Association Tests Using **Co**py **N**umber Profile **Cur**ves (CONCUR) Enhances Power in Rare Copy Number Variant Analysis

Amanda Brucker[1], Wenbin Lu[1], Rachel Marceau West[1], Qi-You Yu[2], Chuhsing Kate Hsiao[2], Tzu-Hung Hsiao[3], Ching-Heng Lin[3], Patrik K. E. Magnusson[4], Patrick F. Sullivan[4,5], Jin P. Szatkiewicz[5], Tzu-Pin Lu[2], Jung-Ying Tzeng[1,2,6,7*],

1 Department of Statistics, North Carolina State University, Raleigh, North Carolina, United States of America

2 Institute of Epidemiology and Preventive Medicine, National Taiwan University, Taipei, Taiwan

3 Department of Medical Research, Taichung Veterans General Hospital, Taiwan

4 Department of Medical Epidemiology and Biostatistics, Karolinska Institutet SE-171 77 Stockholm, Sweden

5 Department of Genetics, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, United States of America

6 Bioinformatics Research Center, North Carolina State University, Raleigh, North Carolina, United States of America

7 Department of Statistics, National Cheng-Kung University, Tainan, Taiwan

* Corresponding author

Email: jytzeng@ncsu.edu (JYT)

Abstract

Copy number variants (CNVs) are the gain or loss of DNA segments in the genome that can vary in dosage and length. CNVs comprise a large proportion of variation in human genomes and impact health conditions. To detect rare CNV association, kernel-based methods have been shown to be a powerful tool because their flexibility in modeling the aggregate CNV effects, their ability to capture effects from different CNV features, and their ability to accommodate effect heterogeneity. To perform a kernel association test, a CNV locus needs to be defined so that locus-specific effects can be retained during aggregation. However, CNV loci are arbitrarily defined and different locus definitions can lead to different performance depending on the underlying effect patterns. In this work, we develop a new kernel-based test called CONCUR (i.e., Copy Number profile Curve-based association test) that is free from a definition of locus and evaluates CNV-phenotype association by comparing individuals' copy number profiles across the genomic regions. CONCUR is built on the proposed concepts of "copy number profile curves" to describe the CNV profile of an individual, and the "common area under the curve (cAUC) kernel" to model the multi-feature CNV effects. Compared to existing methods, CONCUR captures the effects of CNV dosage and length, accounts for the continuous nature of copy number values, and accommodates between- and within-locus etiological heterogeneities without the need to define artificial CNV loci as required in current kernel methods. In a variety of simulation settings, CONCUR shows comparable and improved power over existing approaches. Real data analyses suggest that CONCUR is well powered to detect CNV effects in gene pathways associated with phenotypes using data from the Swedish Schizophrenia Study and the Taiwan Biobank.

Author summary

Copy number variants comprise a large proportion of variation in human genomes. Large rare CNVs, especially those disrupting genes or changing the dosages of genes, can carry relatively strong risks for neurodevelopmental and neuropsychiatric disorders. Kernel-based association methods have been developed for the analysis of rare CNVs and shown to be a valuable tool. Kernel methods model the collective effect of rare CNVs using flexible kernel functions that capture the characteristics of CNVs and measure CNV similarity of individual pairs. Typically kernels are created by summarizing similarity within an artificially defined "CNV locus" and then collapsing across all loci. In this work, we propose a new kernel-based test, CONCUR, that is based on the CNV location information contained in standard processing of the variants and removes the need for any arbitrarily defined CNV loci. CONCUR quantifies similarity between individual pairs as the common area under their copy number profile curves and is designed to detect CNV dosage, length and dosage-length interaction effects. In simulation studies and real data analysis, we demonstrate the ability of CONCUR test to detect CNV effects under diverse CNV architectures with power and robustness over existing methods.

Introduction

Copy number variants (CNVs) are unbalanced structural variants that are typically 1 kilobase pair (kb) in size or larger and lead to more or fewer copies of a region of DNA with respect to the reference genome. CNVs are typically characterized by two descriptive features. The first feature is CNV dosage, or the total number of copies present, with > 2 copies corresponding to duplications and < 2 copies corresponding to deletions. The second is the CNV length, typically measured in base pairs (bp) or kilobase pairs.

CNVs are important risk factors for many human diseases and traits, including Crohn's disease, HIV susceptibility, and body mass index [1–3]. Large and rare CNVs are particularly implicated in neuropsychiatric disorders including autism spectrum disorder, schizophrenia, bipolar disorder, and attention deficit disorder [4]. For example, multiple studies have confirmed a greater burden of rare CNVs in schizophrenia cases compared with normal controls, both genome-wide and in specific neurobiological pathways important to schizophrenia (e.g., calcium channel signaling and binding partners of the fragile X mental retardation protein).

Typically, rare CNVs (e.g., $< 1\%$ frequency) in the genome are intractable to test individually for disease association and instead are examined with collapsing methods. Collapsing methods summarize variant characteristics across multiple variants in a targeted region, typically a gene set or the whole genome, and perform a test of the collective CNV effects. By accumulating information across multiple rare variants, collapsing methods can have enhanced power to detect the effects of rare CNVs that are hard to detect individually but collectively have a significant impact. Collapsing tests for rare CNVs are primarily built on the foundation of rare single nucleotide polymorphism (SNP) association tests but with additional complexity to accommodate the length and dosage features of CNVs. As with SNPs, the effects of CNVs can vary between loci, but CNV collapsing tests must also account for within-locus heterogeneity due to differential dosage effects or length effects within a CNV region.

Similar to SNP collapsing tests, there are also two families of tests for rare CNV analysis: burden-based methods and kernel-based methods. Burden-based tests, e.g., Raychaudhuri et al. [5], summarize the CNV features of an individual via the total CNV counts or average length and model the CNV effects as fixed effects assuming etiological homogeneity of features across multiple CNVs of a targeted region. Kernel-based tests, e.g., CCRET [6] and CKAT [7], aggregate CNV information via genetic similarity based on certain CNV features and model CNV effects as random effects to account for the between-locus etiological heterogeneity. By design, burden tests are optimal when the association signal is driven by homogeneous effects across CNVs, and kernel-based tests are optimal in the presence of etiological heterogeneity. Burden tests often need to subset CNVs by dosage (e.g., deletions only or duplications only) or size (e.g. > 100kb, > 500kb) to increase homogeneity while kernel-based tests do not have such requirements.

In this work, we focus on kernel-based methods because etiological heterogeneity is becoming a more practically encountered scenario as high-resolution CNV detection technologies permit the detection of CNVs with smaller length. In kernel-based association tests, the association between CNVs and the trait is evaluated by examining the correlation between trait similarity and CNV similarity quantified in a kernel. For kernel construction, we can refer to kernel-based tests for SNPs; since SNPs are evaluated at the same single base-pair position (referred to as a locus) across individuals, it is natural to assess similarity locus-by-locus and aggregate the locus-level similarity over all loci in the target region to obtain an overall SNP similarity. A locus unit for CNVs, however, is not so obvious since CNVs span multiple base pairs and may overlap partially between individuals.

To address this issue, standard CNV kernel-based tests construct CNV regions (CNVR). For example, the CNV Collapsing Random Effects Test (CCRET) [6] creates CNVR by clustering CNV segments of different individuals with some arbitrary amount of overlap (e.g., 1 base pair overlap, 50% reciprocal overlap). With CNVRs, the CNV similarity between an individual pair can be quantified first within each CNVR, and this CNVR-level similarity can be summed over all CNVRs in the target region to characterize overall CNV similarity. However, a drawback of this approach is that CNVRs defined in this fashion are contingent on the unique CNV overlapping patterns among individuals in a study, and the defined CNVRs can vary from one study to another. The arbitrary choice of overlapping threshold also impacts the formation of locus units and consequently how the "between-locus" and "within-locus" heterogeneous effects of CNVs are accounted for.

To avoid the issues introduced by arbitrarily defined CNVRs as in CCRET, the CNV Kernel Association Test (CKAT) [7] adopts a different strategy to quantify CNV similarity between two individuals. Specifically, CKAT allows users to define the CNVR as a biologically relevant region, e.g., a chromosome. CKAT also introduces a new kernel function to measure CNV similarity based on both dosage and length features between two CNV events. This CNV-level similarity is then aggregated to derive a measure of CNVR-level similarity using a shift-by-one scanning algorithm that "aligns" CNVs in two profiles based on their ordinal position. A multiple-testing correction is applied if multiple CNVRs are involved in the targeted region. Although the new strategy bypasses the need of an arbitrarily defined locus unit, the scanning alignment may yield unreliable results if CNVRs are too large and distant CNVs contribute to an inaccurate model of profile similarity. In addition, there are computational considerations with a scanning algorithm. Furthermore, CKAT aligns pairs of CNVs based on their ordinal position rather than considering all possible pairs which may not optimally capture similarity.

To address these challenges in quantifying CNV similarity using kernel-based methods, in this work we propose a new approach called the Copy Number profile Curve-based (CONCUR) association test. Based on the concept of copy number (CN) profile curves (introduced below), the CONCUR association test naturally incorporates both CNV dosage and length features and can capture their main effects as well as dosage-length interactions. Additionally, building the kernel based on CN profile curves permits the quantification of CNV similarity without the need for pre-specified locus units. Moreover, CNV length may be incorporated flexibly in units which are supported in good resolution by the sequencing technology or which are computationally stable. Like CCRET and CKAT, the test is built in the framework of kernel machine regression and is powerful under heterogeneous signals and can adjust for confounders. In this analysis, we use simulation studies to demonstrate the improved power CONCUR over existing kernel-based methods in a variety of settings and illustrate the practical utility of CONCUR by conducting pathway analysis on the Swedish Schizophrenia Study data and the Taiwan Biobank data.

Results

Overview of CONCUR

CONCUR assesses the collective effects of rare CNVs on a phenotype in a kernel machine regression framework where the kernel construction does not require a defined CNV locus. As such, CONCUR is built on two major components: (a) the CN profile curve, with which we describe an individual's CNVs across the genome or a region of interest; and (b) the common area under the curve (cAUC) kernel, with which we measure CNV similarity between two individuals and characterize the CNV effects on the phenotype. In a CN profile curve (e.g., Fig 1), CNV dosage is shown on the y-axis as jumps or troughs diverging from a baseline of 2 copies; the start and end points of the jumps and troughs correspond to the start and end locations of the CNV and are shown on the x-axis. At genomic locations where there are no CNV events, the y-axis (dosage) takes value 2 (i.e., the baseline value). CN profile curves are intended to be a visualization of CNV activity and concurrence across samples and contribute to the CONCUR method through the concept of cAUC.

By superimposing two CN profile curves, we identify regions of overlapping CNVs of the same type (i.e., deletion or duplication) and propose to use the common area under the curve (cAUC) to quantify CNV similarity between two individuals. To implement the idea, first the raw dosage values in the CN profile curve are centered and scaled to obtain the duplication profile curve and deletion profile curve. The scaling and centering can be achieved by the dosage (DS) transform functions: $a^{Dup}(DS) = (DS - 2)^d$ for duplications and 0 otherwise, and $a^{Del}(DS) = (2 - DS)^d$ for deletions and 0 otherwise, where d is some pre-specified constant. Second, we superimpose the duplication profile curves of two individuals and note the overlapping regions where both curves are non-zero. Third, for each overlapping region, we multiply the minimum of the two respective transformed dosage values by the length of the overlap, and save this measure of "area of commonality". Finally, we calculate the cAUC between two individuals as the sum of all such areas of commonality in their duplication profile curves plus the sum of all areas in their deletion profile curves. In the special case with $d = 1$ in the dosage transform functions $a^{Dup}(DS)$ and $a^{Del}(DS)$, the cAUCs between various pairs of individuals are illustrated in Fig 1. For individuals with overlapping CNVs of dosage 4 (for duplications; Fig 1 (b)) or dosage 0 (for deletions; Fig 1 (c)), the cAUC is the overlapping length times 2. For individuals with overlapping CNVs of dosage 3 (for duplications; Fig 1 (d)) or dosage 1 (for deletions; Fig 1 (e)), the cAUC is the overlapping length times 1. The cAUC between individuals with overlapping CNVs of the same type but different dosages (e.g., 3 versus 4), is the length of the overlap times 1 (Fig 1 (f)). If there are multiple overlaps in the individuals' CN profile curves, the cAUC between two individuals is the sum of all areas of commonality (e.g., sum of shaded regions in Fig 1 (g)). The cAUC kernel measures similarity in both CNV length and dosage and hence characterizes the joint dosage and length effects. Using the semi-parametric kernel machine regression framework, CONCUR regresses the trait values on CNV effects captured by the cAUC kernel and evaluates the association between traits and CNV profiles via a score-based variance component test.

Fig 1. Diagram of copy number profile curves and common area under the curve. (a) Example of CNV data in standard PLINK format describing profiles of individuals in a small region of chromosome 1. (b)&(c) Copy number (CN) profile curves illustrating the cAUC between individuals with overlapping duplications of dosage 4 in (b) and individuals with overlapping deletions of dosage 0 in (c). (d)&(e) CN profile curves illustrating the cAUC between individuals with overlapping duplications of dosage 3 in (d) and individuals with overlapping deletions of dosage 1 in (e). (f) CN profile curves illustrating the cAUC between individuals with overlapping duplications of dosage 3 and 4. (g) CN profile curves which contain overlapping CNVs in multiple locations, so that the cAUC between the individuals is the sum of the two areas.

Simulation design

The simulations were based on the pseudo-CNV data of 2000 individuals which is publicly available at
`https://www4.stat.ncsu.edu/~jytzeng/Software/CCRET/software_ccret.php`.
Autosome-wide pseudo-CNV data were simulated by mimicking the CNV profiles of unrelated individuals in the TwinGene study [8], and details are described in Tzeng et

al. [6]. Briefly, the TwinGene study used a cross-sectional sampling design and included over 6,000 unrelated subjects born between 1911 and 1958 from the Swedish Twin Registry [9, 10]. CNV calls were generated using Illumina OmniExpress beadchip for 72,881 SNP markers and using PennCNV (version June 2011) [11] as the CNV calling algorithm with recommended model parameters. From the full callset, high quality rare CNVs (frequency $< 1\%$ and size > 100kb) were extracted to form the simulation pool for the pseudo-CNV data. By mimicking the CNV profiles observed in a population dataset such as TwinGene, the pseudo-CNV data are appropriate for the simulation studies in this work. The pseudo-CNV data are stored in PLINK format indicating individual ID, CNV chromosome and starting and ending locations in base pairs (bp), and CNV dosage (e.g., 0, 1, 2, 3, etc.).

For the purpose of simulations we constructed "CNV segments" based on the pseudo-CNV profiles. The endpoints of the segments correspond to locations where a CNV in any one of the samples begins or ends, resulting in segments that contain either one or more intersecting CNVs. Within a segment, CNV dosage of an individual is a constant, and CNVs across individuals may have different dosages but share the same starting and ending positions. Note that different segments will naturally have different lengths. In the simulation study, we built design matrices $\mathbf{Z}^{Dup}, \mathbf{Z}^{Del}$, and $\mathbf{Z}^{Len}$ which codified CNV features by segment in the pseudo-CNV profile data. The dosage matrices took value 0 for those individuals without CNVs in the segment and were coded as 1 or 2 according to the number of additional or missing copies comprising the CNV. Length was the length of the CNV segment in kb for individuals with CNV events and was 0 for individuals without CNVs in the segment.

A case-control phenotype was generated from the logistic model

$$\begin{aligned} \text{logit}(\Pr(Y_i = 1)) \quad &= \quad \gamma_0 + \beta_X X_i + \sum_{j=1}^{R} \beta_j^{Dup} Z_{ij}^{Dup} + \sum_{j=1}^{R} \beta_j^{Del} Z_{ij}^{Del} + \sum_{j=1}^{R} \beta_j^{Len} Z_{ij}^{Len} \\ &+ \quad \sum_{j=1}^{R} \beta_j^{Dup*Len} Z_{ij}^{Dup} Z_{ij}^{Len} + \sum_{j=1}^{R} \beta_j^{Del*Len} Z_{ij}^{Del} Z_{ij}^{Len}, \qquad (1) \end{aligned}$$

where $Z_{ij}{}^{\bullet}$ is the (i, j) entry of matrix $\mathbf{Z}^{\bullet}$, $i = 1, \cdots, N$ indexes individuals, and $j = 1, \cdots, R$ indicates CNV segment. A binary covariate X_i was simulated from Bernoulli(0.5) for each individual. β_j^{Dup} and β_j^{Del} are the log-odds ratios of segment j

for the presence of a CNV versus the absence. Likewise, β_j^{Len} controls the effect of CNV length in segment j, and $\beta_j^{Dup*Len}$ and $\beta_j^{Del*Len}$ allow the effects of CNV length to differ by dosage. $\beta_j^{\bullet} > 0$ (or < 0) corresponds to a deleterious (or protective) CNV effect, and $\beta_j^{\bullet}$ was set to 0 in non-causal segments. We set $\beta_X = \log(1.1)$ and $\gamma_0 = -2$, which corresponds to a disease rate of $\exp(-2) = 13.5\%$ in baseline population. We also fixed $\beta_j^{Len} = 0$ to reflect the observation that length tends to act like an effect modifier of dosage effects.

Among the CNV segments across the genome, we selected 200 segments to be causal, which consist of 100 causal "dup-segments" with at least one duplication and another 100 causal "del-segments" with at least one deletion. A causal dup-segment cannot be a causal del-segment. These causal segments were chosen as a random draw of 50 pairs of adjacent segments which both contained duplications, and another 50 pairs of adjacent segments which both contained deletions. This adjacent causal segment approach was designed to ensure that causal regions had more realistic lengths, since some segments were very short by chance.

We compare the performance of CONCUR with CCRET and CKAT. To implement CCRET, we used the functions from the CCRET package to convert the PLINK data to CCRET design matrices and computed the dosage kernel matrix. For CKAT, following Zhan et al. [7], we designated each chromosome as a CNVR and performed an association test for each chromosome. We reported the Bonferroni-corrected p-value for an overall association by multiplying the minimum p-value among the 22 association tests by 22. CNV lengths within each chromosome were scaled to be in [0,1] by dividing by the range of each chromosome, i.e., the maximal ending position minus the minimal starting position of observed CNVs on each chromosome. The Gaussian kernel scaling parameter was set to be 1.

We examined the methods' performance under two signals: in Scenario I under a dosage×length signal and in Scenario II under a dosage-only signal. We chose these signals to roughly replicate the simulation settings applied to assess CKAT in [7] (dosage×length signal) and to assess CCRET in [6] (dosage signal). Under each scenario, we considered three sub-scenarios: (a) causal duplication effects only (referred to as Scenario I.a or II.a); (b) causal deletion effects only (referred to as Scenario I.b or II.b); and (c) both duplications and deletions to be causal (referred to as Scenario I.c and II.c). Within each sub-scenario, we varied the percentage of deleterious and protective effects by letting a percentage of the causal segments be deleterious or protective. We considered (1) 100% deleterious effects, (2) 50% deleterious and 50% protective, and (3) 10% deleterious and 90% protective. The choice of asymmetric heterogeneity settings was motivated by the rarity of 100% protective CNV effects in a genome-wide analysis, whereas 100% risk-associated effects are not uncommon. The power was evaluated in the range of odds ratios (exp(β)) 1.02-1.10 for Scenario I (dosage×length effects) and 1.1-1.9 for Scenario II (dosage effects). Power estimates are reported for a range of effect sizes such that the power ranges roughly from 0.2 to 0.8.

We implemented case-control sampling to obtain 2000 cases and 2000 controls for each simulation replication. Type I error rates were evaluated based on 5000 replications, and power was estimated based on 300 replications at each effect size. For all methods (i.e., CONCUR, CCRET and CKAT), we adjusted for a simulated binary covariate as a fixed effect in the kernel machine regression. We employed the small-sample variance components test of Chen et al. [12] and obtained p-values using Davies' method [13] as implemented in the CKAT R package.

Simulation Results

The type I error rates of the three tests were examined at nominal levels of 0.01, 0.05, and 0.1 (Table 1). All methods had type I error rates roughly around the nominal level.

Table 1. Type I error rates. Type I error rates of three CNV tests evaluated based on 5000 replications.

Nominal level	CONCUR	CCRET	CKAT
0.01	0.010	0.008	0.009
0.05	0.045	0.047	0.049
0.10	0.096	0.093	0.092

Scenario I: Causal Dosage×Length Effects. Scenario I.a (I.b) considers dosage-length interactions only from causal duplication (deletion) segments, and includes three settings of mixed deleterious and protective effects which are labeled as (D,P)=(100,0), (50,50) and (10,90); (D,P) indicates the proportion of deleterious (D) and protective (P) segments among all causal segments. The results are displayed in Fig 2, with the top row showing power under causal duplication effects and bottom row under causal deletion effects. The CONCUR method has the best or comparable power with the second best method (CCRET) across different settings of deleterious-protective effects. Both CONCUR and CKAT are designed to detect dosage×length signals, but CKAT struggled to pick up this signal perhaps due to applying the method to very large CNVRs (chromosomes) as well as the multiple testing penalty. We also observed a difference in relative performance in the (D,P)=(50,50) setting between I.a (causal duplications) and I.b (causal deletions). This is not unexpected because in the simulated data, there are differences in the features of duplication and deletion events. The proportion of the causal deletion sites out of all deletions was 9.5%, and is 6.9% for duplications. In addition, the 100 causal duplication segments had higher median and mean length compared to the 100 causal deletion segments (median 75kb vs. 32kb; mean 81kb vs. 64kb).

Scenario I.c considers dosage-length interactions from both duplications and deletions and includes four settings of mixed deleterious and protective effects. (Fig 3) These settings are denoted as ($\mathrm{D_{Dup}}$,$\mathrm{P_{Dup}}$,$\mathrm{D_{Del}}$,$\mathrm{P_{Del}}$)=(100,0,100,0), (50,50,50,50), (90,10,10,90), and (10,90,90,10), where $\mathrm{D_{Dup}}$ and $\mathrm{P_{Dup}}$ respectively are the proportions of deleterious and protective segments among causal duplication segments, and $\mathrm{D_{Del}}$ and $\mathrm{P_{Del}}$ are defined similarly for causal deletion segments. These settings allow the assessment of the method performance under multiple sources of effect heterogeneity, including between-locus heterogeneity due to the mixture of deleterious and protective segments, between-locus heterogeneity due to duplication and deletion causal segments, and within-locus heterogeneity due to duplications and deletions with a segment having opposite effects. We observed that CONCUR has the best power among the three tests across different settings, followed by CCRET and then by CKAT.

Fig 2. Power comparison between CONCUR, CCRET, and CKAT under Scenario I (causal dosage×length effects). The top panel shows results from Scenario I.a (causal duplication effects) and the bottom panel from Scenario I.b (causal deletion effects). In each sub-scenario, three different proportions of deleterious vs. protective effects are considered as indicated by (D,P), with D representing the proportions of deleterious segments and P the protective segments among causal segments.

Fig 3. Power comparison between CONCUR, CCRET, and CKAT for Simulation I.c (causal dosage×length effects from both duplications and deletions). Four different proportions of deleterious vs. protective effects are considered as indicated by ($\mathrm{D_{Dup}}$,$\mathrm{P_{Dup}}$,$\mathrm{D_{Del}}$,$\mathrm{P_{Del}}$) with $\mathrm{D_{Dup}}$ and $\mathrm{P_{Dup}}$ reflecting the proportions of deleterious and protective segments among causal duplication segments, and with $\mathrm{D_{Del}}$ and $\mathrm{P_{Del}}$ defined similarly for causal deletion segments.

Scenario II. Causal Dosage Effects. Scenario II.a (II.b) considers dosage effects from causal duplication (deletion) segments, and includes three settings of mixed deleterious and protective effects, i.e., (D,P)=(100,0), (50,50) and (10,90). The results are shown in Fig 4. As expected, the dosage-based CCRET kernel performs the best, with CONCUR following CCRET or having comparable power. Similar results are observed under Scenario II.c (Fig 5), where causal dosage effects are from both duplications and deletions and four varying mixtures of deleterious and protective effects are considered.

Fig 4. Power comparison between CONCUR, CCRET, and CKAT under Scenario II (causal dosage effects). The top panel shows results from Scenario II.a (causal duplication effects) and the bottom panel from Scenario II.b (causal deletion effects). In each sub-scenario, three different proportions of deleterious vs. protective effects are considered as indicated by (D,P), with D representing the proportions of deleterious segments and P the protective segments among causal segments.

Fig 5. Power comparison between CONCUR, CCRET, and CKAT for Simulation II.c (causal dosage effects from both duplications and deletions). Four different proportions of deleterious vs. protective effects are considered as indicated by (D_{Dup},P_{Dup},D_{Del},P_{Del}) with D_{Dup} and P_{Dup} reflecting the proportions of deleterious and protective segments among causal duplication segments, and with D_{Del} and P_{Del} defined similarly for causal deletion segments.

Real data application

In real data applications, we first, as a proof of concept, applied the proposed CONCUR test on a previously analyzed CNV dataset from the Swedish Schizophrenia Study. We next conducted a CNV-triglyceride (TG) association analysis using CONCUR on data from the Taiwan Biobank.

CNV analysis on schizophrenia in the Swedish Schizophrenia Study

We conducted pathway-based CNV analysis on data from the Swedish Schizophrenia Study [14]. The Swedish Schizophrenia Study used a case-control sampling design. Genotyping was done in six batches using Affymetrix 5.0 (3.9% of the subjects), Affymetrix 6.0 (38.6%), and Illumina OmniExpress (57.4%). PennCNV [11] was used to generate CNV calls. After quality control, we obtained a high quality rare CNV (frequency $< 1\%$ and size > 100kb) dataset in 8,547 subjects (3,637 cases and 4,820 controls) [15]. All procedures were approved by ethical committees at the Karolinska Institutet (Dnr No. 04/-449/4 and No. 2015/2081-31/2) and University of North

Carolina (No. 04-1465 and No. 18-1938). All subjects provided written informed consent (or legal guardian consent and subject assent). Previous analyses of this data [15] indicated significant associations of large rare CNVs with schizophrenia risk for both genome-wide dosage effects and gene intersecting effects of selected gene sets.

To evaluate the practical utility of the three kernel-based tests, we performed analysis on the gene sets previously examined in [6], excluding the PSD pathway as it overlaps the other three PSD-related pathways considered. In the eight gene sets, large (> 500kb) rare CNVs were found to be associated with schizophrenia by Szatkiewicz et al. [15], and these associations were corroborated by Tzeng et al. [6] in a gene-interruption analysis with CNVs > 100kb. In each pathway analysis, we performed association tests for joint dosage and length effects of rare CNVs > 100kb, using a fixed effect term to adjust for batch effects. CONCUR and CKAT kernels were constructed from the raw PLINK data and the CCRET dosage kernel was created using the functions available on the CCRET website. For CKAT, we used pathways as the CNVR unit instead of chromosomes because there were multiple chromosomes with only one gene. The results were evaluated against a Bonferroni-adjusted threshold of $0.05/8 = 0.00625$.

Table 2. Association test results for the effects of CNVs with > 100kb in length on schizophrenia risk in the Swedish Schizophrenia Study. Pathways are ordered by the number of tests that found significance (3 tests, 2 tests, 1 test) and then by pathway name. Significant p-values (at threshold 0.05/8=0.00625) are shown in bold.

Gene-sets				P-values		
Gene-set Name	# Genes	# Genes Interrupted in Cases	# Genes Interrupted in Controls	*CONCUR*	*CCRET*	*CKAT*
FMRP targets (Darnell et al. [16])	810	149	152	**2.29E-05**	**0.00044**	**0.00026**
PSD/PSD-95 (Kirov et al. [17])	65	13	10	**0.00052**	**0.00144**	0.00903
Synaptic Proteome (G2Cdb)	1023	121	106	**0.00067**	**0.00010**	0.00736
Cytoplasm (Kirov et al. [17])	266	28	32	**0.00124**	0.01408	**0.00030**
Mental Retardation	503	67	63	**0.00164**	0.10200	**0.00350**
PSD/mGluR5 (Kirov et al. [17])	38	4	7	**0.00040**	0.10540	**0.00129**
PSD/NMDAR (Kirov et al. [17])	61	12	12	**0.00102**	0.00922	**0.00046**
Synaptic genes (Ruano et al. [18])	718	154	164	**5.45E-06**	0.02005	0.00766

CONCUR found significant associations in all pathways, while CCRET and CKAT had alternating significance in some of the pathways (Table 2). In the FMRP pathway, all three tests were significant, and in the remaining seven gene sets, one or both of CCRET and CKAT were significant or near significant. The analyses suggest significant

CNV effects from dosage and/or length affecting schizophrenia risk, and the relative performance of these methods suggest some implications about the underlying effect patterns. CKAT, which is more sensitive to dosage-length interactive effects, found slightly more and different significant associations compared to CCRET, which is more sensitive to dosage effects, while CONCUR appeared to be more encompassing. We also observed stronger power of CKAT in the analysis here compared to the power observed in the simulation studies, which may partially be due to the lack of multiple testing penalty here.

CNV analysis on triglycerides in the Taiwan Biobank

We applied the proposed CONCUR test to the Taiwan Biobank (TWB) data `https://www.twbiobank.org.tw/new_web/` and conducted CNV association analysis with triglyceride (TG) levels on lipid-related pathways. The nationwide biobank project was initiated in 2012 and has recruited more than 15,995 individuals. The study has been approved by the ethical committee at Taichung Veterans General Hospital (IRB TCVGH No. CE16270B-2). The consent was not obtained because the data were analyzed anonymously. Peripheral blood specimens were extracted from healthy donors and genotyped using the Affymetrix Genomewide Axiom TWB array, which was designed specifically for a Taiwanese population. The TWB array contains 653,291 SNPs and was used to generate calls for genome-wide CNVs in the following process. First, Affymetrix Power Tools version 1.18.0 was used to produce a summary file of the intensity values of all probes, and the file was input into the Partek Genomic Suite version 6.6 to call CNVs based on the following criteria: at least 35 consecutive SNP markers, p-values of different CN values between two consecutive segments < 0.001, and signal-to-noise ratio (SNR) ≥ 0.3. A duplication was called if its copy number was > 2.3, whereas a deletion was called if its copy number was < 1.7. Several previous studies [19] [20] have demonstrated appropriate CNV calls with these parameters. After quality control, we obtained CNV data in 14,595 unrelated individuals. Our CNV association analyses focused on a subset of 11,664 individuals who had non-missing TG levels.

We referenced the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway database [21] to identify lipid-related pathways. Among the 17 pathways related to

Table 3. Association test results for the effects of CNVs on triglyceride levels in the Taiwan Biobank. Pathways are ordered by the number of tests that found significance (3 tests, 2 tests, 1 test, and no tests) and then by pathway name. Significant p-values (at threshold of 0.05/15 = 0.00333) are shown in bold.

Gene-sets			P-values		
Gene-set Names	# Genes	# Genes Interrupted	*CONCUR*	*CCRET*	*CKAT*
hsa00120 (Primary acid bile biosynthesis)	17	17	**0.00019**	**0.00314**	**0.00274**
hsa00061 (Fatty acid biosynthesis)	13	12	**0.00171**	0.01187	**0.00197**
hsa00140 (Steroid hormone biosynthesis)	60	58	**0.00030**	0.00623	**0.00159**
hsa00564 (Glycerophospholipid metabolism)	97	86	**0.00322**	0.00398	**0.00209**
hsa00590 (Arachnidonic acid metabolism)	63	62	**0.00212**	0.00883	**0.00211**
hsa00591 (Linoleic acid metabolism)	29	29	**0.00080**	0.01799	**0.00291**
hsa01040 (Biosynthesis of unsaturated fatty acids)	27	23	**0.00012**	0.00394	**0.00158**
hsa00062 (Fatty acid biosynthesis)	30	26	**0.00031**	0.01591	0.00508
hsa00072 (Synthesis and degradation of ketone bodies)	10	10	**0.00008**	0.00459	0.00383
hsa00561 (Glycerolipid metabolism)	61	50	0.00430	0.00494	**0.00198**
hsa00565 (Ether lipid metabolism)	47	43	**0.00018**	0.00859	0.00439
hsa00592 (alpha-Linolenic acid metabolism)	25	25	0.00581	0.00927	**0.00273**
hsa00071 (Fatty acid degradation)	44	43	0.00406	0.01088	0.00631
hsa00100 (Steroid biosynthesis)	19	16	0.01641	0.00618	0.00906
hsa00600 (Sphingolipid metabolism)	47	43	0.00382	0.00512	0.00789

"Lipid metabolism", 15 pathways included genes intersected by the TWB CNV data and were selected. For each pathway we performed the CONCUR test, CCRET, and CKAT. We adjusted for sex, age, BMI, and the top 10 principal components representing the population structure as covariates with fixed effects. As before, CKAT was performed with each pathway comprising a single CNVR. We compared the test results to a Bonferroni threshold of $0.05/15 = 0.00333$.

Out of the 15 pathways, ten pathways were identified as significantly associated with TG by CONCUR, nine pathways by CKAT, and one pathway by CCRET (Table 3). There were a total of 12 pathways found significant by one or more methods, among which one pathway, hsa00120 (primary bile acid biosynthesis), was significant for all methods. Compared to the Swedish Schizophrenia Study analysis, CCRET suffered from lower power and CKAT showed greater power, while the performance of CONCUR was relatively stable. The power loss in CCRET might be due to more dominant length or dosage×length signals and perhaps also a consequence of the stricter significance threshold here. CKAT demonstrated much better power than in the simulation study, which is likely attributable to the treatment of each pathway as a CNVR and hence the absence of multiple testing adjustment needed for multiple CNVRs. However, although CONCUR and CKAT were significant in a roughly equal number of pathways (ten versus nine, respectively), the CONCUR p-values tended to be much smaller than the CKAT p-values. To illustrate, if a more stringent significance threshold was adopted to adjust for the total of 45 tests (15 pathways × 3 methods) at a Bonferroni threshold of $0.05/45{=}0.0011$, then CONCUR would maintain significance in seven pathways while no CKAT p-value would meet the threshold. This behavior somewhat echoes the performance of CKAT in the simulation study.

The relative performance of CONCUR, CKAT and CCRET seems to suggest that CNV length or dosage×length effects dominate in the 12 significant pathways. To illustrate possible CONCUR post hoc analyses so to probe the potential sources of the pathway-level signal, we looked more closely at one pathway, hsa01040 (biosynthesis of unsaturated fatty acids), for which both CONCUR and CKAT were significant while CCRET was borderline significant. Previous studies have reported that monounsaturated fat acids or polyunsaturated fatty acids can effect TG levels [22, 23]. Given the major function of the genes in hsa01040 (i.e., the biosynthesis of unsaturated fatty acids), it is not unexpected that CNVs in these genes were significantly associated with TG levels. We calculated summary statistics describing CNV length and dosage in hsa01040 for individuals with different levels of TG. Based on the TG quantiles from the sample data, we classified individuals as having high TG (>75th percentile [>140 mmHg]), medium TG (25th−75th percentile [68-140 mmHg]) and low TG (<25th percentile [<68 mmHg]). We applied ANOVA to detect differences in CNV length and in dosage characteristics, and applied chi-squared tests to assess differences in the proportion of individuals with CNVs across TG levels. In addition, we examined CNV features in all CNVs together and in duplications and deletions separately.

Table 4. Descriptive statistics for hsa01040 pathway. TG values are classified as Low (<the 25th percentile [<68 mmHg]; n=2,931), Medium (the middle 50% [68 - 140 mmHg]; n=5,844), and High (>the 75th percentile [>140 mmHg]; n=2,889). The percent of individuals with CNVs is with respect to the total number of individuals in each TG category. The mean number of CNVs per individual and mean total length of CNVs (bp) per individual are reported, as well as the mean lengths (bp) and mean dosage per CNV. "Promising" associations with TG are marked with $\star\star$ to indicate p-value< 0.01 and with $\star$ to indicate p-value< 0.05.

CNV Type	TG Level	Pct Individuals with CNV	Mean # CNVs per Individual	# Genes Interrupted	Mean Total CNV Length per Individual (bp)	Mean CNV Length (bp)	Mean CNV Dosage
All	Low	6.18%	3.33	23	25143.71	2433.58	1.63
	Medium	6.07%	3.52	23	24447.30	2473.48	1.63
	High	7.17%	3.48	23	31091.65	2471.43	1.64
Deletion	Low	2.8%	5.84	16	29630.62	2590.55*	1.41
	Medium	2.74%	6.24	17	28107.36	2593.05*	1.40
	High	3.32%	5.79	16	32039.63	2067.96*	1.39
Duplication	Low	3.62%	1.17	20	7811.20	1827.24**	2.50*
	Medium	3.54%	1.22	23	10009.61	2001.81**	2.52*
	High	4.15%	1.38	22	27897.23	3831.02**	2.49*

Taking p-values < 0.05 as a suggestive "promising" association with TG, we did not observe any CNV associations when all CNVs were analyzed together, but for duplications only, there were promising differences in CNV length (p-value=0.0063) and weaker differences in dosage (p-value=0.0255) across TG levels. There were also some weak significance in CNV length for deletions (p-value=0.0423). We were cautious to not over-interpret these "promising" associations since this stratified analysis reflected only marginal associations of a CNV feature, and the tests did not account for the effect heterogeneity that motivates the application of kernel-based methods. We also proceeded with testing using CONCUR on duplications and deletions separately, and found a very significant association with TG in duplications (p-value $< 1 \times 10^{-8}$) and a weaker signal in deletions (p-value=0.0313).

To further explore the signal from duplications, we visualized CNVs in the 23 genes in hsa01040 (Fig 6). Fig 6 displays duplications and deletions in the CNV profiles of individuals categorized by their TG level (low, medium, and high), with profiles clustered so that shared patterns across profiles become apparent. For exploration purposes, we applied CONCUR to duplications in each gene and found that several genes had strong association p-values (i.e., $< 10^{-4}$), *BAAT, ELOVL4, ELOVL6, ELOVL5, HSD17B4*, and *SCD5* (S1 Table). Notably, *BAAT* is an amino acid N-acyltransferase for bile acid. Previous studies have demonstrated that bile acids are important regulators for TG level through crosstalk with farnesoid X receptor (FXR) [24, 25]. Since conversion of cholesterol to bile acid is an essential step in preventing the accumulation of TG, copy number duplications in *BAAT* may directly affect TG levels in the blood. Three *ELO* genes had significant CNV associations. Since the major functions of these genes focus on the elongation of fatty acids, CNV events in these genes are likely to affect the production and metabolism of TG. For example, one study showed that hepatic steatosis was observed in *ELOVL5*-knockout mice due to the activation of SREBP-1c and its target genes [30]. *HSD17B4* is a dehydrogenase, which is able to inhibit the production of DHEA [26]. A previous study showed that TG levels were inversely correlated to DHEA levels in men with type 2 diabetes [27], suggesting a potential link between CNVs in *HSD17B4* and TG levels. *SCD5* serves as a critical enzyme providing a double bond to construct complex lipid molecules such as TG [28, 29], and thus dysregulation of *SCD5* expression may impact TG levels. Further analyses are required to formally localize the sources of the CNV association signal in this pathway and others, but this exploratory analysis nonetheless serves to enrich our understanding of the association in pathway hsa01040 through examination of CNV-level and gene-level features.

Fig 6. Visualization of CNV activity in pathway hsa01040 by level of triglycerides (TG). CNV activity in genes in hsa01040 is shown by level of TG (Low, Medium, and High), with duplications in red and deletions in blue. Columns represent clustered individuals, and genes shown here are the 23 genes in the pathway that contain CNVs, ordered by the number of CNVs contained therein.

Discussion

We introduce CONCUR to leverage the strength of kernel-based methods to access the collective effects of rare CNVs on disease risk and incorporate several desired features. First, CONCUR permits the quantification of CNV similarity in an CNVR-free manner, avoiding the need of arbitrarily defining CNVRs as in current practice. Second, CONCUR incorporates both length and dosage information via the cAUC kernel, and is capable of detecting dosage, length and length-dosage interaction effects. Third, as the technology for detecting smaller CNVs improves, we expect to observe more length variation in CNVs and an increasing need to accommodate length effects in CNV association studies. However, there exist shortcomings in the standard kernel choices for handling CNV length. For example, a linear (or polynomial) kernel, which scores length similarity in a multiplicative fashion, cannot always reflect the true level of length similarity between an individual pair, e.g., a pair of CNVs of length 20kb would be equally similar to two CNVs with lengths 1kb and 400kb (as $20\times20 = 1\times400$). The alternative, e.g., Gaussian kernel as in CKAT, would still require a pre-specified scaling factor. CONCUR addresses these issues by using the common AUC of the CN profile curves of an individual pair and quantifies CNV similarity in dosage and length simultaneously. Finally, unlike current kernel methods, which require discretized copy numbers, CONCUR is directly applicable to continuous and discrete copy numbers. We provide the R functions that perform the CONCUR test at `https://www4.stat.ncsu.edu/~sthollow/JYT/CONCUR/`.

CONCUR shares some philosophy with several CNV analysis strategies in the literature. For example, Aguirre et al. [31] characterized the copy number changes in the pancreatic adenocarcinoma genome by detecting the minimum common regions (MCR) of recurrent copy number changes across tumor samples and using MCRs to prioritize genes that might be involved in pancreatic carcinogenesis. Harada et al. [32] also examined the minimal overlapping/common regions of frequent CNV activities among pancreatic cancer samples and among normal samples to identify candidate regions that might contain critical oncogenes or tumor suppressor genes. Furthermore, Mei et al. [33] proposed algorithms for identifying common CNV regions across individuals of homogeneous phenotypes for downstream association analysis. Built on similar concepts to these "common regions", CONCUR quantifies CNV similarity between sample pairs based on the "size" of the common regions as reflected in congruent location and dosage, and provides an association test to evaluate dosage and length effects.

In the analyses performed in this study, we calculated the cAUC using CNV dosage values transformed by the functions $a^{Dup}(DS) = (DS - 2)$ for duplications and 0 otherwise, and $a^{Del}(DS) = (2 - DS)$ for deletions and 0 otherwise. That is, we used copy number 2 as a reference value, and defined CNV similarity as the overlapping CNV length scaled linearly according to the magnitude of dosage deviation from the reference value. However, CONCUR can be flexibly extended to accommodate other schemes of quantifying common area by adopting different $a(\cdot)$ functions in the calculation of the cAUC. For example, instead of a linear scaling with $a^{\bullet}(DS) = |DS - 2|)$, one may consider a non-linear scaling by setting $a^{\bullet}(DS) = |DS - 2|^d$, with $d < 1$ deflating and $d > 1$ enhancing the contributions of CNVs of more extreme gains/losses. Additionally, one can impose reference values other than 2, such as using 2.3 for duplications (e.g., by setting $a^{Dup}(DS) = (DS - 2.3)$ for duplications and 0 otherwise), and using 1.7 for deletions (e.g., by setting $a^{Del}(DS) = (1.7 - DS)$ for deletions and 0 otherwise). Finally, overlapping area may be further weighted by inverse frequencies when needed, to augment the contribution of overlap in regions of rare CNV activity or of CNVs with uncommon dosage.

Materials and methods

CONCUR method

For individual i, $i = 1, \cdots, n$, denote Y_i the phenotype of individual i. Codify the CNV information in matrix Z_i with dimension $P_i \times 4$ as in the standard PLINK format of CNV data, where P_i is the number of CNVs that individual i has, and each row of Z_i records four features of CNV p, $p = 1, \cdots, P_i$: dosage (denoted as DS_p), chromosome (denoted CHR_p), start location (denoted as $BP1_p$), and end location (denoted as $BP2_p$). The dosage DS_p can be integer or continuous values. Finally let $X_i = (X_{i1}, \cdots, X_{ir})^T$ be the r covariates. Under the kernel machine regression framework, we model the association between phenotypes and CNVs as follows

$$g(\mu_i) = \beta_0 + X_i^T \beta_X + h(Z_i), \tag{2}$$

where $\mu_i = E(Y_i|X_i, Z_i)$, $g(\cdot)$ is the canonical link, and $h(Z_i)$ is an unknown smooth function of the variant features characterized by a kernel function $K(\cdot, \cdot)$. For continuous responses, $g(\mu_i) = \mu_i$; for binary responses, $g(\mu_i) = \log[(\mu_i/(1 - \mu_i)]$.

Profile curves

The proposed cAUC kernel is built on the concept of a CN profile curve. For a given chromosome $k = 1, 2, \cdots, 22$ and individual $i = 1, 2, \cdots, n$, we conceptualize a function $f_{ik}^{CN}(x)$ which returns the copy number of a CNV if x falls in a CNV and returns 2 (i.e., no CNV events) otherwise, e.g., examples shown in Fig 1. Given the CN profile curve, we further define two curves called the duplication profile curve and deletion profile curve, which recenter and rescale the CN values in CN profile curves through the "dosage transform functions" as described below, and allow us to compute cAUC similarity from duplications and from deletions in a more flexible manner.

We further use $q = 1, \cdots, P_{ik}$ to index the CNV features $(DS_q, BP1_q, BP2_q)$ occurring on chromosome k of individual i for $k = 1, \cdots, 22$. Then we construct duplication and deletion profile curves respectively describing duplications and deletions on chromosome k for individual i as follows:

$$f_{ik}^{Dup}(x) = \sum_{q=1}^{P_{ik}} I(BP1_q \leq x \leq BP2_q) a^{Dup}(DS_q) \tag{3}$$

$$f_{ik}^{Del}(x) = \sum_{q=1}^{P_{ik}} I(BP1_q \leq x \leq BP2_q) a^{Del}(DS_q) \tag{4}$$

$$\tag{5}$$

where x is a location on the genome on the same scale as $BP1_q$ and $BP2_q$; I is the indicator function such that $I(\cdot) = 1$ if the condition contained within is satisfied and equals 0 if otherwise; and $a^{\bullet}(DS)$ is a dosage transform function which determines the reference copy number value and controls how different copy number values contribute more or less to similarity in profiles. If an individual has no CNVs in chromosome k, then their duplication and deletion profile curves are identically equal to zero, i.e.,

$f_{ik}^{Dup}(x) = f_{ik}^{Del}(x) \equiv 0$ for all x. Although not explicitly shown, f_{ik}^{Dup} and f_{ik}^{Del} are functions of Z_i as the information of $DS_q, BP1_q, BP2_q$ and chromosome k for subject i is obtained from Z_i.

In this study, we designated $a^{Dup}(DS_q) = (DS_q - 2)$ if DS_q is from a duplication and 0 otherwise and $a^{Del}(DS_q) = (2 - DS_q)$ if DS_q is from a deletion and 0 otherwise. That is, for a given chromosome k and individual i, the function $f_{ik}^{Dup}(x)$ equals the magnitude of the duplication (i.e., number of additional copies compared to the reference copy number 2) for x inside a duplication and equals 0 otherwise, with analogous logic for $f_{ik}^{Del}(x)$. Other options of the dosage transform functions are described in the discussion section.

cAUC kernel

We propose to quantify the similarity between individuals i and j by comparing f_{ik}^{Dup} vs. f_{jk}^{Dup} and f_{ik}^{Del} vs. f_{jk}^{Del} over chromosomes $k = 1, \cdots, 22$ using the following kernel function

$$k_{cAUC}(Z_i, Z_j) = \sum_{k=1}^{22} \int_{\mathbb{N}} \left[\min\left(f_{ik}^{Dup}(x), f_{jk}^{Dup}(x)\right) + \min\left(f_{ik}^{Del}(x), f_{jk}^{Del}(x)\right)\right] \mathrm{d}\mu(x) \quad (6)$$

where $\min\left(f_{ik}^{\bullet}(x), f_{jk}^{\bullet}(x)\right)$ captures the minimum of the two functions evaluated at x and $\mu(x)$ is the counting measure. We refer to the kernel function as the cAUC kernel as it computes the minimal common area under the two individuals' duplication and deletion profile curves. The cAUC kernel function is a valid kernel as shown in S1 Appendix.

The intuition of the cAUC kernel is to quantify similarity using the length of overlapping CNVs between two individuals, with dosage information of the two overlapping CNVs determining how the overlapping length is scaled. The similarity between CNVs of different types (i.e., duplication vs. deletion) is 0, and the similarity between CNVs of the same type depends on the copy number values and the dosage transform function $a^{\bullet}(DS)$. For legal choices of $a^{\bullet}(DS)$, the similarity between a shared duplication (or deletion) event of larger magnitude will be higher than the similarity between a duplication of smaller magnitude, while the minimum operator enforces that the overlapping length is scaled by the CNV of smaller magnitude in a pair with different magnitudes.

Legal choices of $a^{\bullet}(DS)$ will upweight the contribution from similar CNVs of greater magnitude in duplication or deletion, which are often more rare and have higher impact. As proposed in the Discussion section, the family of dosage transform functions $a^{\bullet}(DS) = |DS - 2|^d$ provides a spectrum of weighting schemes, with $d < 1$ down-weighting and $d > 1$ upweighting the contribution of higher magnitude CNVs. Across copy number data of varying types and varying sample-level characteristics, the $a^{\bullet}(\cdot)$ dosage transform function allows for flexible scaling of dosage to appropriately customize the cAUC measure of similarity.

Association test

The association between phenotype and CNVs is examined by testing the hypothesis $H_0 : h(\cdot) = 0$. To do so, we define the vector of subject-specific CNV effects $H = (h(Z_1), \cdots, h(Z_n))$ and treat H as random effects which follow $N(0, \tau\mathbf{K})$, where $\tau \geq 0$ is a variance component and $\mathbf{K}$ is a $n \times n$ kernel matrix with its (i, j)th entry being $K(Z_i, Z_j)$. Following Liu et al. [34] [35], testing $H_0 : h(\cdot) = 0$ is equivalent to testing $\tau = 0$ under a generalized linear mixed model. As in [7] [6], we use a score-based test, which has the form of

$$T = \frac{(Y - \mu_0)\Delta\mathbf{WKW}\Delta(Y - \mu_0)}{2}\bigg|_{\tau=0, \mu_0=\hat{\mu}_0, \phi=\hat{\phi}} \tag{7}$$

where Y is $n \times 1$ vector of responses; $\mu_0 = E(Y)$ under H_0; ϕ is a dispersion factor parameterizing the variance of Y; $\Delta \in \mathbb{R}^{n\times n}$ is a diagonal matrix with its ith diagonal element being $\delta_i = 1/g'(\mu_i)$; $\mathbf{W} \in \mathbb{R}^{n\times n}$ is a diagonal weight matrix with its ith diagonal element being $w_i = [v(\mu_i)]^{-1}\delta_i^2$ where $v(\cdot)$ comes from $\text{Var}(Y_i) = v(\mu_i)\phi$ per the exponential dispersion family of probability density functions. The score statistic asymptotically follows a weighted chi-square distribution [34] [35]. Recently, Chen et al. [12] derived the corresponding small-sample distribution, which is used to calculate the p-value in this work.

Supporting information

S1 Table. Gene-level CONCUR tests on genes in pathway hsa01040.

S1 Appendix. Proof of symmetry and positive semi-definiteness of cAUC kernel.

Acknowledgments

References

1. McCarroll SA, Huett A, Kuballa P, et al. Deletion polymorphism upstream of IRGM associated with altered IRGM expression and Crohn's disease. Nat Genet. 2008;40(9):1107-12.

2. Liu S, Yao L, Ding D, et al. CCL3L1 copy number variation and susceptibility to HIV-1 infection: a meta-analysis. PLoS One. 2010;5(12):e15778.

3. Macé A, Tuke MA, Deelen P, et al. CNV-association meta-analysis in 191,161 European adults reveals new loci associated with anthropometric traits. Nat Commun. 2017;8(1):744.

4. Malhotra D, Sebat J. CNVs: Harbinger of a Rare Variant Revolution in Psychiatric Genetics. Cell. 2012;148(6):1223-41.

5. Raychaudhuri S, Korn JM, McCarroll SA, et al. Accurately assessing the risk of schizophrenia conferred by rare copy-number variation affecting genes with brain function. PLoS Genet. 2010;6(9):e1001097.

6. Tzeng JY, Magnusson PK, Sullivan PF, et al. A new method for detecting associations with rare copy-number variants. PLoS Genet. 2015;11(10):e1005403.

7. Zhan X, Girirajan S, Zhao N, et al. A novel copy number variants kernel association test with application to autism spectrum disorders studies. Biometrics. 2016;32(23):3603-3610.

8. Beekman M, Heijmans BT, Martin NG, et al. Two-locus linkage analysis applied to putative quantitative trait loci for lipoprotein(a) levels. Twin Res. 2003;6(4):322-4.

9. Pedersen NL, Lichtenstein P, Svedberg P. The Swedish Twin Registry in the Third Millennium. Twin Res. 2002;5:427-32.

10. Lichtenstein P, Bjork C, Hultman CM, et al. Recurrence risks for schizophrenia in a Swedish national cohort. Psychol Med. 2006;36(10):7-26.

11. Wang K, Li M, Hadley D, et al. PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. Genome Res. 2007;17(11):1665-74.

12. Chen J, Chen W, Zhao N, et al. Small Sample Kernel Association Tests for Human Genetic and Microbiome Association Studies. Genet Epidemiol. 2016;40(1):5-19.

13. Davies RB. Algorithm AS 155: The Distribution of a Linear Combination of chi-2 Random Variables. Journal of the Royal Statistical Society Series C (Applied Statistics). 1980;29(3):323-33.

14. Ripke S, O'Dushliane C, Cahmbert K, et al. Genome-wide association analysis identifies 13 new risk loci for schizophrenia. Nat Genet. 2013;45(10):1150-9.

15. Szatkiewicz JP, O'Dushlaine C, Chen G, et al. Copy number variation in schizophrenia in Sweden. Mol Psychiatry. 2014 July;19(7):762-773.

16. Darnell JC, Van Driesche SJ, Zhang C, et al. FMRP stalls ribosomal translocation on mRNAs linked to synaptic function and autism. Cell 2011;146(2):247-261.

17. Kirov G, Pocklington AJ, Holmans P, et al. De novo CNV analysis implicates specific abnormalities of postsynaptic signalling complexes in the pathogenesis of schizophrenia. Mol. Psychiatry. 2012 Feb;17(2):142-153.

18. Functional gene group analysis reveals a role of synaptic heterotrimeric G proteins in cognitive ability. Ruano D, Abecasis GR, Glaser B, et al. Am J Hum Genet. 2010;86(2):113-125.

19. Lu TP, Lai LC, Tsai MH et al. Integrated analyses of copy number variations and gene expression in lung adenocarcinoma. PloS One. 2011;6(9):e24829.

20. Lai LC, Tsai MH, Chen PC, et al. SNP rs10248565 in HDAC9 as a novel genomic aberration biomarker of lung adenocarcinoma in non-smoking women. J BiomedSci. 2014;21(1):24.

21. Kanehisa M, Sato Y, Kawashima M, et al. KEGG as a reference resource for gene and protein annotation. Nucleic Acids Res. 2016;44(D1):D457-462.

22. Grundy SM. Monounsaturated fatty acids and cholesterol metabolism: implications for dietary recommendations. J Nutr. 1989;119(4):529-533.

23. Ooi EM, Watts GF, Ng TW, et acl. Effect of dietary Fatty acids on human lipoprotein metabolism: a comprehensive update. Nutrients. 2015;7(6):4416-25.

24. Lien F, Berthier A, Bouchaert E, et al. Metformin interferes wtih bile acid homeostasis through AMPK-FXR crosstalk. J Clin Invest. 2014;124(3):1037-1051.

25. Watanabe M, Houten SM, Wang L, et al. Bile acids lower triglyceride levels via a pathway involving FWR, SHP, and SREBP-1c. J Clin Invest. 2004;112(10):1408-18.

26. de Launoit Y, Adamski J. Unique multifunctional HSD17B4 gene product: 17beta-hydroxysteroid dehydrogenase 4 and D-3-hydroxyacyl-coenzyme A dehydrogenase/hydratase involved in Zellweger syndrome. J Mol Endocrinol. 1999;22(3):227-40.

27. Boudou P, de Kerviler E, Erlich D, et al. Exercise training-induced triglyceride lowering negatively correlates with DHEA levels in men with type 2 diabetes. Int J Obes Relat Metab Disord. 2001;25(8):1108-12.

28. Castro LF, Wilson JM, Gonçalves O, et al. The evolutionary history of the stearoyl-CoA desaturase gene family in vertebrates. BMC Evol Biol. 2001;11:132.

29. Flowers MT, Ntambi JM. Role of stearoyl-coenzyme A desaturase in regulating lipid metabolism. Curr Opin Lipidol. 2008;19(3):248-256.

30. Sassa T, Kihara A. Metabolism of very long-chain Fatty acids: genes and pathophysiology. Biomol Ther (Seoul). 2014;22(2):83-92.

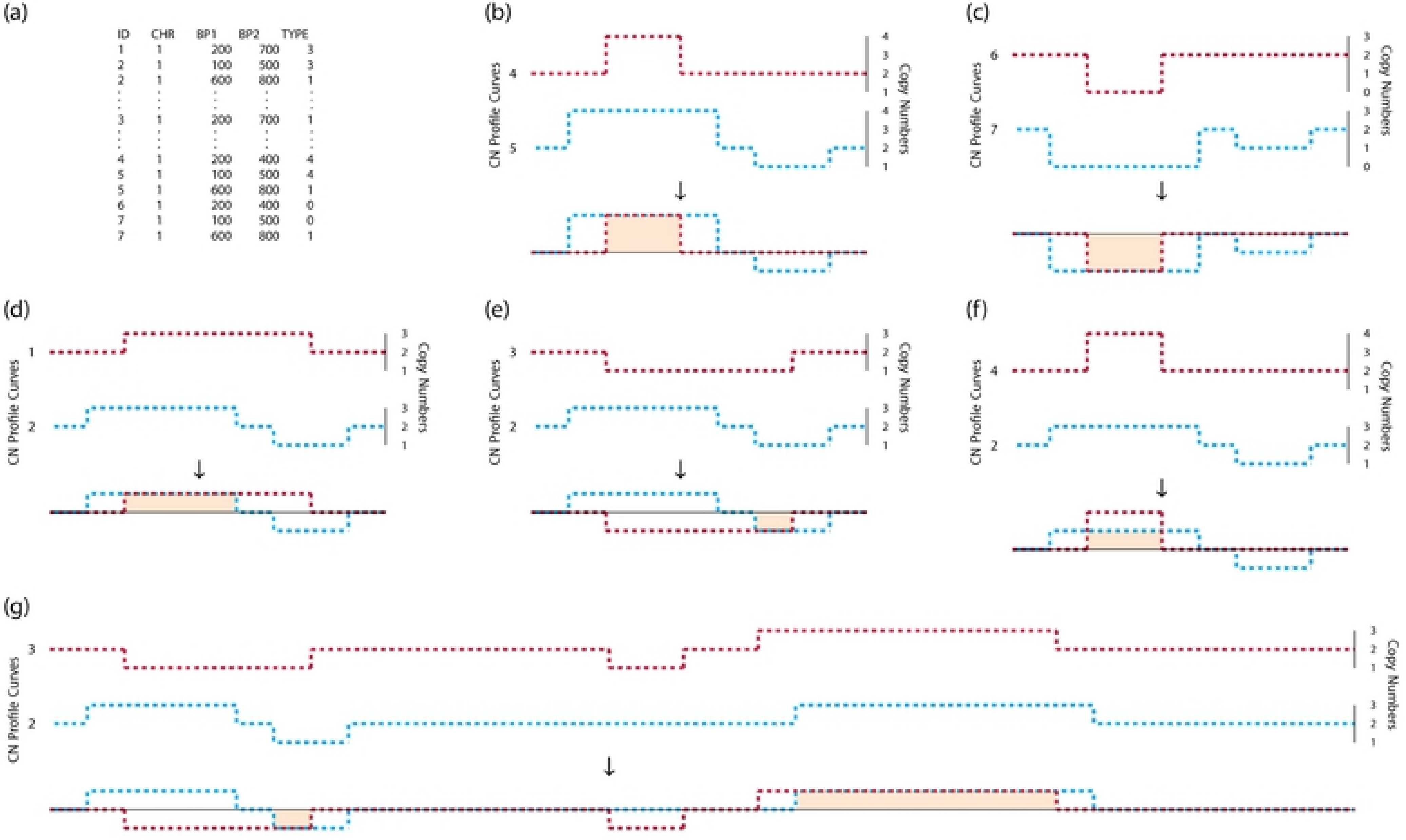
31. Aguirre AJ, Brennan C, Bailey G, et al. High-resolution characterization of the pancreatic adenocarcinoma genome. Proc Natl Acad Sci USA. 2004;101(24):9067-9072.

32. Harada T, Chelala C, Bhakta V, et al. Genome-wide DNA copy number analysis in pancreatic cancer using high-density single nucleotide polymorphism arrays. Oncogenomics. 2008;27(13):1951-1960.

33. Mei TS, Salim A, Calza S, et al. Identification of recurrent regions of Copy-Number Variations cross multiple individuals. BMC Bioinformatics. 2010;11:147.

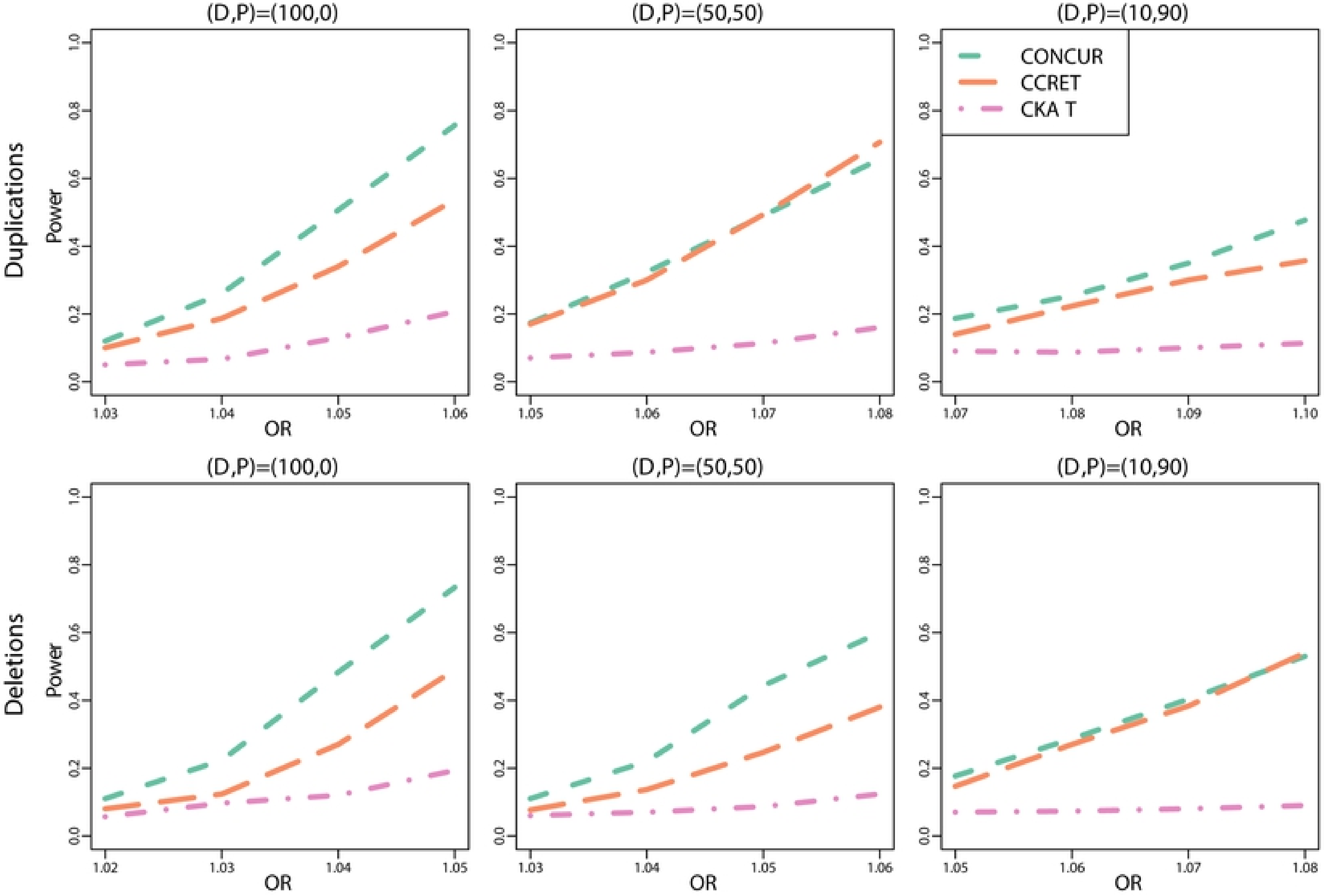
34. Liu D, Lin X, Ghosh D. Semiparametric Regression of Multidimensional Genetic Pathway Data: Least-Squares Kernel Machines and Linear Mixed Models. Biometrics. 2007;64(4):1079-1088.

35. Liu D, Ghosh D, Lin X. Estimation and testing for the effect of a genetic pathway on a disease outcome using logistic kernel machine regression via logistic mixed models. BMC Bioinformatics. 2009;9(1):292.



ID	CHR	BP1	BP2	TYPE
1	1	200	700	3
2	1	100	500	3
2	1	600	800	1
⋮	⋮	⋮	⋮	⋮
3	1	200	700	1
⋮	⋮	⋮	⋮	⋮
4	1	200	400	4
5	1	100	500	4
5	1	600	800	1
6	1	200	400	0
7	1	100	500	0
7	1	600	800	1

Figure1



(D,P)=(100,0)
(D,P)=(50,50)
(D,P)=(10,90)
CONCUR
CCRET
CKA T
Duplications
Deletions
Power
0.0
0.2
0.4
0.6
0.8
1.0
OR
1.03
1.04
1.05
1.06
1.07
1.08
1.09
1.10
1.02

Figure2

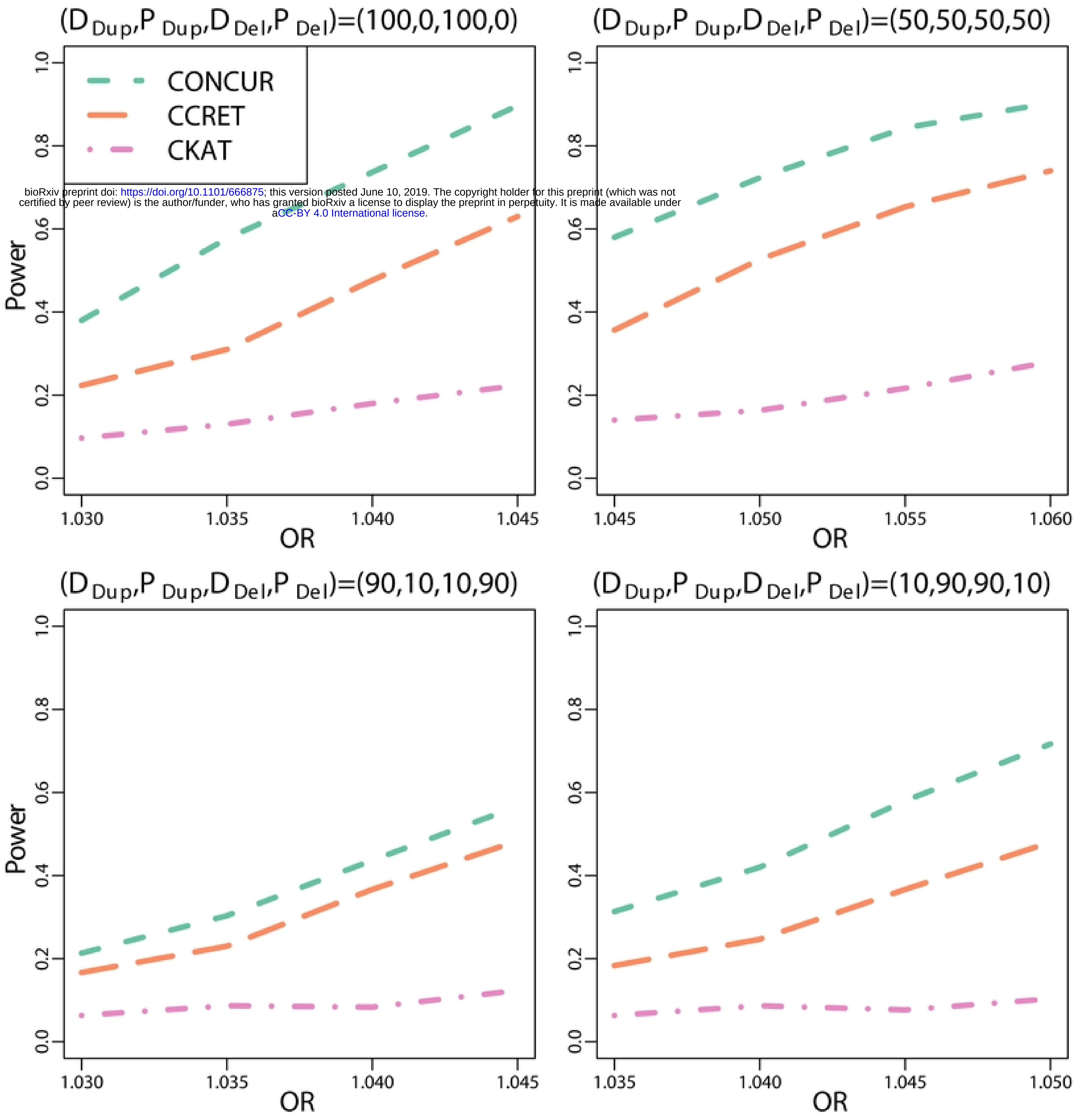

(D Dup,P Dup,D Del,P Del)=(100,0,100,0)
(D Dup,P Dup,D Del,P Del)=(50,50,50,50)
(D Dup,P Dup,D Del,P Del)=(90,10,10,90)
(D Dup,P Dup,D Del,P Del)=(10,90,90,10)
CONCUR
CCRET
CKAT
Power
0.0
0.2
0.4
0.6
0.8
1.0
OR
1.030
1.035
1.040
1.045
1.050
1.055
1.060

Figure3

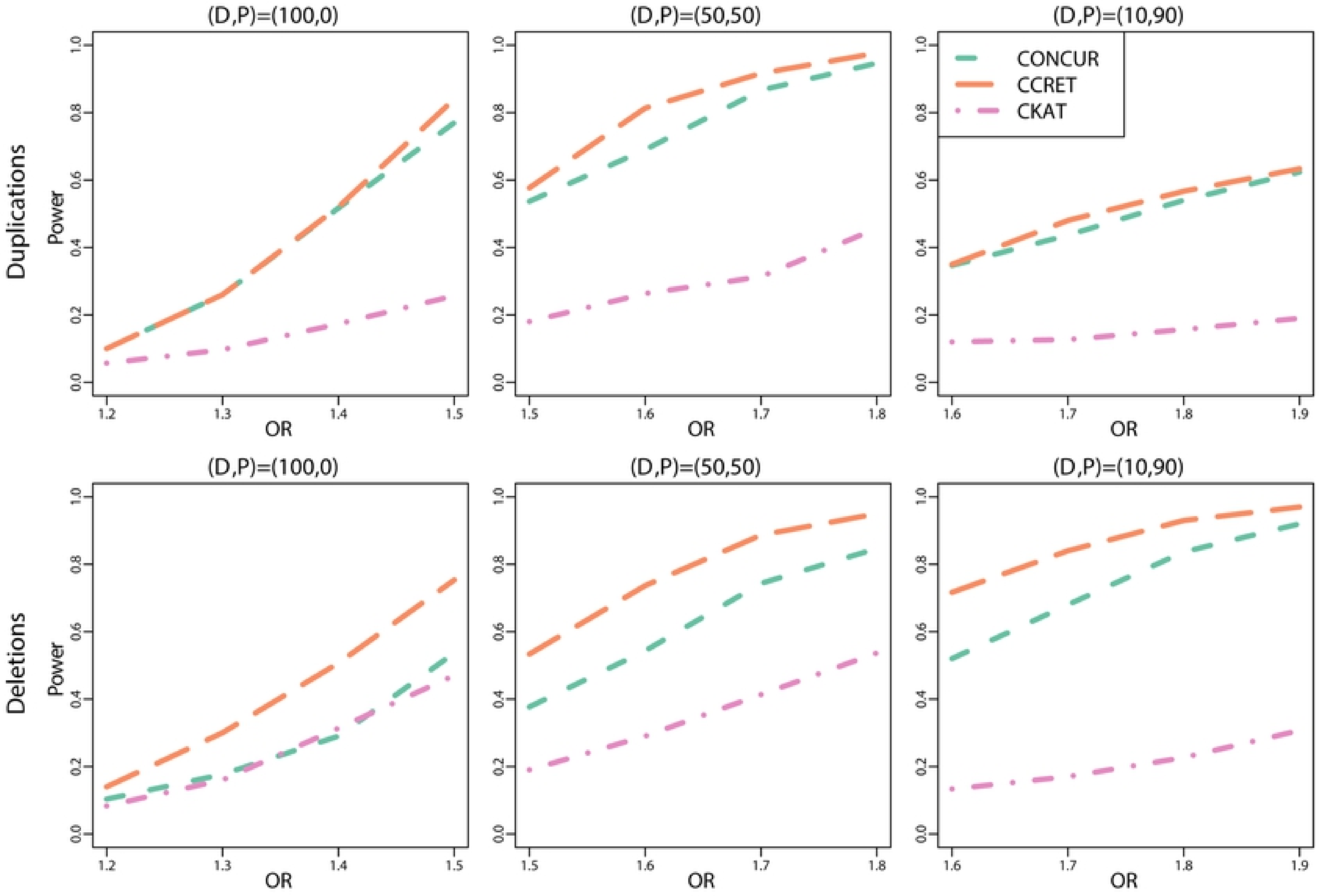

(D,P)=(100,0)
(D,P)=(50,50)
(D,P)=(10,90)
Duplications
Deletions
Power
OR
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.3
1.4
1.5
1.6
1.7
1.8
1.9
CONCUR
CCRET
CKAT

Figure4

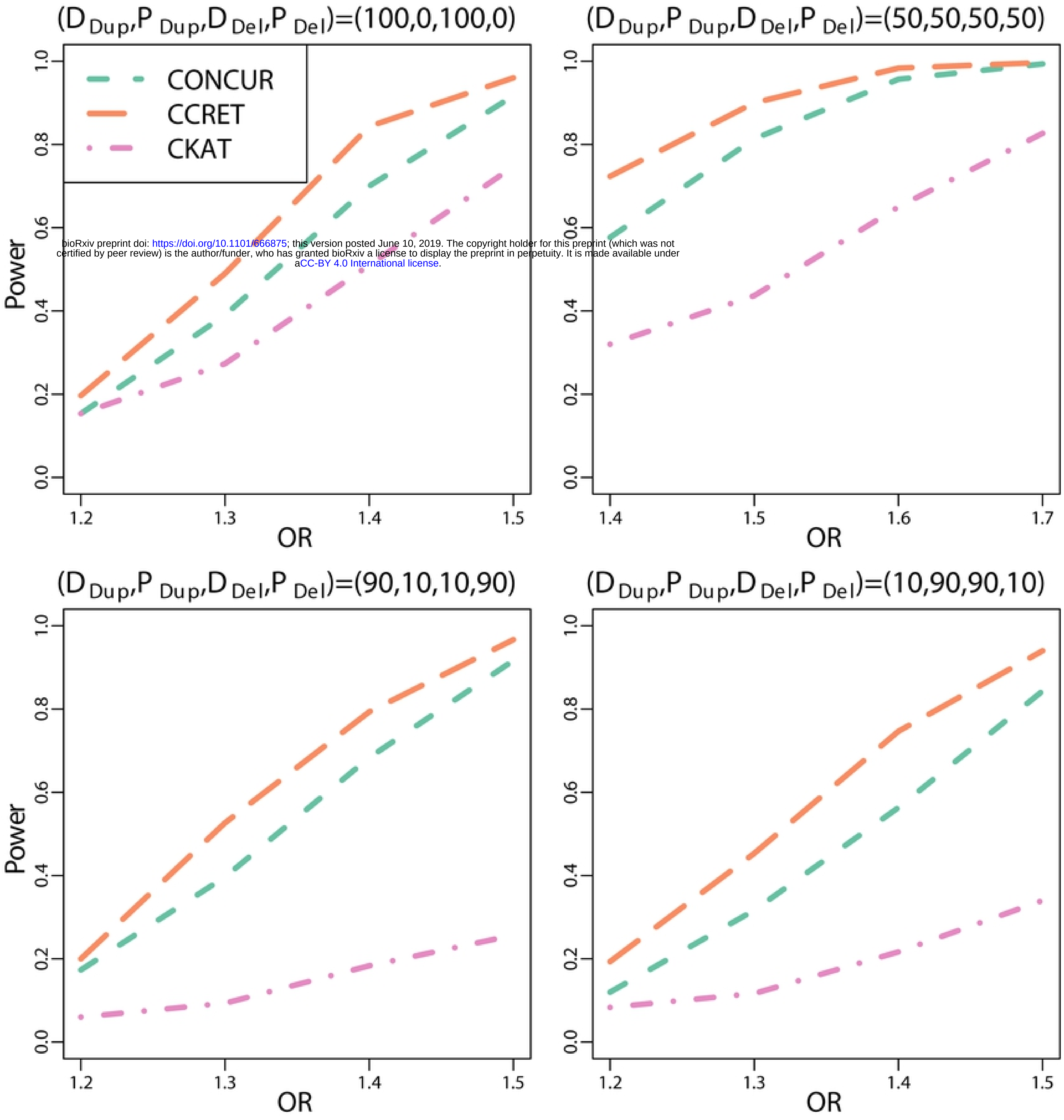

$(D_{Dup},P_{Dup},D_{Del},P_{Del})=(100,0,100,0)$
$(D_{Dup},P_{Dup},D_{Del},P_{Del})=(50,50,50,50)$
$(D_{Dup},P_{Dup},D_{Del},P_{Del})=(90,10,10,90)$
$(D_{Dup},P_{Dup},D_{Del},P_{Del})=(10,90,90,10)$
CONCUR
CCRET
CKAT
Power
0.0
0.2
0.4
0.6
0.8
1.0
OR
1.2
1.3
1.4
1.5
1.6
1.7

Figure5

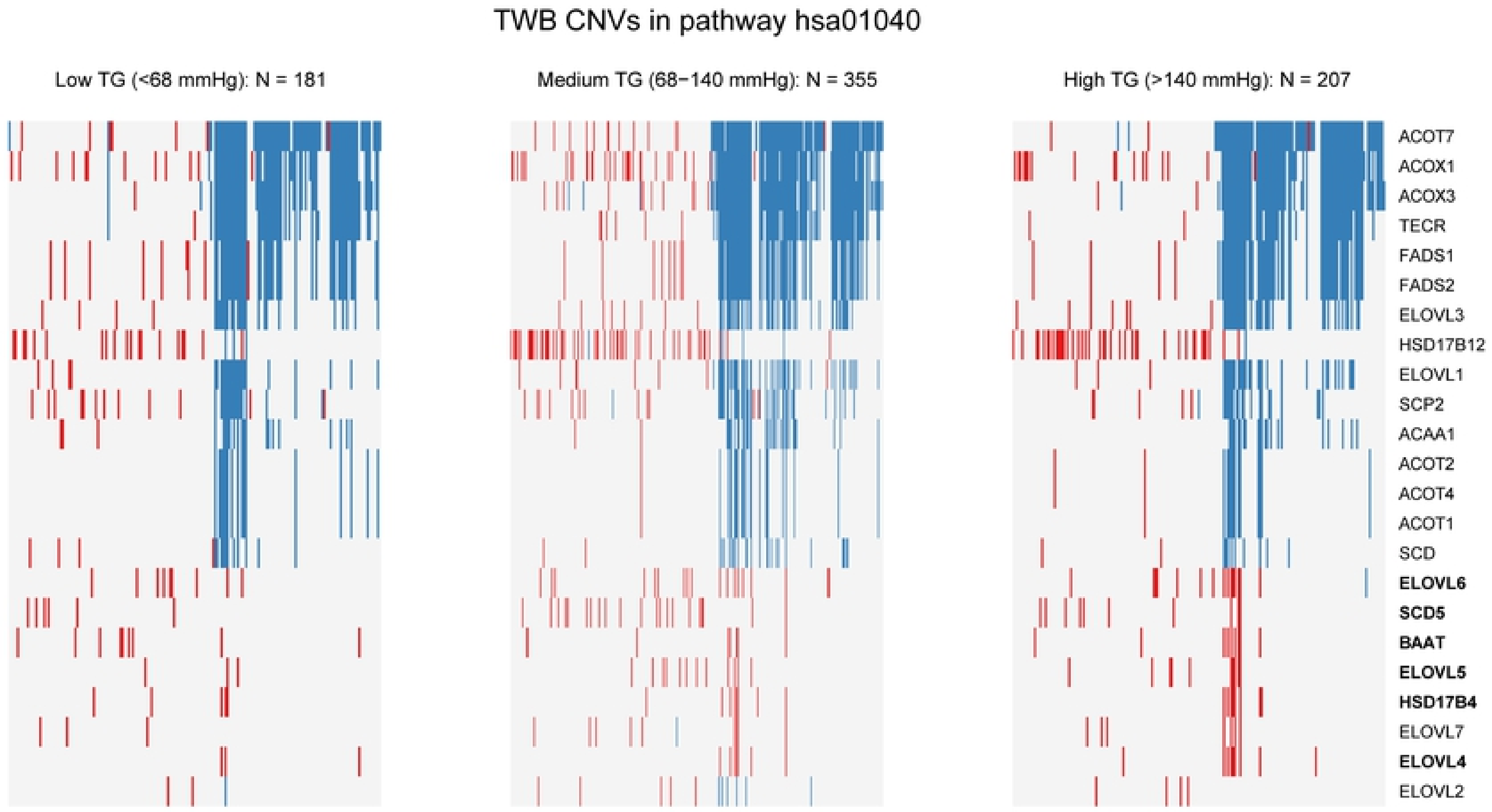

TWB CNVs in pathway hsa01040
Low TG (<68 mmHg): N = 181
Medium TG (68−140 mmHg): N = 355
High TG (>140 mmHg): N = 207
ACOT7
ACOX1
ACOX3
TECR
FADS1
FADS2
ELOVL3
HSD17B12
ELOVL1
SCP2
ACAA1
ACOT2
ACOT4
ACOT1
SCD
ELOVL6
SCD5
BAAT
ELOVL5
HSD17B4
ELOVL7
ELOVL4
ELOVL2

Figure6