents/products/technotes/technote_cnv_algorithms.pdf). PennCNV is a HMM based software developed to call CNVs using LRR and BAF values for each sample genotyped on Illumina platform. A 'population BAF' (PFB) file was created for each cohort based on 200 randomly selected final reports. The *clean_cnv.pl* script was called with default parameters to merge adjacent CNVs with small gaps. Samples with more than 200 CNVs were excluded from further analysis. QuantiSNP is also a HMM-based software developed to call CNVs using LRR and BAF values. CNVpartition, developed by Illumina, defines 14 different copy number states and for each of them, jointly models the LRRs and BAFs as a bivariate Gaussian distribution. Based on these distributions, it calculates the likelihood of each of the 14 states for a given LRR and BAF. For all three software, CNVs were called using default parameters.

2.3 Consensus CNV

For each CNV detected by PennCNV, we looked at the fraction detected by the two other software, a percentage of agreement for each CNV is then calculated. Zero percent agreement means that none of the CNV probes called by PennCNV is detected by the two other software, 50% means that half of the CNV probes are detected by all the other software and 100% means that all the probes within the CNV are retrieved by both CNVpartition and QuantiSNP. We considered a CNV to be a *consensus call* if its percentage of agreement is above 70%, meaning that the two other software detect at least 70% of it. We used this working definition of a consensus CNV. Other definitions could be imagined (see Section 4), but we chose this mainly due to data availability in large samples.

2.4 Modeling consensus calls

For each CNV detected by PennCNV we would like to predict whether the two other software would confirm it, i.e. whether it is a

consensus call. To this end, we modeled the dichotomized overlap percentage, as a function of various PennCNV parameters using a logistic model (see Equation 1). Ten available parameters were considered: confidence score, CNV length, number of probes, LRR mean, LRR standard deviation, BAF mean, BAF standard deviation, waviness factor (WF) and the total number of CNVs per individual. For more information regarding the definition of the quality metrics, please refer to the Supplementary Table S1, to the PennCNV description (Wang *et al.*, 2007) and the associated website (http://penncnv.openbioinformatics.org/en/latest/). The first three parameters characterize each CNV in a specific sample, while the others correspond to the global signal quality for each individual.

$$\Pr(Y = 1|V_1, V_2, \ldots, V_{10}) = \text{logit}\left(\beta_0 + \sum_{i=1}^{10} \beta_i V_i\right) \quad (1)$$

Variable Y was defined as 1 if the CNV was confirmed by the two other software and zero otherwise. It is the collection of values for all probes and all samples. Variables V_i represent the various CNV/sample parameters provided by the PennCNV software. Step-wise logistic regression using the R function *step* was performed separately for deletions and duplications. Coefficients β_i with a corresponding P-value below 10^{-5} were considered as significant and were set to zero otherwise. Coefficients were estimated separately for each test cohort and used for cross-validation purposes. In the future, the mean of these coefficients can be used for new cohorts. Furthermore, as the coefficients are correlated, we are less interested in their actual values but rather in their combination.

2.5 QS calculation

The coefficients from the logistic model obtained from one/some cohort(s) can be used to estimate the probability of a CNV being a consensus call in other cohorts. We termed this quantity as QS, hence its value indicates the probability of a CNV called by PennCNV parameters being a consensus call.

$$QS_{\text{cnv}} = \frac{1}{1 + \exp(-(\beta_0 + \sum_{i=1}^{n} \beta_i V_i))} \quad (2)$$

These values are multiplied by -1 in case of deletions to retain both quality and copy number information. Note that the variables used in the formula are only the ones retained after stepwise selection (in the independent data set).

2.6 Previous measures of CNV quality

As PennCNV calls admittedly contain many false positives, popular filtering criteria have been recommended and applied in many studies (Chettier *et al.*, 2014; Glessner *et al.*, 2013; Palta *et al.*, 2015; Pinto *et al.*, 2011; Wang *et al.*, 2007). We tested different quality metrics filtering combinations to compare with our QS (Table 1).

2.7 Quality metric comparison through correlation

To evaluate the performance of our newly proposed CNV QS and previously applied metrics, we compared how well the different CNV quality metrics agree with the consensus calls defined by software overlap. First, for each cohort, Spearman correlation between the consensus CNV status and CNV quality metrics were calculated. We compared the confidence score, the different filters and the QS. The QS coefficients were based on the average coefficients leaving out the test cohort.

Table 1. Quality metrics used in different articles to filter CNVs

	Filter A	Filter B	Filter C	Filter D	Filter E
LRR sd	≤0.25	<0.3	<0.24	<0.3	≤0.27
BAF drift	≤0.002	<0.01			
BAF sd	≤0.05				≤0.13
|WF|	≤0.04	<0.05	<0.05	<0.05	
No. CNVs		<100		<100	
Call rate			>0.99	>0.98	≥0.98
Confidence	≥5				
No. Probes			≥10		≥5
Length	≥1 kb				≥1 kb

These filters were used to evaluate the performance of our QS in comparison to what is usually used in the literature: filters A (Palta *et al.*, 2015), B (Wang *et al.*, 2007), C (Chettier *et al.*, 2014), D (Glessner *et al.*, 2013) and E (Pinto *et al.*, 2011).

2.8 Quality metric comparison through receiver operating characteristic curve

To evaluate the discriminatory power of the proposed QS, we computed the receiver operating characteristic (ROC) curve and estimated the area under the curve (AUC). The QS (based on the leave-one (cohort)-out cross-validation) was used as the predictor and the consensus CNV status as the response. Two different definitions of the consensus CNV status were used. The first one considers all the CNVs and defines a consensus CNV as one detected by the three software and false CNV all the others. The second definition was the same as the first, but ignored all CNVs that were called by two software only.

2.9 Quality metric comparison through simulated CNV-phenotype association

Simulations were done at the probe level. When two probes have the same profile across all samples, we kept only one representative. Then, for each cohort, we assessed the performance of the QS. We first converted the CNVs metrics into (no. of probes × no. of samples) tables in order to overcome the problem of different CNVs boundaries across samples. We then simulated association between an *in silico* phenotype and: (i) PennCNV raw CNV calls, (ii) PennCNV confidence score, (iii) the different filters, (iv) the QS based on the average coefficients leaving out one cohort. Then simulations were done for each probe separately. For a specific probe, an *in silico* phenotype is simulated based on the consensus CNV status (defined as the overlap between the three software), an effect size and noise (Equation 3). The effect sizes (β) ranged from 0 to 3 by a step of 0.05 and the noise (ε) was set to follow a standard Gaussian distribution (mean = 0, variance = 1):

$$y = CNV_{consensus} * \beta + \varepsilon \quad (3)$$

Linear regressions were then performed to derive association between this *in silico* phenotype and the estimated CNV status (based on filtered CNVs, confidence score or QS). We ran thousand simulations for each effect size, and we estimated the statistical power to detect an association at $\alpha = 10^{-3}$. Simulations were done separately for deletions and duplications. Within each cohort, we calculated, for each probe, the CNV frequency (Equation 4) and the precision (Prec) (Equation 5).

$$CNV_{Freq} = \frac{\#calledCNV_S}{\#samples} \quad (4)$$

$$CNV_{Prec} = \frac{\#consensusCNV_S}{\#calledCNV_S} \quad (5)$$

CNVs probes were classified according to their precision and frequency. Power computations are presented for 30 (=5 × 6) different bins of frequency (5 bins) and precision (6 bins) combinations. CNV frequency bins boundaries were set at 0, 1, 5, 10, 15, 20 and 100%. Precision bins boundaries were set at 0, 10, 20, 30, 40, 50 and 100%.

2.10 QS validation for other consensus definition

We used pennCNV to call CNV on the 169 trios from the SKIPOGH study. We defined ***consensus CNVs*** as those called by pennCNV both for the proband and at least one of its parents. For all CNVs called for the probands we calculated the QS and also applied the different filters. To assess the agreement between consensus CNV status and the different measures we used logistic regression with consensus status as outcome and (i) QS value, (ii) confidence score, (iii) the different filters as predictors.

2.11 Recovering the known 16p11.2 CNV-BMI association

We called CNVs in the 16p11.2 region in 119 873 UK BioBank participants using pennCNV. For each CNV a QS was estimated and the different filters were applied. Then for the entire cohort, we calculated the association between the inverse quantile normalized BMI and (i) QS, (ii) the confidence score, (iii) the raw copy number state and (iv) the different filters.

3 Results

3.1 Percentage of agreement

The percentage of CNVs detected by PennCNV and confirmed by the two other software is between 20 and 30% (Supplementary Table S2, Supplementary Figs. S1–S3). The results are consistent across the three cohorts with a slightly better percentage of agreement for the duplications than for the deletions. These results mean that three quarters of the PennCNV calls may be false positives, adding noise to downstream analyses. It confirms the necessity to develop a new quality measure that estimates the probability of a CNV call to be a consensus call, in order to increase detection power and avoid spurious associations due to systematic artifacts.

3.2 QS coefficients

For each CNV (called by PennCNV), its QS is computed to estimate the probability of being a consensus call (see Section 2). The QS is derived from a logistic model, whose coefficients (for each CNV and sample characteristic) were estimated separately for deletions and duplications using three cohorts. Coefficients with significant P-value ($<10^{-5}$) were kept for the final model (Supplementary Tables S3–S5). As expected, the confidence score provided by PennCNV is the parameter with the strongest contribution (coefficient ranging from 2.21 to 4.77, $P < 10^{-300}$). It confirms that CNVs with high confidence scores are more likely to be consensus calls. The number of CNVs per individual has a negative coefficient (ranging from −0.57 to −0.17), which means that CNVs called from sample with few CNVs are most probably consensus calls. Interestingly, the coefficients of the mean LRR have opposite values between deletions and duplications. This behavior is consistent across the three cohorts and is unlikely by chance. This is due to the fact that, for a deletion, it is easier to detect a drop in the intensity signal if the global LRR mean value is high. The contrary applies for duplications. The negative coefficients for the number of probes might appear counterintuitive. But in the case of multivariate regression, the estimates of correlated variables (e.g. the correlation between PennCNV confidence score and the number of probes was 0.56) are difficult to interpret. Furthermore, in a univariate model, the number of probes has a significant positive estimate for the duplications in all the three cohorts. Regarding the deletions, estimates are either positive or not significant.

3.3 QS distribution

QS values have been computed separately for deletions and duplications on the three samples groups coming from two cohorts. The distribution of the QSs (Panels A and B Supplementary Figs. S4–S6) shows a bimodal distribution: most of the CNVs are either of bad or good quality, and only few stand in between. As seen for the percentage of agreement, the majority of CNVs are most likely false positive. If one prefers to work with dichotomized CNV calls, we recommend to declare CNVs with QS > 0.5 as consensus CNVs (Mannik *et al.*, 2015).

The CNV frequency distribution (Panels C–D Supplementary Figs. S4–S6) reveals that the majority of deletions with low frequency (≤3%) are of bad quality (QS < 0.5). On the contrary, most of the duplications with a frequency lower than 0.4% have a high-QS (QS ≥ 0.9). A large part of the deletions with frequency ≥3% have intermediate (0.5 ≤ QS < 0.9) or high (QS ≥ 0.9) QS, likewise for duplications with frequency ≥ 10%. Since CNVs are called per sample, frequency might be a predictor. These results show that using our QS to filter CNVs may be particularly useful in the case of rare deletions. In the case of duplications, the advantage is less apparent and would be more relevant for intermediate frequencies (between 0.4 and 10%).

Regarding the length (Panels E and F Supplementary Figs. S4–S6), almost all the CNVs with high QS are longer than 10 kb. It confirms the difficulty to detect small CNVs on SNP array platforms. When applied to CNV-phenotype associations, our QS has the potential to reduce the noise created by false calls especially for CNVs with length <75 kb. It would increase the power to detect associations with short CNVs, which have not yet been implicated with any complex traits.

3.4 Quality metric comparison through correlation

The correlation of the consensus CNV status (defined by the overlap between the three software) with the QS (based on the leave-one (cohort)-out cross-validation) is twice as high as with the classical filters for all types of CNVs and all cohorts (e.g. 0.53 versus 0.26 for Hypergenes GVA and all CNV types; 0.45 versus 0.22 for SCCS and all CNV types; see Table 2 and Supplementary Tables S6–S7). This score has also a ∼20% higher correlation than the confidence score for the deletions and has similar performance for the duplications. The correlations are consistent over the three cohorts and are slightly higher for deletions than for duplications. Table 2 points out the unexpectedly low performance of the classical filters, which perform much worse than the confidence score provided by PennCNV. To summarize, in case of deletions, our QS is a better proxy for the consensus calls than the classical filtering or the PennCNV confidence score.

Although the above table reports only average performance for the whole genome, we give a graphical overview of the distribution of correlations according to CNV frequency and detectability (precision). The box plots (Fig. 1) show detailed performance for CNVs

Table 2. Correlation, for HYPERGENES Geneva samples, between the consensus CNV status and the different quality metrics and filters

HYPERGENES Geneva	QS	Conf score	Filter A	Filter B	Filter C	Filter D	Filter E
All	0.53	0.37	0.24	0.26	0.25	0.26	0.17
Deletions	0.55	0.36	0.27	0.31	0.25	0.30	0.18
Duplications	0.45	0.42	0.16	0.11	0.23	0.12	0.10

A (Palta et al., 2015), B (Wang et al., 2007), C (Chettier et al., 2014), D (Glessner et al., 2013) and E (Pinto et al., 2011).

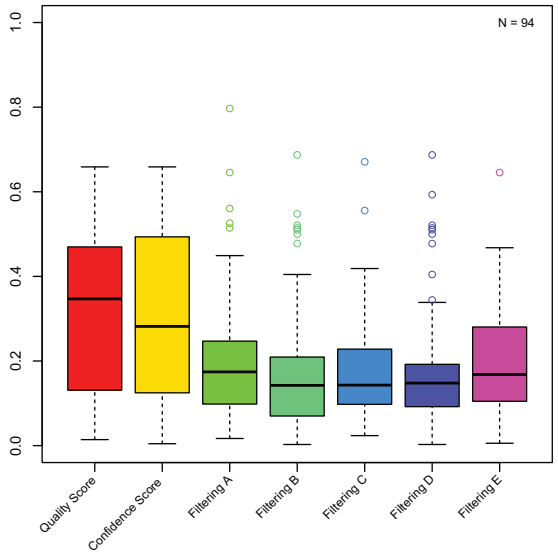

Fig. 1. Boxplots corresponding to the correlations between the consensus CNV status (defined by the overlap between the three software) and quality metrics such as QS, PennCNV confidence score and classical filters: A (Palta *et al.*, 2015), B (Wang *et al.*, 2007), C (Chettier *et al.*, 2014), D (Glessner *et al.*, 2013) and E (Pinto *et al.*, 2011), as defined in the method section. Calculations were done for deletions on the Hypergenes Geneva samples. CNVs were classified according to their precision and frequency. The frequency range for this boxplot is between 1 and 5% while the precision range is between 10 and 20%. The N number corresponds to the number of unique probes being in the precision and CNV frequency window

with frequency 1–5% and detectability 10–20% in the Hypergenes Geneva samples. Results were comparable for other frequency ranges and other cohorts (Supplementary Figs. S7–S9). In summary, classical filters perform worse than our QS and the confidence score from PennCNV. In general, our QS performs discernably better than the confidence score mainly for low frequency CNVs (≤5%) and poor detectability (≤30%). When the frequency and detectability increase, the performances of these two quality metrics become more and more similar. These results are concordant with the frequency distribution figure (Supplementary Figs. S4–S6) and correlation table (Table 2).

3.5 Quality metric comparison through ROC curve

To estimate the discrimination power of our QS to retrieve consensus calls, we computed the ROC curves for all the CNVs, only deletions and only duplications for the Hypergenes Geneva samples (Fig. 2—results for the other cohorts are presented in the Supplementary Figs. S13–S14). AUC are respectively equal to 0.845, 0.871 and 0.774 confirming that the QS recovers well the consensus

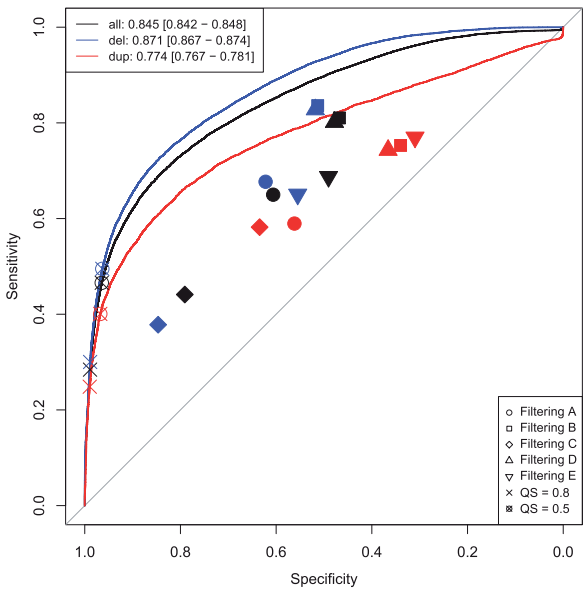

Fig. 2. ROC curve using the QS (based on the leave-one (cohort)-out cross-validation) as predictor and the consensus CNV status as the response. The consensus CNVs are the ones detected by the three software, all the others are defined as false. The plot is based on the Hypergenes Geneva individuals. In black are the results for all the CNVs, in blue for the deletions only and in red for the duplications only. The cross and the star dots give the specificity and sensitivity using a QS threshold of 0.8 and 0.5 respectively. Other symbols correspond to the different filters: A (Palta *et al.*, 2015), B (Wang *et al.*, 2007), C (Chettier *et al.*, 2014), D (Glessner *et al.*, 2013) and E (Pinto *et al.*, 2011)

calls. A threshold of 0.8 will give a specificity of ∼99% with sensitivity between 25 and 30%, while a threshold of 0.5 will have specificity around 96% for sensitivity between 40 and 50%. Based on these results, we recommend using a QS threshold between 0.5 and 0.8 to filter CNVs, if necessary. As comparison, the others filters stand all below the ROC curves (Fig. 2), showing that our QS offers better sensitivity and specificity.

3.6 Quality metric comparison through simulated CNV-phenotype association

The previously reported comparisons only reflect how different filters compare to each other, but do not reveal how much power improvement they could offer in association studies relative to each other or to no filtering. To this end, we performed simulation studies for deletions and duplications separately. Exhaustively sampling CNVs of different characteristics shows an increase up to 20% in statistical power of our QS in comparison to the other quality metrics, for probes with low frequency and detectability (CNVs frequency ≤5%; detectability ≤30%). Figure 3 illustrates power curves for CNVs with frequency between 1 and 5% and precision 10-20% for samples of the Hypergenes Geneva cohort As these two parameters increase, all the CNV quality metrics start to perform similarly and the power curves overlap. Results for other frequency ranges and other cohorts are shown in Supplementary Figures S15–S17. Surprisingly, for most scenarios, classical filters have inferior performance compared with keeping all CNVs without any filtering. It seems that classical filters also remove too many consensus calls and hence decrease power (this explains why the detection power reaches a plateau below 100% in some cases). These simulations show that our QS offers considerable advantage over other quality metrics in particular for rare deletions overlapping poor quality

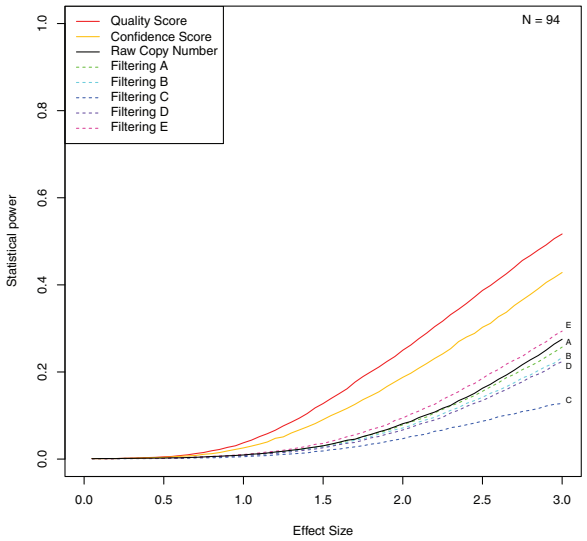

Fig. 3. CNV-phenotype association simulation based on deletions detected in the Hypergenes Geneva samples. Associations were done with different quality metrics such as the QS, the confidence score, the classical filters: A (Palta *et al.*, 2015), B (Wang *et al.*, 2007), C (Chettier *et al.*, 2014), D (Glessner *et al.*, 2013) and E (Pinto *et al.*, 2011), as defined the method section, and the raw CNV calls. CNVs were classified according to their precision and frequency. The frequency range for this boxplot is between 1 and 5% while the precision range is between 10 and 20%. This plot shows the statistical power as a function of the effect size for the different quality metrics. N corresponds to the number of unique probes being in the precision and CNV frequency window

probes. Comparable simulations have been made for duplications (see Supplementary Figs. S18–S20) and, in accordance with the correlation analysis, our QS performs similarly to the confidence score or the raw copy number. Here again, the classical filters perform the worst. Note that our simulations were based on small samples ($n <$ 2000), hence large simulated effect sizes were necessary to reach high meaningful power.

3.7 QS validation for other consensus definition

We estimated the performance of the QS by using a different definition of the consensus CNV status. A CNV detected in an offspring was defined to be consensus if it was also detected in at least one of the two parents. This definition doesn't take into account *de novo* CNVs but their occurrence is relatively low (Veltman and Brunner, 2012). Using this consensus definition, our QS ($P = 4*10^{-80}$) clearly outperforms all classical filters ($P > 2*10^{-40}$) and the confidence score ($P = 7*10^{-60}$) from pennCNV, considering all the CNVs together (Supplementary Table S8) as well as when looking at deletions and duplications separately (Supplementary Tables S9 and S10).

3.8 Recovering the known 16p11.2 CNV-BMI association

Two CNVs in the 16p11.2 region are robustly associated with BMI (Jacquemont *et al.*, 2011; Zufferey et al., 2012). Therefore, they can be used to benchmark the performance of the different quality measures: Stronger association of these CNVs with BMI is an indicator of performance. We computed the association between BMI and (i) the QS, (ii) the confidence score, (iii) the raw copy number state and (iv) the different filters for CNVs detected in the 16p11.2 region using 119 873 individuals from the UK BioBank. The association *P*-values for QS were on average >10-fold smaller than for any that of the other CNV measures. The QQplot (Supplementary Fig. S21) clearly shows inflation for low *P*-values in this region. However, this inflation is stronger when using the QS as genotypic data compared with using any raw data or filters. Furthermore, the fraction of probes, in this region, that pass the genome-wide significance level at $2*10^{-7}$ is clearly higher when using our QS (44.4%) than any of the other CNV measures (39.6% at best, Supplementary Table S11).

4 Discussion

Our QS has been built in order to estimate the probability of a CNV to be a consensus call based on the fact that consensus CNV is defined as the overlap of CNV calls from three distinct software applied to Illumina genotyping arrays. We have demonstrated through ROC, correlation and power analyses that our QS can better recover consensus CNVs than other CNV quality metrics. Naturally, many other consensus definitions could be used. For example, CNVs called by only one other software may not necessarily be false and hence could have been classified as neither false, nor true calls. This definition of consensus, however, did not change the conclusions (Supplementary Figs. S22–S24). Using trio data, we have also explored a definition where CNV calls confirmed in (at least) one of the parents are deemed as consensus CNVs. Notably, our QS recovered these consensus calls significantly better than other quality measures and filters. Finally, we demonstrated the advantage of the QS through (re-)discovering the known 16p11.2 CNV-BMI associations in the UK Biobank. Here again, association *P*-values for the QS were >10-fold smaller than those of any other CNV quality metric.

We thus recommend the use of this metric, as a probabilistic CNV dosage for CNV association studies. This metric has been used to search association between CNV load and cognitive phenotypes in unselected populations (Mannik *et al.*, 2015). Our CNV calls reliably led to stronger association results and retrieved the same CNV frequencies as in the discovery cohort [where some CNVs have been manually confirmed (Mannik *et al.*, 2015)].

It would have been possible to define the consensus as the overlap between CNV calls from multiple genotyping technologies [e.g. genotyping arrays, array CGH (aCGH) or whole genome sequencing (WGS)] applied for the same study participants. Unfortunately, even though aCGH is widely used for CNVs detection in clinical settings, there is no gold standard software (Hupe *et al.*, 2004; Pique-Regi *et al.*, 2008; Valsesia *et al.*, 2012; van Houte *et al.*, 2010; Venkatraman and Olshen, 2007) for genome-wide CNV calling in large population samples. Similarly, algorithms calling CNVs from WGS are in early stages (Abyzov *et al.*, 2011; Chen *et al.*, 2009; Klambauer *et al.*, 2012; Zhu *et al.*, 2012). Another problem is data availability: there are over hundred-fold more population-based samples available with genotyping chip data than those with aCGH or WGS. An additional problem for WGS is that the coverage is usually below 10, which insufficient for reliable CNV calling (http://www.haplotype-reference-consortium.org/). In the future, when such data become available on larger scale one can apply our approach to other consensus CNV definitions.

Over the past years, many GWAS based on Illumina SNP array data have been performed for different traits (Feuk *et al.*, 2005; Locke *et al.*, 2015; Wood *et al.*, 2014). All these data could be reanalyzed to perform CNV association studies using our QS as CNV allele dosage for associations just like the currently used imputed genotype dosages (Kutalik *et al.*, 2011). This would allow a better filtering and would increase statistical power especially for rare deletions as it has been demonstrated in the Results. Such rare variants

have to be meta-analyzed in a way such that the imputation quality is taken into account in the meta-analysis.

As shown in our simulations, the power advantage offered by the QS may not be remarkable for individual cohorts, but much more so in the context of meta-analysis facilitating the collection of large samples. Although larger and larger association studies of common/rare single nucleotide variants reveal increasing proportions of the 'missing heritability', rare CNVs are relatively neglected apart from a few successful examples (Jacquemont *et al.*, 2011; Walters *et al.*, 2010; Zufferey et al., 2012). Therefore, as a future work, we will apply this quality measure in the context of large meta-analytic CNV-association studies on anthropometric traits. In particular, for BMI, the heritability of which is estimated to be 40–70%, but common variants seem to account for <25% (Locke *et al.*, 2015; Maes *et al.*, 1997; Zaitlen *et al.*, 2013). To this purpose, we wrapped up the QS calculation in a pipeline designed to run CNV trait associations. This pipeline has been tested with 18 analysts (on 70 000 samples) and is available online (http://goo.gl/T6yuFM).

Acknowledgements

The Swiss Hepatitis C Cohort Study; The Hypergenes Consortium; We thank the SKIPOGH team for providing access to the SKIPOGH genetic data. This research has been conducted using the UK Biobank Resource; Thanks to Teumer A for allowing access to the SHIP data and to Metspalu A. and Männik K. for allowing access to the EGCUT data for preliminary tests; we thank to Valsesia A. and Porcu E. for their valuable advice and critical review of the article. We thank to Ang W., Deelen P., Hayward C., Kristiansson K., Lenzini P., Liu X., Männik K., Mattsson H., Nõukas M., Rosengren A., Sapkota Y., Schick U., Shrine N., Van Der Most P., Venturini C., Winkler T. and Zhang W. for testing the QS pipeline and reporting valuable feedback.

Funding

This work was supported by grants from the Swiss National Science Foundation (31003A_160203 to AR, 31003A-143914 to Z.K, 33CM30-124087/140333 to SKIPOGH, 148417 to SCCS.), the Leenaards Foundation (Z.K.), the SystemsX.ch (51RTP0_151019 to Z.K.) and FP7 EU grant (HEALTH-F4-2007-201550).

Conflict of Interest: none declared.

References

Abyzov,A. *et al.* (2011) CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Res.*, **21**, 974–984.

Brasch-Andersen,C. *et al.* (2004) Possible gene dosage effect of glutathione-S-transferases on atopic asthma: using real-time PCR for quantification of GSTM1 and GSTT1 gene copy numbers. *Hum. Mut.*, **24**, 208–214.

Carter,N.P. (2007) Methods and strategies for analyzing copy number variation using DNA microarrays. *Nat. Genet.*, **39**, S16–S21.

Chen,K. *et al.* (2009) BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nat. Methods*, **6**, 677–681.

Chettier,R. *et al.* (2014) Endometriosis is associated with rare copy number variants. *PloS One*, **9**, e103968.

Colella,S. *et al.* (2007) QuantiSNP: an Objective Bayes Hidden-Markov Model to detect and accurately map copy number variation using SNP genotyping data. *Nucleic Acids Res.*, **35**, 2013–2025.

Conrad,D.F. *et al.* (2006) A high-resolution survey of deletion polymorphism in the human genome. *Nat. Genet.*, **38**, 75–81.

Craddock,N. *et al.* (2010) Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. *Nature*, **464**, 713–720.

Dowjat,K. and Wlodarska,I. (1981) G-banding patterns in mouse lymphoblastic leukemia L1210. *J. Natl. Cancer Inst.*, **66**, 177–182.

Fanciulli,M. *et al.* (2007) FCGR3B copy number variation is associated with susceptibility to systemic, but not organ-specific, autoimmunity. *Nat. Genet.*, **39**, 721–723.

Feuk,L. *et al.* (2006) Structural variation in the human genome. *Nat. Rev.. Genet.*, **7**, 85–97.

Feuk,L. *et al.* (2005) Discovery of human inversion polymorphisms by comparative analysis of human and chimpanzee DNA sequence assemblies. *PLoS Genet.*, **1**, e56.

Glessner,J.T. *et al.* (2013) ParseCNV integrative copy number variation association software with quality tracking. *Nucleic Acids Res.*, **41**, e64.

Gonzalez,E. *et al.* (2005) The influence of CCL3L1 gene-containing segmental duplications on HIV-1/AIDS susceptibility. *Science*, **307**, 1434–1440.

Grozeva,D. *et al.* (2010) Rare copy number variants: a point of rarity in genetic risk for bipolar disorder and schizophrenia. *Arch. Gen. Psychiatry*, **67**, 318–327.

Hupe,P. *et al.* (2004) Analysis of array CGH data: from signal ratio to gain and loss of DNA regions. *Bioinformatics*, **20**, 3413–3422.

Iafrate,A.J. *et al.* (2004) Detection of large-scale variation in the human genome. *Nat. Genet.*, **36**, 949–951.

Itsara,A. *et al.* (2009) Population analysis of large copy number variants and hotspots of human genetic disease. *Am. J. Hum. Genet.*, **84**, 148–161.

Jacquemont,S. *et al.* (2011) Mirror extreme BMI phenotypes associated with gene dosage at the chromosome 16p11.2 locus. *Nature*, **478**, 97–9102.

Kallioniemi,A. *et al.* (1992) Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. *Science*, **258**, 818–821.

Klambauer,G. *et al.* (2012) cn.MOPS: mixture of Poissons for discovering copy number variations in next-generation sequencing data with a low false discovery rate. *Nucleic Acids Res.*, **40**, e69.

Kutalik,Z. *et al.* (2011) Methods for testing association between uncertain genotypes and quantitative traits. *Biostatistics*, **12**, 1–17.

Locke,A.E. *et al.* (2015) Genetic studies of body mass index yield new insights for obesity biology. *Nature*, **518**, 197–206.

Maes,H.H. *et al.* (1997) Genetic and environmental factors in relative body weight and human adiposity. *Behav. Genet.*, **27**, 325–351.

Mannik,K. *et al.* (2015) Copy number variations and cognitive phenotypes in unselected populations. *Jama*, **313**, 2044–2054.

McCarroll,S.A. *et al.* (2006) Common deletion polymorphisms in the human genome. *Nat. Genet.*, **38**, 86–92.

McKinney,C. *et al.* (2008) Evidence for an influence of chemokine ligand 3-like 1 (CCL3L1) gene copy number on susceptibility to rheumatoid arthritis. *Ann. Rheum. Dis.*, **67**, 409–413.

Nguyen,D.Q. *et al.* (2008) Reduced purifying selection prevails over positive selection in human copy number variant evolution. *Genome Res.*, **18**, 1711–1723.

Nister,M. *et al.* (1987) Evidence for progressional changes in the human malignant glioma line U-343 MGa: analysis of karyotype and expression of genes encoding the subunit chains of platelet-derived growth factor. *Cancer Res.*, **47**, 4953–4960.

Palta,P. *et al.* (2015) Haplotype phasing and inheritance of copy number variants in nuclear families. *PloS One*, **10**, e0122713.

Pepler,W.J. *et al.* (1968) An unusual karyotype in a patient with signs suggestive of Down's syndrome. *J. Med. Genet.*, **5**, 68–71.

Pinto,D. *et al.* (2011) Comprehensive assessment of array-based platforms and calling algorithms for detection of copy number variants. *Nat. Biotechnol.*, **29**, 512–520.

Pique-Regi,R. *et al.* (2008) Sparse representation and Bayesian detection of genome copy number alterations from microarray data. *Bioinformatics*, **24**, 309–318.

Ponte,B. *et al.* (2014) Reference values and factors associated with renal resistive index in a family-based population study. *Hypertension*, **63**, 136–142.

Prasad,L. *et al.* (2007) Cohort Profile: the Swiss Hepatitis C Cohort Study (SCCS). *Int. J. Epidemiol.*, **36**, 731–737.

Pruijm,M. *et al.* (2013) Heritability, determinants and reference values of renal length: a family-based population study. *Eur. Radiol.*, **23**, 2899–2905.

Ravnan,J.B. *et al.* (2006) Subtelomere FISH analysis of 11 688 cases: an evaluation of the frequency and pattern of subtelomere rearrangements in individuals with developmental disabilities. *J. Med. Genet.*, **43**, 478–489.

Redon,R. *et al*. (2009) Comparative genomic hybridization: DNA labeling, hybridization and detection. *Methods Mol. Biol*., **529**, 267–278.

Redon,R. *et al*. (2006) Global variation in copy number in the human genome. *Nature*, **444**, 444–454.

Salvi,E. *et al*. (2012) Genomewide association study using a high-density single nucleotide polymorphism array and case-control design identifies a novel essential hypertension susceptibility locus in the promoter region of endothelial NO synthase. *Hypertension*, **59**, 248–255.

Stefansson,H. *et al*. (2014) CNVs conferring risk of autism or schizophrenia affect cognition in controls. *Nature*, **505**, 361–366.

Sudmant,P.H. *et al*. (2015) Global diversity, population stratification, and selection of human copy-number variation. *Science*, **349**, aab3761.

Tan,R. *et al*. (2014) An evaluation of copy number variation detection tools from whole-exome sequencing data. *Hum. Mutat*., **35**, 899–907.

Valsesia,A. *et al*. (2013) The Growing Importance of CNVs: New Insights for Detection and Clinical Interpretation. *Front. Genet*., **4**, 92.

Valsesia,A. *et al*. (2012) Identification and validation of copy number variants using SNP genotyping arrays from a large clinical cohort. *BMC Genomics*, **13**, 241.

van Houte,B.P. *et al*. (2010) CGHnormaliter: a Bioconductor package for normalization of array CGH data with many CNAs. *Bioinformatics*, **26**, 1366–1367.

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

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

Walters,R.G. *et al*. (2010) A new highly penetrant form of obesity due to deletions on chromosome 16p11.2. *Nature*, **463**, 671–675.

Wang,K. *et al*. (2007) PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Res*., **17**, 1665–1674.

Wood,A.R. *et al*. (2014) Defining the role of common variation in the genomic and biological architecture of adult human height. *Nat. Genet*., **46**, 1173–1186.

Yang,Y. *et al*. (2007) Gene copy-number variation and associated polymorphisms of complement component C4 in human systemic lupus erythematosus (SLE): low copy number is a risk factor for and high copy number is a protective factor against SLE susceptibility in European Americans, *Am. J. Hum. Genet*., **80**, 1037–1054.

Ylstra,B. *et al*. (2006) BAC to the future! or oligonucleotides: a perspective for micro array comparative genomic hybridization (array CGH). *Nucleic Acids Res*., **34**, 445–450.

Zaitlen,N. *et al*. (2013) Using extended genealogy to estimate components of heritability for 23 quantitative and dichotomous traits. *PLoS Genet*., **9**, e1003520.

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

Zhang,F. *et al*. (2009) Copy number variation in human health, disease, and evolution. *Annu. Rev. Genomics Hum. Genet*., **10**, 451–481.

Zhao,M. *et al*. (2013) Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives. *BMC Bioinformatics*, **14** (**Suppl 11**), S1.

Zhu,M. *et al*. (2012) Using ERDS to infer copy-number variants in high-coverage genomes. *Am. J. Hum. Genet*., **91**, 408–421.

Zufferey,F. *et al*. (2012) A 600 kb deletion syndrome at 16p11.2 leads to energy imbalance and neuropsychiatric disorders. *J. Med. Genet*., **49**, 660–668.