Copy number variation profile-based genomic subtyping of premenstrual dysphoric disorder in Chinese

Hong Xue[1,2,3*], Zhenggang Wu[2,3], Xi Long[2], Ata Ullah[2], Si Chen[2], Wai-Kin Mat[2], Peng Sun[1], Ming-Zhou Gao[1], Jie-Qiong Wang[1], Hai-Jun Wang[1], Xia Li[1], Wen-Jun Sun[1], and Ming-Qi Qiao[1*]

[1]Shandong University of Traditional Chinese Medicine, Jinan, Shandong, People's Republic of China

[2]Division of Life Science, Hong Kong University of Science and Technology, Hong Kong, People's Republic of China

[3]School of Basic Medicine and Clinical Pharmacy, China Pharmaceutical University, Nanjing, Jiangsu, People's Republic of China

* Correspondence:

Hong Xue

hxue@ust.hk

Mingqi Qiao

qmingqi@163.com

Abstract

Premenstrual dysphoric disorder (PMDD) affects nearly 5% women of reproductive age. The symptomatic heterogeneity, along with largely unknown genetics, of PMDD have greatly hindered its effective treatment. In the present study, 127 Chinese PMDD patients of the ‘invasion’ and ‘depression’ subtypes clinically differentiated by us earlier were analyzed together with 108 non-PMDD controls for genome-wide copy number variations (CNVs). Germline genomic DNA samples from white blood cells were subjected to AluScan sequencing-based CNV profiling, which enabled clustering of patient samples readily into the V and D groups, dominated by the “invasion” and “depression” clinical subtypes, respectively; the CNVs obtained with 100-kb windows yielded two clusters that were correlated with these subtypes with a consistency of up to 89.8%. Diagnostic correlation- and frequency-based CNV features of either CNV-gain (CNVG) or CNV-loss (CNVL) that could differentiate between V and D subtypes were selected and analyzed. CNVG features located preferentially in S2-phase replicating regions and enriched with steroid hormone biosynthesis pathway of genes were found protective against PMDD. Moreover, machine learning employing the correlation-based CNV features could predict with >80% accuracy whether a genomic sample was D-type, V-type or control. In terms of their CNV profiles, the D- and V-types differed more from one another than from the controls, thereby providing a genomic basis for the clinical D-V subtyping of PMDD. Genome-wide profiling of CNVs, as a new approach to complex disease genetics, has revealed recurrent CNVs and genomic features beyond individual genes and mutations underlying PMDD clinical diversity.

Introduction

Premenstrual dysphoric disorder (PMDD) is a syndrome that afflicts 5-10% of women in their reproductive years (1). The severity of the syndrome is typically highest just before the menstruation period, suggesting that the symptoms were linked to hormonal changes. This has been confirmed by the findings of premenstrual neurosteroid fluctuations, and alterations in the sensitivity of $GABA_A$ receptors to neurosteroids giving rise to mood instability (2, 3). Cortical gamma-aminobutyric acid (GABA) levels also declined during the menstrual cycle in healthy women but increased in women with PMDD from the follicular phase to the mid-luteal and late luteal stages (4). Furthermore, PMDD has been associated with the estrogen receptor alpha gene *ESR1* (5), and the ESC/E(Z) genes affecting the interactions of sex hormones with other genes (6). Five major contributors to the etiology of PMDD include: (1) genetic susceptibility; (2) progesterone and its metabolite ALLO; (3) estrogen, serotonin and brain-derived neurotrophic factor (BDNF); (4) brain structure and function; and (5) the hypothalamic-pituitary-adrenal axis and hypothalamic-pituitary-gonadal axis (7). The schizophrenia-associated SNPs in *GABRB2*, located in introns 8 and 9 near an AluYi6 insertion, have been associated with both schizophrenia and bipolar disorder (8, 9), heroin addiction (10), altruism (11), autism and mental retardation (12). Deletion of gabrb2 genes from knockout mice also brought about schizophrenic symptoms that were alleviated by the antipsychotic Risperidone (13). Recently, analysis of germline copy-number-variations (CNVs) at the nsv1177513 site in Exon 11, and the esv2730987 site in Intron 6, of *GABRB2* in PMDD and schizophrenia patients showed that CNV alterations at both esv2730987 and nsv1177513 were significantly associated with schizophrenia in Chinese and Germans as well as PMDD in Chinese (14). Moreover, subjects with different levels of susceptibility to cancer could be distinguished by means of diagnostic CNV marker features selected from the germline genomes with the application of machine learning (15).

It is recognized that the symptoms of PMDD are consistent with multiple clinical subtypes. A Delphi survey led to the proposal of three symptoms-based types of PMDD, *viz.* a predominantly physical type, a predominantly emotional type, and a mixed type (16); and DSM-V proposed that PMDD is defined by one or more of the symptoms of marked affective lability, marked irritability or anger, marked depressed mood and hopelessness, and marked anxiety and tension, plus at least one of seven other symptoms. At the School of Basic Medicine, Shandong University of Traditional Chinese Medicine, the medical records also pointed to at least two major types of PMDD, viz. an irritability-marked 'invasion' type (58.9%) and a depressive mood-marked 'depression' type (27.5%) (17). In view of the spectrum of PMDD symptoms, the objective of the present study was to enquire whether the two major clinical subtypes of PMDD could be corelated with genomic profiles. Through genome-wide CNV profiling by AluScan next-generation sequencing (18, 19), the results revealed two large clusters of CNV profiles that were highly correlated with the clinical "depression" and "invasion" subtypes. Furthermore, CNV-gain (CNVG) and CNV-loss (CNVL) features diagnostic of PMDD or each of the two clinical subtypes were uncovered among CNVs called from sequence windows of different sizes, which were variously distributed in genomic regions of different replication timing and overlapped with genes in various genetic pathways of potential clinical relevance. These results provided genomic verification for the invasion- and depression-subtypes employed by us previously (17), which corresponded to part of the complex symptoms stipulated by DSM-V (20) as diagnostic criteria for PMDD.

Methods

Clinical assessments

Clinical diagnosis of PMDD patients (P-type subjects) from asymptomatic controls (C-type subjects) was performed in accordance to the protocol in Diagnostic and Statistical Manual of Mental Disorders (DSM-IV) by two psychiatrists independently. The identifications of ‘depression-type’ and ‘invasion-type’ subjects were carried out as previously described (17).

Genomic DNA samples

Peripheral white blood cell DNA samples were collected from PMDD patients and non-PMDD control subjects with approval by the institutional ethic committee of Shandong University of Chinese Medicine. The patients and healthy volunteers who participated in this study all signed the informed consent form. The samples consist of a control cohort of 108 subjects and a PMDD cohort of 127 cases. The latter cohort was further divided into the depression-subtype (71 cases) and invasion-subtype (56 cases). The subtypings of the 127 PMDD cases were given in Table S1.

AluScan sequencing and CNV calling

Samples of ~0.1μg DNA were subjected to inter-Alu PCR amplification using the four Alu-consensual primers AluY278T18, AluY66H21, R12A/267 and L12A/8 (18). The 200 bp to ~6 kb amplicons in each sample were employed to build a library for sequencing on the Illumina platform with 100 bp paired-end reads. According to the standard framework, all the reads were mapped to reference human genome hg19 downloaded from UCSC by BWA, followed by base recalibration and local realignment by GATK (21). CNVs were called from the AluScan sequences with the method of AluScanCNV2 (19, 22) based on sequence windows of 50-500 kb in 50-kb increments on the 22 autosomes and the X chromosome. The CNV profiles of all 108 control and 127 PMDD subjects were available in Table S2.

Clustering and grouping of patient samples based on CNV profiles

The profiles of CNVG and CNVL called from the 127 P-group samples using different CNV-calling window sizes were separately subjected to correlation analysis and hierarchical clustering with 1,000 bootstraps using the ‘pvclust’ R package (23). The derived correlation heatmaps as well as the CNVG-based and CNVL-based dendrograms obtained for each window size were employed to determine the two subgroups of CNV profiles using two different grouping methods for cross validation.

In the first method, *viz*. the straightforward ‘*cutree*’ method, the sub-clustering was carried out using the ‘*cutree*’ function from the ‘dendextend’ R package (24) to cut each dendrogram into 2-8 sub-clusters (Figure S1). The DNA samples located in the sub-cluster populated with the highest number of clinical depression-type samples among all the sub-clusters was referred as D-type genomic samples; and the DNA samples located in the remaining sub-clusters were combined and referred as V-type genomic samples. In the second, or ‘*semi-supervised*’ method, some branches on the dendrograms were first rotated around their respective nodes to bring the closely co-localized samples into tightly knit sub-clusters enclosed by black square boxes on the diagonal of each heatmap. Thereupon, all the samples within the same block box were all designated as D-type or V-type genomic samples depending on whether the majority clinical subtype of the samples were depression-type or invasion-type. The designated D- and V-type genomic samples derived using the two grouping methods for ten different window sizes are shown in Table S3 and exemplified by the blue and red branches in Figure 1 and Figure S2 respectively for the 100-kb CNV profiles.

For either the ‘*cutree*’ method or the ‘*semi-supervised*’ method, let the number of CNV-based D-type samples that also belonged to the clinical depression-subtype be represented by $True_D$, and the number of CNV-based V-type samples that also belonged to the clinical invasion-subtypes be represented by $True_V$. Accordingly, the consistency (Y) between CNV-based classification and the clinical classification of PMDD patient samples could be estimated by:

$$Y = (True_D + True_V)/127$$

On this basis, the levels of consistency between CNV-based and clinical subtypings for the different CNVs called using different window sizes for both the '*cutree*' and '*semi-supervised*' methods are shown in Table S4.

Selection of diagnostic CNV features

The selection of diagnostic CNV features was performed using either (a) correlation-based method or (b) frequency-based method as described (15). CfsSubsetEval from the Weka package was employed together with BestFirst search method to select the correlation-based diagnostic CNV features. Fisher's exact tests were employed to select the frequency-based CNV features that showed significantly different occurrence frequencies between a pair of sample groups (e.g. P-vs-C or D-vs-V) with a false discovery rate (FDR) less than 0.01.

Predictive subtyping of genomic samples by machine learning

Earlier, diagnostic germline CN-gains and CN-losses from leucocyte DNA samples of subjects with or without past episodes of cancers in tissues other than leucocytes were found to provide a useful basis to predict the propensity of the subject to cancer (15). Since the 127 P-group and 108 C-group DNA samples from PMDD and control subjects consisted of a mixture of D-type, V-type and C-type DNAs, the question arose whether it was possible to predict the typing of DNA samples between the D-vs-V, D-vs-C and V-vs-C choices employing the diagnostic CNVG and CNVL features obtained with the correlation-based method.

For example, in a choice between the P-vs-C types, a mixture of P- and C-type samples were randomly separated into a labeled Learning Band and an unlabeled Test Band, with equal or near equal number of samples in the two bands. Diagnostic CNVG and CNVL features were selected from the labeled Combined Learning Band with machine learning using the correlation-based method and employed to estimate the risk factor *R* for each DNA sample in the Test Band according to Eqn. 1.

$$R = \log\left(\frac{Pr(PMDD|Features)}{Pr(Control|Features)}\right) \qquad \text{Eqn. 1}$$

$$\Pr(PMDD|Features) = \Pr\,(Features|PMDD) \times \Pr\,(PMDD)/\Pr\,(Features)$$

$$\Pr(Control|Features) = \Pr\,(Features|Control) \times \Pr\,(Control)/\Pr\,(Features)$$

where Pr(PMDD|Features) was the posterior probability of membership in the PMDD group given the CNV data of a particular Test Band sample; Pr(Control|Features) was its posterior probability of membership in the Control group given the same test CNV data; Pr(Features|PMDD) was the likelihood function of the test CNV data given membership in the PMDD group; Pr(Features|Control) was the likelihood function of the test CNV data given membership in the Control group; Pr(PMDD) and Pr(Control) were the prior distributions of PMDD and Control samples respectively within the Learning Band; and Pr(Features) was the prior distribution of CNV-features among all the CNVs within the Learning Band.

For every sample in the Test Band, its value of R estimated using Eqn. 1 would predict whether the sample belonged to the Control group or PMDD group: it would predictively belong to Control group (*viz*. ‘non-PMDD’) if $R < 0$; belong to PMDD group if $R > 0$; or no prediction could be made if $R = 0$. For every PMDD sample in the Test Band, $R > 0$ represented a ‘true’ prediction whereas $R < 0$ represented a ‘not true’ prediction. On the other hand, for any Control sample in the Test Band, $R > 0$ represented a ‘not true’ prediction whereas $R < 0$ represented a ‘true’ prediction. Accuracy of prediction was therefore given by:

$$\text{Accuracy} = \frac{[\text{True predictions of Control}]+[\text{True predictions of PMDD}]}{[\text{Total predictions of Control}]+[\text{Total predictions of PMDD}]} \times 100\% \qquad \text{Eqn. 2}$$

Repetition of this procedure 1,000 times would yield 1,000 Accuracy estimates, and in turn the Average Accuracy regarding the P-vs-C typing.

Functional annotation of genes overlapping with diagnostic CNV features

By comparing the genomic coordinates of all the frequency-based diagnostic CNV features to those of the known genes retrieved from the R package 'TxDb.Hsapiens.UCSC.hg19.knownGene' version 3.2.2 (25), and considering any gene to be 'overlapping' with a CNV feature if any proportion of its sequence (from > 0% to 100% in 10% increments) coincided with part or all of the CNV feature, the list of CNV-overlapping genes obtained was uploaded to DAVID Bioinformatics Resources as a *test-list* employing the 'RDAVIDWebService' R package (26). All the known genes on chromosomes 1-22 and X were also uploaded as the *background-list*. Comparison of the two lists using the 'getFunctionalAnnotationChart' of the 'RDAVIDWebService' R package revealed gene pathways or categories, as defined in the GO, KEGG and INTERPRO databases, that were enriched with the *test-list* of genes among the *background-list* of genes. The pathways or categories yielding <0.05 Benjamini-corrected *p*-values were regarded to be significantly enriched in the genes on the *test-list* (Table S5).

Genomic-feature content of diagnostic CNV features in different replication phases

DNA sequences on 22 autosomes and chromosome X were subject to replication-time segmentation according to Long & Xue (27). Briefly speaking, experiment-assessed replication timing of all 1-kb sequence windows in the genomes of fifteen human cell lines were retrieved from the 'UW Repli-seq track' in the UCSC Table Browser (28), and the representative replication phase of each sequence window was identified as one of the six types of sequencing segments (viz. G1b, S1, S2, S3, S4 and G2) based on their experiment-assessed replication timing in all fifteen human cell lines.

The density or intensity of genomic features were quantified as described in Ng et al (29) in the diagnostic CNV features and the non-diagnostic-CNV regions in each type of replication phase. The genomic feature content of diagnostic CNV features is indicated by the fold change of the density or intensity of the genomic feature in diagnostic CNV features relative to the non-diagnostic-CNV regions.

Statistical analysis

All comparisons of CNV frequencies were conducted using Fisher's exact tests, and the *p*-values were adjusted by false discovery rate for multiple comparisons. In functional annotation of genes, *p*-values from DAVID web service were subject to Benjamini-correction for multiple comparisons. When annotating the genes that overlapped with any diagnostic CNV feature, empirical *p*-values were estimated using Monte Carlo methods with 1,000 simulations to validate the significant gene pathways/categories based on the 50-, 100- or 450-kb size groups of CNV features. In each round of simulation, sequence windows of the same size as the targeted group of CNV features were randomly selected from chromosomes 1-22 and X, with the number of selected windows being equal to the average number of CNV features in the different type-comparisons to be analysed (see Table S6). For each simulation, the genes that overlapped with any of the selected sequence windows were functionally annotated. The empirical *p*-value of a targeted pathway was given by (r+1)/(n+1), where n = 1,000 and r = number of simulations that displayed significant enrichment (<0.05 Benjamini-corrected *p*-values) in the targeted pathway.

Software for data processing and visualization

Data processing tasks were carried out using custom R codes, except that tasks requiring machine learning were processed using Weka package. All figures were drawn under R environment using the ‘ggplot2’ (30), ‘pheatmap’ (31) and ‘quantsmooth’ (32) packages, except for Figure 3 which was drawn using http://bioinformatics.psb.ugent.be/webtools/Venn/, and Figure 5 using Integrative Genomics Viewer 2.3.69 (33).

Results

Correlations between clinical diagnosis and CNV profiles

In order to examine whether there might be significant correlation between the clinical symptoms of PMDD patients and their germline CNV profiles, the CNVGs and CNVLs called from different sizes of sequence windows on the 127 P-type DNA samples, were subjected to hierarchical clustering in each instance. The CNVGs and CNVLs called from 100-kb sequence windows of the 71 depression-subtype and 56 invasion-subtype patient samples were segregated using the cutree and semi-supervised methods into distinct D-type and V-type clusters in the dendrograms as shown in Figure S2 and Figure 1 respectively. The clusters obtained from the CNVG dendrograms were designated as D_G and V_G clusters, and the clusters obtained from the CNVL dendrograms were designated as D_L and V_L clusters. Notably, the cutree method yielded 72 V_G-type and 55 D_G-type CNVG profiles with 81.10% consistency between the invasion-vs-depression clinical classification and the V-vs-D CNVG-based classification (Figure S2A); whereas the semi-supervised method yielded 61 V_G-type and 66 D_G-type CNVG profiles with 89.76% consistency between the invasion-vs-depression clinical classification and the V-vs-D CNVG-based classification (Figure 1A). Therefore, using either the cutree method or the semi-supervised method, the CNVG-based classification was highly correlated with the clinical symptom-based classification of the

PMDD genomes; this was likewise the case with the D_L-type and V_L-type CNVLs. Altogether, for the CNVGs and CNVLs in, the 50-500 kb window sizes, the cutree method yielded consistencies of 68-91%, and the semi-supervised method yielded consistencies of 88-98%, between the CNV-based and symptom-based classifications. The semi-supervised classifications of P-type samples based on CNVs called from 50-500 kb window sizes were available in Figure S3. These results demonstrated that both the CNVGs and CNVLs contributed to the etiology of the depression-type and the invasion-type symptoms. Moreover, the comparable results obtained using the cutree and semi-supervised methods confirmed the robustness of the CNV-symptom correlations. When the CNVGs or CNVLs called from 100-kb sequence windows of the 108 C-type control samples were subject to hierarchical clustering along with the P-type samples, ~40% of the C-type CNV profiles formed a tight sub-cluster and ~60% were dispersely distributed in the dendrogram, forming sub-clusters with the depression-subtype or invasion-subtype PMDD samples (Figure S4).

Use of diagnostic CNV-features for predictive subtyping

The correlation between germline CNV profiles and clinical subtypes of PMDD suggests that it would be practicable to predict from the germline CNVs of women their propensity to develop PMDD, as well as the likely subtype of the PMDD clinical condition. Toward this objective, the method developed earlier by us through the use of diagnostic CNV-features selected with machine learning to assess a subject's propensity for cancer (15) could be employed as described in 'Selection of diagnostic CNV features' under Method. Figure 2 shows the diagnostic CNV features selected by either the correlation method or the frequency method for prediction the propensity of a test subject's germline CNVs for which of the P, C, V_G, D_G, V_L and D_L genomic groups: the P-type and C-type outcomes would be assessed based on PMDD symptoms; V_G and D_G would be based on the distinction between the V and

D clusters in the CNVG dendrogram in Figure 1A; and V_L and D_L would be based on the distinction between the V and D clusters in the CNVL dendrogram in Figure 1B. The diagnostic CNVG and CNVL features selected using the correlation and frequency methods are given in Table S7.

Figure 2A shows the sets of diagnostic CNV features selected using the correlation-based (red triangles) or frequency-based (black circles) method to enable a choice between a pair of genomic groups. For example, the D_G-vs-C panel of Figure 2A contained a mixture of 66 D_G-type samples and 108 C-type samples. The diagnostic CNV features selected from the total of 144 samples by means of either the correlation method (red triangles) or the frequency method (grey circles) were distributed in a crescent near the y-axis and another crescent near the x-axis. Accordingly, any DNA sample in the mixture that was enriched with near-y diagnostic CNV features would be predicted to be endowed with a greater propensity for C-type over D_G-type, whereas any DNA sample that was enriched with near-x diagnostic CNV features would be predicted to be endowed with a propensity for D_G-type over C-type. In the D_G-vs-C panel of Figure 2B, diagnostic CNV features selected using the correlated method was employed to predict the D_G-vs-C nature in the 174-sample mixture as described under the ‘Predictive subtyping of genomic samples by machine learning’ section in Methods. After 1,000 trial runs, each with a random partition of the samples into an 87-sample Learning Band and an 87-sample Test Band, the average prediction accuracy obtained was 83.0%. Altogether, the seven panels in Figure 2B yielded average prediction accuracies ranging from 81.0% to 88.4%. Interestingly, the list of correlation-based CNV features useful for differentiating between the propensities toward the D and V subtypes (Table S8) showed that the CNV features biased in favor of V-type samples were mostly CNVL features (27/42 for

V_G and 14/17 for V_L). The accuracies of sample-classification predictions derived from the cutree method are available in Figure S5.

Favorable diagnostic CNV-features were often shared by more than one PMDD types, as indicated by the overlaps between the colored circles for the P-vs-C (blue), D-vs-C (red) and V-vs-C (green) comparisons in the Venn diagrams (Figure 3A and B). A range of CNV features were shared by all three kinds of circles, suggesting that they represented key CNV features differentiating between the control and PMDD patient samples (Table S9). Notably also, in all the panels in Figure 3, there was no CNV feature was shared only by the red circles for D-vs-C and the green circles for V-vs-C, which suggests that the CNV-features favoring the D-type genomes differed diametrically from the CNV-features favoring the V-type genomes. As well, there were more D-favoring CNVG features than CNVL features, but more V-favoring CNVL features than CNVG features.

Genome-wide distribution of diagnostic CNV features

In order to have a global view of CNV profiles, the locations and replication timing of all frequency-based diagnostic CNV features, whether overlapping with any known genes or not, were plotted on Figure S6. The results showed that the CNV features were widely spread on all the somatic chromosomes and chromosome X. Chromosomes 4, 13, 18 21 and X were particularly abundant in CNV features that replicated in the G2 phase. Given the correlation between the clinical symptom-based typing of PMDD cases and the clustering of germline diagnostic CNV features, these CNV features could be useful guides in a search for some of genomic sites underlying PMDD.

In Figure 4, the distributions of the CNV features among DNA regions replicating at different cell cycle phases exhibited a number of characteristics: (a) In terms of the number of CNV features that differed between a pair of CNV-types, the P-vs-C panel (viz. P>C or P<C) gave rise to the smallest difference, whereas the D-vs-V pair ($D_G>V_G$ or $D_G<V_G$) gave rise to the largest difference; (b) the ratio of CNVL features relative to CNVG features (viz. L/G on chart) that favored the C-type over P-type were 1.32 for 50-kb CNV features, 2.13 for 100-kb ones and 2.05 for 450-kb ones, all greater than unity (Figure 4A); (c) the P-vs-C comparisons were suggestive of protective effects of smaller size CNVLs in the early replication phases and larger CNVLs in the later phases (Figure 4A); (d) the CNVLs captured by 50-kb windows included significantly more V-favoring than either C-favoring or D-favoring ones (L/G = 2.41 in Figure 4C and 1.80 in Figure 4D); (e) the CNVGs were significantly enriched in D-favoring features compared to C-favoring or V-favoring ones, whereas CNVLs were significantly enriched in V-favoring features compared to C-favoring or D-favoring ones; (f) D-vs-V comparisons suggest that V-type PMDD was correlated with smaller CNVG features belonging to the early replication phases and large CNVGs belonging to the later phases; (g) Large CNVG features were enriched in the G2-phase replicating sequences, especially among the features selected for the D-vs-C and D-vs-V comparisons (see G2-phase columns marked with red asterisks in Figure 4B and D); (h) More than half of the large G2-phase CNVG features in the $C>D_G$ group are identical to those of the $V_G>D_G$ group, suggesting the shared genetic variations in G2 phase underlying V and C types; (i) Large CNVL feature were enriched in S3-phase replicating sequences in the $C>V_G$ and $D_G>V_G$ groups but not in the $V_G>C$ or $V_G>D_G$ groups. The replication-phase distributions of CNV features obtained based on the D_L- or V_L-type samples derived from the CNVL dendrogram in Figure 1B were available in Figure S7.

Pathways and genes enriched in diagnostic CNV features

A wide range of genes showed sequence overlaps with the frequency-based diagnostic CNVG and CNVL features of a range of KEGG pathways in PMDD and its subtypes (Table 1) which pointed to their possible contributions to the PMDD disorder, and some major genes were contained in more than one pathway (Table 2). It was striking that, as indicated in lines 1-5 of Table 2, the control C-type was favored by high frequencies of CNVG features relative to the diseased P-, D- or V-type, suggesting that a major causal factor of the PMDD disorder could be decreased levels of the CNVG features overlapping with the steroid hormone biosynthesis pathway, with the involvement of *CYP*- and *UGT*-genes replicating in phases S2 and S1. As shown in lines 9-17 of Table 2, the C-type and V-type profiles were favored over the D-type by high frequencies of CNVL features in the *GRI*-genes, which were involved in pathways of nicotine addiction, circadian entrainment, serotonergic synapse, dopaminergic synapse and cAMP signaling. The chromosomal sites of these genes and their overlaps with the 100-kb CNV features are shown in Figure 5.

The 50-kb CNV features overlapped with the genes in the glutamatergic-synapse, alcoholism, and systemic lupus erythematosus pathway genes, as well as steroid hormone biosynthesis pathway genes replicating in S2 and S3 (Table S5). On the other hand, the 450-kb CNV features overlapped with chemokine signaling pathway genes (Table S5). Because high regional density of genes could impact on gene annotations by yielding false-positive co-localizations when a CNV feature incidentally captured a gene cluster belonging to a pathway, empirical *p*-values based on Monte Carlo simulations were also estimated for the 100-kb CNV features (Table S10), which provided additional support for some of the pathway in Table 2 through the elimination of such false positives (see ‘Statistical analysis’ in Methods).

Genomic features enriched in diagnostic CNV features

Co-localization analysis revealed various associations between 100-kb frequency-based CNV features and a wide spectrum of genomic features in different replication phases (Table 3 and Table S11). D versus V differences in genomic feature contents can be identified from the thermal scale plots of co-localization scores illustrated in Figure 6 and Figure S8. The genomics features apparently differed between D and V types included: (1) In terms of retrotransposons, D-favoring CNVG features enriched with more of the subfamily of evolutionarily very young short transposons SVAef, while V-favoring CNVG and D-favoring CNVL features enriched with the very young long transposon subfamily, L1vy. (2) With respect to genetic markers, P-favoring, especially D-favoring CNVL features were enriched with recombination events as well as genetic variation hotspots and clusters (27). GWAS reported markers were co-localized with D-favoring CNVGs in S1, V-favoring CNVLs in S4 and C-favoring CNVLs G1b. As well, ClinVar markers were enriched in V-favoring CNVG of S2 phase and D-favoring CNVL of S1. (3) In respect to the group of CpG-related genomic features, the main difference between the two types was that D-favoring CNVL features were more enriched with CpG features such as MeBS in S4 replicating sequences. Compared with D- and V-favoring, the C-favoring CNV features were more prominently enriched with CpG features, especially for C-favoring CNVG in S3 and CNVL in G2 and S4 phases. (4) In regard to non-coding RNA, LINC was enriched in V-favoring CNVL as well as C-favoring CNV features, but not in D-favoring features. (5) To a lesser extent, the enrichment of histone binding sites in D-favoring CNVG features of G2 and S2 phases. In contrast, histone sites were enriched in V-favoring CNVL features of S4 phase. This trend was clearly visible from Figure 6, where twelve kind histone binding sites were analyzed separately and displayed side-by-side.

Some of the strongly enriched genomic features with great than one-fold enrichment was listed in Table 3. For example, enrichment of DNase I hypersensitive sites (DNase) was found in C>P CNVL and C>V_L CNVG features in G2-phase replicating sequences and P>C (as well as D>C and V>C) CNVL features in S4-phase replicating sequences. Regulatory elements isolated by formaldehyde (FAIRE) were found to enrich in C>V, D>C and D>V CNVL features that located in the late-replicating S4 and G2 phases. Disease- or trait-associated SNPs identified by genome-wide association studies (GWAS) were enriched in C>V CNVLs in G1b phase, and P>C CNVGs in S1 phase reaching a fold-change of 5.5 relative to non-diagnostic-CNV regions in S1 phase. The C-favoring (C>P and C>D_L) CNVG features and C-favoring CNVL (C>P and C>V_G) features tend to co-localize with CpG islands (CpGi) in median to late-replicating S3-G2 phases. A range of methylation-related features (Me450, MeBS, and MeMRE) were found to enrich in C-favoring CNVG features mainly in early to median G1b-S3 phases, and C-favoring CNVL features in late-replicating S4-G2 phases. Long intergenic non-coding RNAs (LINC) were found to be enriched in C-favoring CNVGs mainly in S2 phase or CNVLs mainly in G2 phase, D-favoring CNVGs in G1b phase, and V-favoring CNVLs in S3-G2 phases.

Discussion

Application of either the cutree method or the semi-supervised method to the hierarchically clustered CNVGs or CNVLs in the germline genomes of PMDD subjects enabled the distinction between the D-type and V-type CNV profiles. The high degree of consistency between the clinical depression-subtype and D-type CNV profiles, and between the clinical invasion-subtype and V-type CNV profiles, indicated that the two clinical PMDD subtypes were intrinsically correlated with the two dissimilar types of CNV profiles. This was further conformed when diagnostic CNVG and CNVL features were selected by means of machine learning using the correlation method, and employed as abundance markers to predict whether a given germline genomic sample belonged to the control group, the PMDD group, the V-type CNV group or the D-type CNV group, yielding average accuracies of prediction of 81.0-88.4% (Figure 3), which in turn validated the use of diagnostic CNVG and CNVL features to identify the genes and pathways that overlapped with such diagnostic features as potential contributors to the PMDD disorder.

In this regard, there exists overall accord between the cutree and the semi-supervised methods in terms of diagnostic CNV features identified, replication-phase distribution and pathway enrichments (Table S5, S12 and S13). As indicated in DSM-V, PMDD is defined by a complex system of symptoms. In the present study, limited data allowed the analysis of only the depression-type and invasion-type symptoms. Nevertheless, the Venn diagrams in Figure 3 clearly showed that the CNVs underlying the D-type and V-type CNV profiles were strikingly more divergent from one another than their separate divergences from the CNVs underlying the C-type. This finding was also consistent with the results in Figure 2A, which showed that there were more correlation-based or frequency-based CNV features that could be employed to distinguish between D_G-vs-V_G or D_L-vs-V_L compared to CNV-features that could distinguish between D_G-vs-C, V_G-vs-C, D_L-vs-C or V_L-vs-C. As well, the mixed distribution of control CNV profiles among depression- or invasion-type CNV profiles (Figure S3) indicated that the difference between the CNVs in the two subtypes of PMDD was larger than their individual differences from the control. This surprising genome condition, as illustrated in Figure 7, raises the question of whether the depression-type and invasion-type conditions of PMDD might represent two distinct clinical disorders.

A faithful temporal order of DNA replication is fundamental to normal cellular function, and aberrant replication timings were observed in complex diseases including cancers (34, 35). Accordingly, the relative abundances of diagnostic CNVG and CNVL features among genomic DNA sequence regions preferentially replicating in each one of the six phases of cell cycle, namely G1b, S1, S2, S3, S4 and G2 were examined in Figure 4. The peaks of C-favoring CNVL features in P-vs-C comparisons (downward hollow green bars in Figure 4A) shifted clearly from the early S1 phase among the 50-kb features to the late G2 phase among the 450-kb features, pointing to the enrichment of some small CNVL features in the early-replicating regions and larger CNVL features in the late-replicating regions among the determinants of the C-type, *viz*. in the prevention of PMDD occurrence. As well, more D-favoring CNVG features were located in G2-replicating sequences compared to genomic DNA sequences replicating in other cell-cycle phases within the D>C and D>V groups, which was particularly notable in view of the enrichment of G2 phase-replicating sequences in non-coding sequences (27, 29). In addition, the abundance of V-favoring 50-kb CNVL features in the V>C and V>D comparisons (Figure 4B and D) suggests that small-size CNVLs also played important roles in the development of V-type PMDD.

When diagnostic CNVs were analyzed for their genomic feature enrichment with reference to replication phases, interesting observations were obtained (Figure 6; Table 3). It has been revealed that the late-replicating S4-G2 phases in the gene-distal zones are found to be depleted of functional genomic features (27). However, the present study observed associations of open chromatin signals, regulatory elements and epigenetic regulation sites with the diagnostic CNV features in these late-replicating sequences (Table 3), indicating that the diagnostic CNV features might represent pivotal genomic sites in the late-replicating sequences that sequence alterations may give raise to functional perturbations underlying PMDD and its two subtypes. As illustrated herein, genomic feature content analysis, implemented with replication phase information, has pointed to the likelihood of genomic events underlying the subtyping of PMDD and hence a genomic nature of the disorder and its clinical diversity. Since genomic features included in the analysis were broad in spectrum and well beyond the boundary of known genes, the feature enrichment analysis performed herein may complement with and surpass genetic pathway analysis as a powerful tool for genomic studies on complex traits and disorders.

Previously, a number of genes was proposed to be PMDD suspectable genes, including those of steroid hormone biosynthesis (2, 3), and estrogen signaling (36, 37), and these proposals were supported by the presence of these genes in Table 1. The overlaps of genes of nicotine addiction, glutamatergic synapses, olfactory transduction, alcoholism, systemic lupus erythematosus, hypogonadism, premature ovarian failure, and breast cancer with PMDD might be suggestive of hitherto hidden aspects of central nervous system or endocrine system involvements with PMDD. The *GRIA4* gene, overlapping with the 100-kb CNV features for the C>D and V>D comparisons, groups, has also been found to be associated with schizophrenia (38), in accordance with the shared CNVs between schizophrenia and PMDD (14).

In conclusion, through CNV profiling, the present study provided evidence for strong correlation of the clinical depression-subtype or invasion-subtype with the D-type and V-type germline genomes, marked by the overlaps between their CNVs and the machine-selected diagnostic CNV features that favored one or another type of genomes. On account of this correlation, the diagnostic CNV features could be employed as frequency markers to predict the propensity to PMDD and one of its clinical subtypes, as well as position markers to identify candidate PMDD genes and pathways. Moreover, the genetic difference between the depression-favoring and invasion-favoring CNV profiles was found to exceed their individual divergences from the normal controls (Figure 7), raising the question of how this outcome might have been evolved. Future studies will be required to determine how many of the array of PMDD symptoms besides the depression-subtype and invasion-subtype ones could be significantly correlated with CNVs, and what complex diseases other than PMDD would embody CNV-symptom correlations as strong as those encountered with PMDD.

Acknowledgements

We thank Ms. Peggy Lee for the technical support. The study was supported by grants to HX from University Grants Council (SRF116SC01; UROP18SC06; UROP20SC07) and Innovation Technology Council (ITS/085/10; ITS113/15FP; ITCPD/17-9; ITT/023/17GP; ITT/026/18GP) of Hong Kong SAR; Shenzhen Municipal Council of Science and Technology, Guangdong (JCYJ20170818113656988); Shandong Province First Class Disciple Development Grant and Tai-Shan Scholar Program, Shandong; and Ministry of Science and Technology (National Science and Technology Major Project, No. 2017ZX09301064), People's Republic of China, as well as grants from National Natural Science Foundation of China to MQ (No. 8157151623) and JW (No. 81603510), respectively.

Conflict of Interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Author Contributions

HX and MQ conceived and designed the experiments, ZW, XLo, AU, SC and WM performed the AluScan sequencing related experiments and analysis of the sequencing data. PS, MG, JW,

HW, XLi, WS and MQ coordinated the collection of PMDD and control cohorts, and HX, ZW, XLo, SC and MQ wrote the paper.

Supplementary Information

Supplementary materials are available online, including Figures S1-S8 and Table S1-S15.

References

1. Cohen LS, Soares CN, Otto MW, Sweeney BH, Liberman RF, Harlow BL. (2002): Prevalence and predictors of premenstrual dysphoric disorder (PMDD) in older premenopausal women - The Harvard Study of Moods and Cycles. *J Affect Disorders*. 70(2):125-132.

2. Smith SS, Ruderman Y, Frye C, Homanics G, Yuan M. (2006): Steroid withdrawal in the mouse results in anxiogenic effects of 3alpha,5beta-THP: a possible model of premenstrual dysphoric disorder. *Psychopharmacology (Berl)*. 186(3):323-333.

3. Turkmen S, Backstrom T, Wahlstrom G, Andreen L, Johansson IM. (2011): Tolerance to allopregnanolone with focus on the GABA-A receptor. *Br J Pharmacol*. 162(2):311-327.

4. Epperson CN, Haga K, Mason GF, Sellers E, Gueorguieva R, Zhang W, et al. (2002): Cortical gamma-aminobutyric acid levels across the menstrual cycle in healthy women and those with premenstrual dysphoric disorder: a proton magnetic resonance spectroscopy study. *Arch Gen Psychiatry*. 59(9):851-858.

5. Huo L, Straub RE, Roca C, Schmidt PJ, Shi K, Vakkalanka R, et al. (2007): Risk for premenstrual dysphoric disorder is associated with genetic variation in ESR1, the estrogen receptor alpha gene. *Biol Psychiatry*. 62(8):925-933.

6. Dubey N, Hoffman JF, Schuebel K, Yuan Q, Martinez PE, Nieman LK, et al. (2017): The ESC/E(Z) complex, an effector of response to ovarian steroids, manifests an intrinsic difference in cells from women with premenstrual dysphoric disorder. *Mol Psychiatry*. 22(8):1172-1184.

7. Raffi ER, Freeman MP. (2017): The etiology of premenstrual dysphoric disorder: 5 interwoven pieces. Current Psychiatry. 16(9):20-28.

8. Lo WS, Lau CF, Xuan Z, Chan CF, Feng GY, He L, et al. (2004): Association of SNPs and haplotypes in GABAA receptor beta2 gene with schizophrenia. *Mol Psychiatry*. 9(6):603-608.

9. Chen J, Tsang SY, Zhao CY, Pun FW, Yu Z, Mei L, et al. (2009): GABRB2 in schizophrenia and bipolar disorder: disease association, gene expression and clinical correlations. *Biochem Soc Trans*. 37(Pt 6):1415-1418.

10. Kim YS, Yang M, Mat WK, Tsang SY, Su ZH, Jiang XF, et al. (2015): GABRB2 Haplotype Association with Heroin Dependence in Chinese Population. *PLoS One*. 10(11).

11. Tsang SY, Zhong SF, Mei LL, Chen JH, Ng SK, Pun FW, et al. (2013): Social Cognitive Role of Schizophrenia Candidate Gene GABRB2. *PLoS One*. 8(4).

12. Lew AR, Kellermayer TR, Sule BP, Szigeti K. (2018): Copy Number Variations in Adult-onset Neuropsychiatric Diseases. *Curr Genomics*. 19(6):420-430.

13. Yeung RK, Xiang ZH, Tsang SY, Li R, Ho TYC, Li Q, et al. (2018): GABRB2 knockout mice displayed schizophrenia-like and comorbid phenotypes with interneuron-astrocyte-microglia dysregulation. *Transl Psychiatry*. 8:128.

14. Ullah A, Long X, Mat WK, Hu T, Khan MI, Hui L, et al. (2020): Highly recurrent copy number variations in GABRB2 associated with schizophrenia and premenstrual dysphoric disorder. *Front Psychiatry*. 11:572.

15. Ding X, Tsang SY, Ng SK, Xue H. (2014): Application of machine learning to development of copy number variation-based prediction of cancer risk. *Genomics Insights*. 7:1-11.

16. Ismail KM, Nevatte T, O'Brien S, Paschetta E, Backstrom T, Dennerstein L, et al. (2013): Clinical subtypes of core premenstrual disorders: a Delphi survey. *Arch Womens Ment Health*. 16(3):197-201.

17. Qiao M, Sun P, Wang H, Wang Y, Zhan X, Liu H, et al. (2017): Epidemiological distribution and subtype analysis of premenstrual dysphoric disorder syndromes and symptoms based on TCM theories. *Biomed Res Int*. 2017:4595016.

18. Mei L, Ding X, Tsang SY, Pun FW, Ng SK, Yang J, et al. (2011): AluScan: a method for genome-wide scanning of sequence and structure variations in the human genome. *BMC Genomics*. 12:564.

19. Yang JF, Ding XF, Chen L, Mat WK, Xu MZ, Chen JF, et al. (2014): Copy number variation analysis based on AluScan sequences. *J Clin Bioinforma*. 4(1):15.

20. American Psychiatric Association. (2013): Diagnostic and statistical manual of mental disorders (Fifth Edition, DSM-5TM). American Psychiatric Publishing, Washington, DC and London England. p. 171-174.

21. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. (2010): The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res*. 20(9):1297-1303.

22. Hu T, Chen S, Ullah A, Xue H. (2019): AluScanCNV2: An R package for copy number variation calling and cancer risk prediction with next-generation sequencing data. *Genes Dis*. 6(1):43-46.

23. Suzuki R, Shimodaira H. pvclust: Hierarchical Clustering with P-Values via Multiscale Bootstrap Resampling. R package version 2.0-0. ed2015.

24. Galili T. (2015): dendextend: an R package for visualizing, adjusting and comparing trees of hierarchical clustering. *Bioinformatics*. 31(22):3718-3720.

25. Carlson M, Maintainer BP. (2015): TxDb.Hsapiens.UCSC.hg19.knownGene: Annotation package for TxDb object(s).

26. Fresno C, Fernandez EA. (2013): RDAVIDWebService: a versatile R interface to DAVID. *Bioinformatics*. 29(21):2810-2811.

27. Long X, Xue H. (2020): Genetic-variant hotspots and hotspot clusters in the human genome facilitating adaptation while increasing instability. *bioRxiv* doi: https://doi.org/10.1101/2020.10.16.342188.

28. Hansen RS, Thomas S, Sandstrom R, Canfield TK, Thurman RE, Weaver M, et al. (2010): Sequencing newly replicated DNA reveals widespread plasticity in human replication timing. *Proc Natl Acad Sci U S A*. 107(1):139-144.

29. Ng SK, Hu T, Long X, Chan CH, Tsang SY, Xue H. (2016): Feature co-localization landscape of the human genome. *Sci Rep*. 6:20650.

30. Wickham H. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York; 2009.

31. Kolde R. pheatmap: Pretty Heatmaps. 2015.

32. Oosting J, Eilers P, Menezes R. quantsmooth: Quantile smoothing and genomic visualization of array data. 2014.

33. Robinson JT, Thorvaldsdottir H, Winckler W, Guttman M, Lander ES, Getz G, et al. (2011): Integrative genomics viewer. *Nat Biotechnol.* 29(1):24-26.

34. Macheret M, Halazonetis TD. (2018): Intragenic origins due to short G1 phases underlie oncogene-induced DNA replication stress. *Nature.* 555(7694):112-116.

35. Watanabe Y, Maekawa M. (2010): Spatiotemporal regulation of DNA replication in the human genome and its association with genomic instability and disease. *Curr Med Chem.* 17(3):222-233.

36. Huo L, Straub RE, Roca C, Schmidt PJ, Shi K, Vakkalanka R, et al. (2007): Risk for premenstrual dysphoric disorder is associated with genetic variation in ESR1, the estrogen receptor alpha gene. *Biol Psychiat.* 62(8):925-933.

37. Dubey N, Hoffman JF, Schuebel K, Yuan Q, Martinez PE, Nieman LK, et al. (2017): The ESC/E(Z) complex, an effector of response to ovarian steroids, manifests an intrinsic difference in cells from women with premenstrual dysphoric disorder. *Mol Psychiatr*. 22(8):1172-1184.

38. Makino C, Fujii Y, Kikuta R, Hirata N, Tani A, Shibata A, et al. (2003): Positive association of the AMPA receptor subunit GluR4 gene (GRIA4) haplotype with schizophrenia: linkage disequilibrium mapping using SNPs evenly distributed across the gene region. *Am J Med Genet B Neuropsychiatr Genet.* 116B(1):17-22.

Figure Legends

Figure 1. **Hierarchical clustering of PMDD samples based on their pairwise similarities in genome wide CNV profiles.** For all 127 P-group samples, all CNVs were identified from

AluScan sequencing data with 100-kb non-overlapping scanning windows across the genome and used in the plots of similarity scores for CNVGs (A) and CNVL (B), respectively. The dendrograms on top of the heat maps were bootstrapped 1,000 times. The color of each square in the heat map indicates the correlation coefficient (r) of a pair of samples according to the blue-red thermal scale. The semi-supervised classification of samples based on (A) CNVGs and (B) CNVLs was indicated by the red dendrogram branches for V-type and blue ones for D-type genomes. The bands below the dendrograms and on the left-hand side of the heat maps portrayed the subtyping of PMDD samples based on clinical symptoms, with purple bands representing the clinically determined invasion subtype (n = 56) and orange bands the depression subtype (n = 71). Each of the square diagonal boxes in panels (A) and (B) enclosed a group of genomes with close correlations between each other in the group, such that they could be identified as a coherent block of genomes belonging to either the V-type or D-type CNV profiles depending on their enrichment in the invasion- or depression-subtype samples (see ‘Clustering of patient samples based on CNV profiles’ in Methods). Comparable heat maps obtained using sequence window sizes of 50 to 500 kb for CNV-calling are shown in Figure S3.

Figure 2. Occurrence frequencies of diagnostic CNV features and their prediction accuracies for seven pairs of sample groups. Panel (A) shows the frequency distribution of diagnostic CNV features for different pairs of sample groups. The x-axis represents the frequency of CNVs in the first-named group (Group 1 as shown on x-axis), and y-axis the frequency of CNVs in the second-named group (Group 2 as shown on y-axis) in a given pair of sample groups. Diagnostic CNV features with higher frequencies in Group 1 relative to Group 2 (located in lower right crescent) are referred to as ‘Group 1-favoring’ features, whereas diagnostic CNV features with higher frequencies in Group 2 relative to Group 1

(located in upper left crescent) are ‘Group 2-favoring’ features. Black circles are CNV features selected using the frequency-based method with FDR < 0.01 (Fisher’s exact tests), and red triangles are CNV features selected using the correlation-based method. Panel (B) shows the prediction accuracies (estimated using Eqn.2 in Methods) of sample classification in seven sample-pairs based on CNV features selected using the correlation method. For each of the seven pairs, prediction accuracy was estimated 1,000 times and the average accuracy (Av.) was given in the pertinent panel. Subscript G denotes that the D- or V-type samples were derived from the dendrogram of CNVGs (Figure 1A), while subscript L denotes that the D- or V-type samples were derived from the dendrogram of CNVLs (Figure 1B).

Figure 3. Overlaps between the diagnostic CNV features differentiating the two subtypes of PMDD collectively and individually from the control. CNV features identified using (A) correlation-based method, and (B) frequency-based method. Circled ‘G’ indicates CNVG features and circled ‘L’ indicates CNVL features. The ‘>’ and ‘<’ signs portray the relative frequencies of the CNV features for a pair of sample groups, e.g. P>C represents diagnostic CNV features that occurred in higher frequencies in P-group compared to C-group. Subscript G denotes that the D- or V-type samples were derived from the CNVG dendrogram in Figure 1A, whereas subscript L denotes that the D- or V-type samples were derived from the CNVL dendrogram in Figure 1B.

Figure 4. Distribution of frequency-based diagnostic CNV features among genomic sequences of different DNA replication phases. Number of base pairs of the CNV features called using 50, 100 and 450-kb windows for (A) P-vs-C, (B) D_G-vs-C, (C) V_G-vs-C, and (D) D_G-vs-V_G groups. The solid bars represent CNVG features and hollow bars represent CNVL features in each panel. The replication phases G1b to G2 are color coded as shown. The ‘>’ or

'<' sign portrays larger or smaller frequencies of the CNV features in favor of the first-named group over the second-named one. L/G represents the ratio of the number of CNVLs over the number of CNVGs. Significant enrichment of CNV features in a particular replication phase in the genome is indicated by asterisks that are color coded according to the replication phase, or in black asterisks for comparison between an L/G value in the upper half of a panel and an L/G value in the lower half (Bonferroni-corrected, *** $p < 0.005$, ** $p < 0.01$, * $p < 0.05$). Numerical *p*-values are shown in Table S14. Subscript G denotes that the D- or V-type samples were derived from the CNVG dendrogram in Figure 1A. See Figure S7 for the results obtained based on the D- or V-type samples derived from the CNVL dendrogram in Figure 1B.

Figure 5. Selected genes overlapping with frequency-based diagnostic CNV features. Expanded views of chromosomal segments on (A) chromosomes 2 and 7 for steroid biosynthesis pathway genes, (B) chromosomes 5, 11, 12, 16 and 17 for *GRI*-genes of the glutamatergic synapse and nicotine addiction pathways, and (C) chromosomes 6 and X for the non-pathway *TRERF1* and *POF1B* genes with color-coded representation of the DNA replication phase in the 'Phase' track, and aligned gene sequence(s) in blue (e.g. *UGT1A8* or *TRERF1*) as described in RefSeq Genes in UCSC Genome Browser. Green rectangular boxes either below the genes indicate the presence of diagnostic CNVG or CNVL feature(s). Inside each box, colored stripes are indicative of CNVL features(s), and solid coloring is indicative of CNVG features(s): purple for predominantly V-favoring features, orange for D-favoring features, and green for C-favoring features.

Figure 6. Enrichment analysis of genomic-feature contents in different replication phases for control and PMDD subtypes. Frequency-based CNV features diagnostic for C

group, i.e., control, as well as that for D and C groups of PMDD samples clustered by CNVG dendrogram, identified with 100-kb scanning windows, were used in the analysis. Enrichment analysis results were plotted for CNVG features in the upper two panels and that for CNVL features in the bottom two panels. A similar analysis performed in parallel for clustered by CNVL dendrogram can be found in Figure S8. Fold-change of each genomic feature in the diagnostic CNV features relative to the non-diagnostic-CNV regions was estimated according to 'Genomic-feature content of diagnostic CNV features in different replication phases' in Methods, and was color-coded based on the thermal scale. Fold-change greater than 2-fold was capped at 2 in the heat map. 'Group 1' indicated the first-named group and 'Group 2' the second-named group in a given pair of samples. Genomic features were grouped into Retrotransposon (SVAef, SVAcd, SVAab, AluYvy, AluYy,s AluS, AluJ, FLAM, L1vy, L1y, L1m, L1o, MIR, L2), Genetic markers (RecD, RecH, RecK, GWAS, ClinVar, GV hotspot, Cluster, CNVG), Regulatory sites (H3k27me3, H4k20me1, H3k9me1, H2az, H3k79me2, H3k36me3, H3k4me3, H3k9ac, H3k4me2, H3k27ac, H3k4me1, H3k9me3, MeMRE, MeDIP, MeBS, CpGi, CpGe, Me450, TFBS, REG, FAIRE, DNase) and Gene/Transcription (Gene, EXPS, LRNA+, LRNA-, LINC) groups on the x-axis based on their sequence and functional properties. The descriptions of genomic features and numeric data were available in Table S11.

Figure 7. Genetic distances between the two subtypes of PMDD and the control. Pairwise distances were estimated based on the abundance of diagnostic CNV features between C-, D- and V-type genomes. The numbers of frequency-based CNV features were employed as an approximate index of the genetic distance between the D-vs-C, V-vs-C or D-vs-V sample pairs in Table S15, which comprised the 50-500 kb frequency-based CNV

features. Notably, the D-vs-V distance was larger than the D-vs-C distance or the V-vs-C distance.

Table 1. Selected genes overlapping with 100-kb frequency-based CNVG and CNVL features[1] with adjusted *p*-values less than 0.01.

Gene	**Ratio**	**H%**[2]	**L%**[3]	**CNV**	**CNV location (x 100 kb)**[4]	***p*-value**[5]	**CNV in replication phase (%)**[6]					
							G1b	**S1**	**S2**	**S3**	**S4**	**G2**
GABA$_A$ Receptor Family (KEGG: Nicotine addiction)												
GABRR1, 2	$C > D_G$	26	2	L	6:899-900	2.05E-04	0	0	99	0	0	0
GABRG3	$C > D_G$	61	27	L	15:275-276	5.55E-04	0	0	0	0	0	100
GABRG3	$V_G > D_G$	59	27	L	15:275-276	3.46E-03	0	0	0	0	0	100
Glutamate Metabotropic Receptor (KEGG: Glutamatergic synapse)												
GRM4	$D_L > V_L$	80	51	L	6:339-340	6.99E-03	0	87	0	0	0	0
GRM8	$D_L > V_L$	91	45	L	7:1,267-1,268	1.08E-06	0	0	0	37	45	0
GRM5	$V_G > D_G$	16	0	L	11:882-883	3.98E-03	0	90	0	0	0	0
Glutamate Ionotropic Receptor (KEGG: Glutamatergic synapse, Nicotine addiction)												
GRIA1	$V_G > D_G$	31	0	L	5:1,530-1,531	6.10E-06	0	0	0	12	78	0
GRIK2	$V_G > D_G$	16	0	L	6:1,025-1,026	3.29E-03	0	0	0	0	0	100
GRIA4	$V_G > D_G$	25	3	L	11:1,057-1,058	3.98E-03	0	0	0	0	100	0
GRIN2B	$V_G > D_G$	23	0	L	12:138-139	2.56E-04	0	0	0	0	100	0
GRIN2A	$V_G > D_G$	39	6	L	16:98-99	1.61E-04	0	0	0	77	16	0
GRIN2A	$V_G > D_G$	26	5	L	16:99-100	6.84E-03	0	0	0	94	0	0
GRIN2C	$V_G > D_G$	39	12	L	17:728-729	4.36E-03	3	92	0	0	0	0
UDP Glucuronosyltransferase 1 Family (KEGG: Steroid hormone biosynthesis)												
UGT1A1, 3-10	$V_L > D_L$	36	11	L	2:2,346-2,347	6.62E-03	0	0	98	0	0	0
Cytochrome P450 (KEGG: Steroid hormone biosynthesis)												
CYP3A4, 5, 7	$C > D_G$	30	5	G	7:993-994	3.05E-04	0	100	0	0	0	0
CYP3A4, 5, 7	$C > D_L$	30	9	G	7:993-994	6.15E-03	0	100	0	0	0	0
CYP3A4, 5, 7	$V_L > D_L$	55	9	G	7:993-994	6.37E-07	0	100	0	0	0	0
CYP11B1, 2	$V_L > D_L$	19	0	G	8:1,439-1,440	6.55E-04	0	1	68	0	0	0
CYP11A1	$V_L > D_L$	51	22	L	15:746-747	7.80E-03	100	0	0	0	0	0
Premature Ovarian Failure Protein 1B												
POF1B	$C > P$	29	9	G	X:845-846	2.13E-03	0	0	0	0	6	92
POF1B	$C > D_G$	29	0	G	X:845-846	2.19E-06	0	0	0	0	6	92
POF1B	$V_G > D_G$	18	0	G	X:845-846	1.20E-03	0	0	0	0	6	92
POF1B	$D_G > C$	23	5	G	X:846-847	2.94E-03	0	0	0	0	0	100
POF1B	$D_G > V_G$	23	2	G	X:846-847	1.67E-03	0	0	0	0	0	100
Transcriptional Regulating Factor 1 (Breast cancer anti-estrogen resistance 2)												
TRERF1	$C > D_G$	20	0	L	6:423-424	5.55E-04	8	92	0	0	0	0
TRERF1	$V_G > D_G$	16	0	L	6:423-424	3.98E-03	8	92	0	0	0	0
TRERF1	$D_L > V_L$	42	15	G	6:424-425	7.63E-03	97	0	0	0	0	0

Opioid Binding Protein/Cell Adhesion Molecule Like (Hypogonadotropic Hypogonadism 14)												
OPCML	D_G > C	100	79	L	11:1,331-1,332	3.17E-04	0	0	0	0	0	100
OPCML	D_G > V_G	100	85	L	11:1,331-1,332	7.57E-03	0	0	0	0	0	100
OPCML	D_G > C	61	24	L	11:1,332-1,333	1.07E-04	0	0	0	0	0	100
OPCML	D_G > V_G	61	15	L	11:1,332-1,333	6.24E-06	0	0	0	0	0	100
MACRO Domain Containing 2 (Mono-ADP Ribosylhydrolase 2, Hypogonadotropic Hypogonadism 21)												
MACROD2	C > D_G	21	2	G	20:158-159	7.43E-04	0	0	0	31	34	0
MACROD2	D_G > C	97	77	G	20:151-152	1.45E-03	0	0	0	100	0	0
MACROD2	C > D_G	64	35	G	20:144-145	2.07E-03	0	0	0	74	15	0
MACROD2	C > D_G	64	24	L	20:156-157	2.82E-05	0	0	0	75	23	0
MACROD2	D_G > V_G	97	62	G	20:151-152	8.43E-06	0	0	0	100	0	0
MACROD2	V_G > D_G	31	2	G	20:158-159	2.79E-05	0	0	0	31	34	0
MACROD2	V_G > D_G	21	2	G	20:145-146	2.03E-03	0	0	0	100	0	0
MACROD2	V_G > D_G	64	35	G	20:144-145	6.56E-03	0	0	0	74	15	0
MACROD2	V_G > D_G	74	24	L	20:156-157	1.17E-06	0	0	0	75	23	0
MACROD2	V_G > D_G	15	0	L	20:146-147	7.25E-03	0	0	0	100	0	0

[1] See Table S7 for data on 50-500 kb CNV features;

[2] CNV frequency in first-named, higher frequency group (H), e.g. C-group of C>P pair;

[3] CNV frequency in second-named, lower frequency group (L), e.g. P-group of C>P pair;

[4] Chromosome number with start and end coordinates (to be multiplied by 100 kb);

[5] FDR-corrected *p*-value obtained using Fisher's exact test on counts of CNV features in the two groups compared, as specified in the 'Ratio' column;

[6] % base pairs in 6 replication phases; the % in genomic regions with unknown replication timing are not shown.

Table 2. Representative pathways enriched in 100-kb frequency-based CNV features.

No.	**Group**[1]		**CNV**[2]	**KEGG pathway**[3]	**Chromosome**[4]	**Gene distribution in replication phases**[5]						**Proportion (%)**[6]	***p*-value**[7]
	H	**L**				**G1b**	**S1**	**S2**	**S3**	**S4**	**G2**		
1	C	P	G	Steroid hormone biosynthesis	2	-	-	++++++	-	-	-	> 0 - 100	[5.7E-04, 1.5E-09]
2	C	V_G	G	Steroid hormone biosynthesis	2	-	-	++++++	-	-	-	> 0 - 100	[2.7E-04, 9.3E-10]
3	C	V_L	G	Steroid hormone biosynthesis	2	-	-	++++++	-	-	-	> 0 - 100	[2.7E-04, 1.3E-09]
4	C	D_G	G	Steroid hormone biosynthesis	2,7	-	++	++++	-	-	-	> 0 - 100	[5.6E-05, 1.4E-11]
5	C	D_L	G	Steroid hormone biosynthesis	2,7	-	++	++++	-	-	-	> 0 - 100	[3.1E-05, 7.6E-12]
6	V_L	D_L	L	Steroid hormone biosynthesis	2,15	+	-	+++++	-	-	-	> 0 - 100	[4.8E-02, 6.7E-05]
7	V_L	D_L	G	Steroid hormone biosynthesis	7,8,1	-	+++	++	-	-	-	60	[4.9E-02, 4.9E-02]
8	D_G	V_G	L	Ovarian steroidogenesis	15,10,14,16,7	+++	+	+	+	-	-	40	[4.8E-02, 4.8E-02]
9	C	D_G	L	Nicotine addiction	6,11,12,15,16,17,5	-	+	++	+	++	+	> 0 - 20	[3.2E-02, 1.9E-02]
10	C	D_L	L	Nicotine addiction	6,11,12,15,16,17,5	-	+	++	+	++	+	> 0 - 20	[4.9E-02, 1.2E-02]
11	C	D_G	L	Circadian entrainment	11,17,12,16,20,5,8	+	+	-	+	++++	+	20	[2.9E-02, 2.9E-02]
12	V_G	D_G	L	Serotonergic synapse	1,11,15,12,17,20,21,5,6,7	+	+	+	+++	-	+	> 0 - 30	[3.8E-02, 1.2E-02]
13	V_G	D_G	L	Glutamatergic synapse	11,15,17,1,12,16,20,5,6	+	++	+	+	++	+	> 0, 10	[4.6E-02, 3.6E-02]
14	V_G	D_G	L	Nicotine addiction	15,11,12,16,17,5	-	+	-	++	+++	+	10	[4.6E-02, 4.6E-02]
15	V_G	D_G	L	Dopaminergic synapse	1,11,12,16,17,2,20,21,4,5,7,8	+	-	++	++	++	+	> 0 - 20	[4.4E-02, 3.0E-02]
16	V_G	D_G	L	Circadian entrainment	1,11,17,12,16,20,21,4,5	+	+	++	+	++	+	> 0 - 20	[3.0E-02, 1.7E-02]
17	V_G	D_G	L	cAMP signaling pathway	1,11,5,10,12,16,17,3,4,6,7	+	+	+	++	++	-	20	[3.3E-02, 3.3E-02]

[1] Significant difference in CNV frequencies between compared groups, with ‘H’ and ‘L’ indicating higher- and lower-frequency group respectively. The subscripts ‘G’ and ‘L’ indicate sample groups clustered based on CNVG and CNVL respectively;

[2] ‘G’ indicates copy-number-gains and ‘L’ indicates copy-number-losses;

[3] KEGG pathway IDs are, in order of appearance in table, hsa00140, hsa04913, hsa05033, hsa04713, hsa04726, hsa04724, hsa04728 and hsa04024;

[4] Chromosomes where pathway genes overlapped with CNV feature(s);

[5] Approximate distribution of pathway genes in different replication phases, with ‘-’ indicating 0%, and one ‘+’ indicating 0-20%, up to six ‘+’ indicating 100%;

[6] Proportion of gene sequence overlapping with the CNV feature(s) ranging from > 0% to 100%;

[7] Range of Benjamini-adjusted *p*-values of pathway enrichment pertaining to bottom and top figures referred to in footnote 6.

Table 3. Selected genomic features in 100-kb frequency-based diagnostic CNV features with fold-change greater than 1 in at least one replication phase(s).

Group[1]		CNV[2]	Fold-change in different replication phases[3]					
H	**L**		**G1b**	**S1**	**S2**	**S3**	**S4**	**G2**
CpGi (CpG island)								
C	P	G	-0.32	0.06	-0.06	**1.62**	-0.18	**1.15**
C	P	L	-0.18	0.05	0.39	0.46	**1.10**	**2.20**
C	D_L	G	-0.45	0.02	-0.08	0.59	**1.18**	0.57
C	V_G	L	0.14	0.06	0.69	0.71	0.47	**1.42**
P	C	L	-0.34	0.12	-0.02	0.38	-0.23	**1.08**
Me450 (Methylation status using HumanMethylation450)								
C	P	G	-0.51	0.54	**1.04**	0.88	-0.01	0.43
C	P	L	-0.29	0.23	-0.02	0.33	**1.16**	**1.05**
C	V_L	G	**1.00**	0.04	-0.02	-0.35	-0.48	-0.25
MeBS (cytosine methylation using bisulfite sequencing)								
C	P	L	-0.12	0.45	0.12	0.49	**1.53**	**1.55**
C	D_L	G	-0.53	0.10	0.06	0.64	**1.65**	0.59
C	V_G	L	0.27	0.26	0.69	0.94	**1.51**	0.29
D_G	C	L	**1.03**	0.54	0.35	0.24	**1.20**	-0.39
D_L	C	L	**1.38**	0.20	0.25	0.37	0.13	-0.34
D_G	V_G	L	**1.15**	0.41	0.37	0.46	**1.69**	-0.24
D_L	V_L	L	0.82	0.45	0.53	0.59	**1.47**	-0.15
V_L	D_L	G	0.15	-0.17	-0.14	0.45	**1.35**	**1.21**
MeMRE (Methylation using MRE-								
C	P	G	-0.26	0.23	-0.30	**2.00**	-0.12	-0.19
C	P	L	-0.23	-0.06	0.31	0.47	**1.44**	0.90
P	C	L	-0.13	0.10	0.10	0.33	-0.22	**1.70**
DNase (DNase I hypersensitive sites)								
C	P	L	-0.12	0.02	0.11	0.55	0.46	**1.02**
C	V_L	G	0.22	-0.30	0.25	-0.55	0.04	**1.14**
P	C	L	-0.10	0.08	0.51	0.38	**1.80**	-0.13
D_G	C	L	0.28	0.27	0.24	-0.05	**1.37**	-0.52
D_L	C	L	0.47	0.22	0.39	-0.12	**1.77**	-0.14
V_G	C	L	-0.16	0.22	0.05	0.00	**1.38**	-0.57
V_L	C	L	-0.15	0.14	0.25	0.09	**1.08**	-0.66
FAIRE (Formaldehyde-assisted isolation of regulatory elements)								
C	P	L	-0.18	0.17	-0.10	0.15	0.71	**2.69**
C	V_G	L	-0.23	0.12	0.03	0.14	0.46	**2.97**
C	V_L	L	0.02	0.08	-0.04	0.04	0.33	**2.54**
P	C	L	-0.09	0.16	0.53	-0.03	0.86	**1.56**
D_G	C	L	0.32	0.28	0.16	-0.17	**1.55**	0.14
D_L	C	L	0.34	0.29	0.33	-0.24	**2.11**	**1.25**
D_G	V_G	L	0.22	0.17	0.11	0.11	**1.15**	0.92
D_L	V_L	L	0.20	0.21	0.02	0.03	0.77	**2.08**
LINC (Large intergenic non-coding RNA)								
Continued								
C	P	G	-0.97	-0.52	**18.30**	**4.23**	0.32	-0.11
C	P	L	-0.99	**1.05**	0.17	**1.02**	-0.73	0.92
C	V_L	G	0.11	-0.96	0.03	**1.75**	-0.55	**2.42**
C	V_G	L	-1.00	-0.77	-0.79	-0.93	-0.47	**1.59**
C	V_L	L	-0.98	-0.74	-0.72	-0.87	-0.50	**1.06**
V_G	C	L	0.48	-0.70	0.32	**2.14**	**1.61**	**1.36**
V_L	C	L	0.31	-0.75	-0.02	**1.80**	**1.01**	0.74
V_G	D_G	L	0.37	-0.67	**1.19**	0.52	0.33	**1.34**
RecD (Sex-averaged rates of recombination)								
P	C	L	**1.31**	0.39	0.26	0.44	0.36	0.97
D_G	C	L	0.23	0.16	0.43	0.42	0.35	**1.24**
D_L	C	L	0.40	0.30	0.52	0.50	0.51	**1.56**
D_L	V_L	L	0.25	0.13	0.24	0.22	0.45	**1.06**
GWAS (GWAS-identified SNPs)								
C	V_G	L	**2.08**	-0.01	-0.02	-0.35	0.18	0.60
C	V_L	L	**1.63**	-0.13	-0.01	-0.23	0.10	0.38
P	C	G	-0.26	**5.48**	-0.38	0.27	-0.26	-0.34
D_G	C	G	-0.16	**1.61**	-0.13	-0.28	-0.05	-0.01
D_L	C	G	-0.08	**1.78**	-0.17	-0.22	-0.03	-0.16
V_G	C	L	-0.34	-0.20	-0.22	-0.23	**2.78**	-0.14
V_L	C	L	-0.28	-0.37	-0.29	-0.17	**2.00**	0.01
D_G	V_G	G	-0.16	**1.25**	-0.02	0.33	-0.28	-0.25
D_L	V_L	G	-0.08	**1.27**	-0.03	0.24	-0.24	-0.11
GV hotspot (Density-based genetic-variant hotspot)								
C	V_G	G	**1.97**	-0.34	-0.21	-0.27	-0.15	**1.22**
C	V_L	G	**1.63**	0.01	-0.27	-0.09	-0.28	0.69
P	C	L	0.12	-0.01	-0.11	0.70	**1.58**	0.29
D_G	V_G	G	-0.03	-0.23	-0.44	0.14	0.20	**1.03**
D_L	V_L	G	0.01	-0.31	-0.43	0.28	0.00	**1.06**
D_G	V_G	L	0.30	0.27	0.09	0.42	0.61	**1.22**
Cluster (Cluster of genetic-variant hotspots)								
C	D_L	L	0.03	-0.84	0.38	-0.87	-0.30	**2.42**
C	V_G	G	**3.68**	-1.00	-1.00	0.42	-0.15	-0.29
C	V_L	G	**2.90**	-1.00	-1.00	0.37	-0.21	-0.52
C	V_G	L	-0.23	**1.42**	0.03	0.24	0.26	0.83
P	C	L	-1.00	-1.00	-0.27	**4.33**	**4.26**	-1.00
D_G	C	L	**2.15**	0.73	-0.05	0.57	**1.34**	-1.00
D_L	C	L	**2.46**	0.90	0.07	0.73	**1.11**	-1.00

LINC (Large intergenic non-coding RNA)								
C	D_G	G	0.50	-0.64	**7.56**	**1.38**	**1.46**	0.11
C	D_L	G	-0.98	-0.76	**8.90**	**1.53**	**1.51**	0.16
C	D_G	L	-0.86	0.44	0.46	-0.79	-0.88	**3.10**
D_G	C	G	**1.84**	-0.92	-0.68	0.21	-0.99	-0.80
D_G	C	L	-0.85	-0.84	-0.95	**1.89**	-1.00	-1.00
D_L	C	L	-0.81	-0.82	-0.95	**1.85**	-1.00	-1.00
V_G	D_G	G	0.87	-0.74	**4.34**	-0.66	0.81	0.16
V_L	D_L	G	0.76	-0.75	**3.45**	-0.69	0.65	-0.08
V_G	C	G	-1.00	-0.26	**1.48**	0.46	0.12	-0.83
V_L	C	G	-1.00	**1.45**	0.72	0.19	-0.08	-0.38
V_G	C	L	-0.41	-0.72	-0.70	**1.68**	0.46	-0.97
V_L	C	L	-0.21	-0.78	-0.41	**1.20**	0.27	-0.89
D_G	V_G	G	0.02	-0.65	-0.55	-0.50	**1.52**	0.89
D_L	V_L	G	-0.05	-0.64	-0.55	-0.53	**1.41**	0.95
D_G	V_G	L	0.91	0.42	0.43	0.64	**1.71**	**1.25**
V_G	D_G	L	-0.64	-0.36	**3.00**	0.50	0.12	0.20
V_L	D_L	L	-0.44	-0.52	-0.15	0.41	-0.38	**1.07**

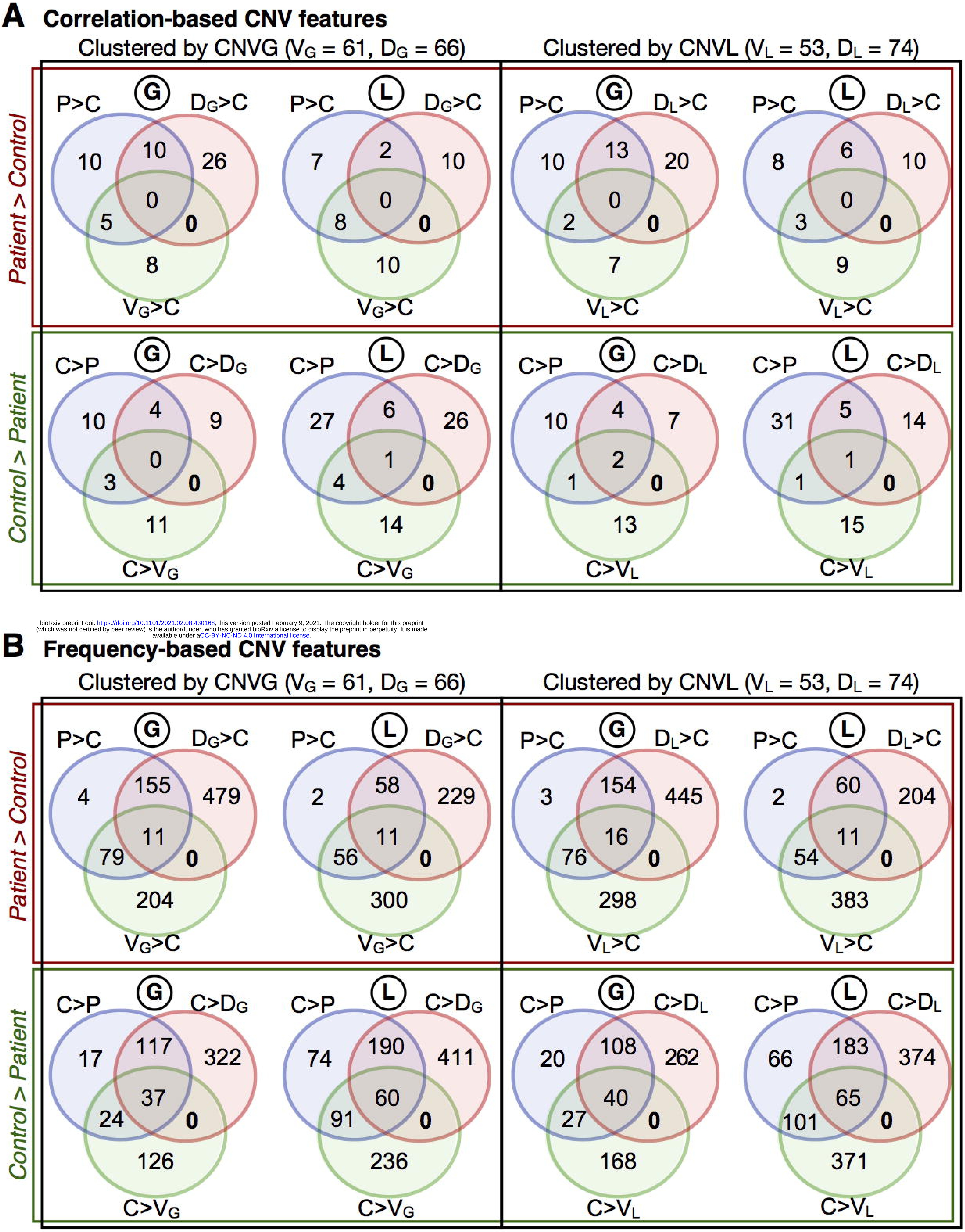
[1] Significant difference in CNV frequencies between compared groups, with ‘H’ and ‘L’ indicating higher- and lower-frequency group respectively. The subscripts ‘G’ and ‘L’ indicate sample groups clustered based on CNVG and CNVL respectively;

[2] ‘G’ indicates copy-number-gains and ‘L’ indicates copy-number-losses;

[3] Fold-change (> 1-fold in bold) of genomic feature density or intensity in diagnostic CNV features relative to non-diagnostic-CNV regions in replication phase.

A
Clustered by 100-kb CNVG
r
1
0.6
0.2
$V_G = 61$
$D_G = 66$
B
Clustered by 100-kb CNVL
r
1
0.6
0.2
$V_L = 53$
$D_L = 74$

A
P vs. C
CNV Features:
△ Correlation-based
○ Frequency-based
DG vs. C
VG vs. C
DG vs. VG
DL vs. C
VL vs. C
DL vs. VL
CNV Frequency in Group 2 (%)
CNV Frequency in Group 1 (%)
0
50
100
B
P vs. C
Av. = 81.1
DG vs. C
Av. = 83.0
VG vs. C
Av. = 85.0
DG vs. VG
Av. = 84.6
DL vs. C
Av. = 81.0
VL vs. C
Av. = 88.4
DL vs. VL
Av. = 85.3
Count (x 100)
0
2
4
60
80
100
Prediction Accuracy (%)



A Correlation-based CNV features

Clustered by CNVG (V_G = 61, D_G = 66)

Clustered by CNVL (V_L = 53, D_L = 74)

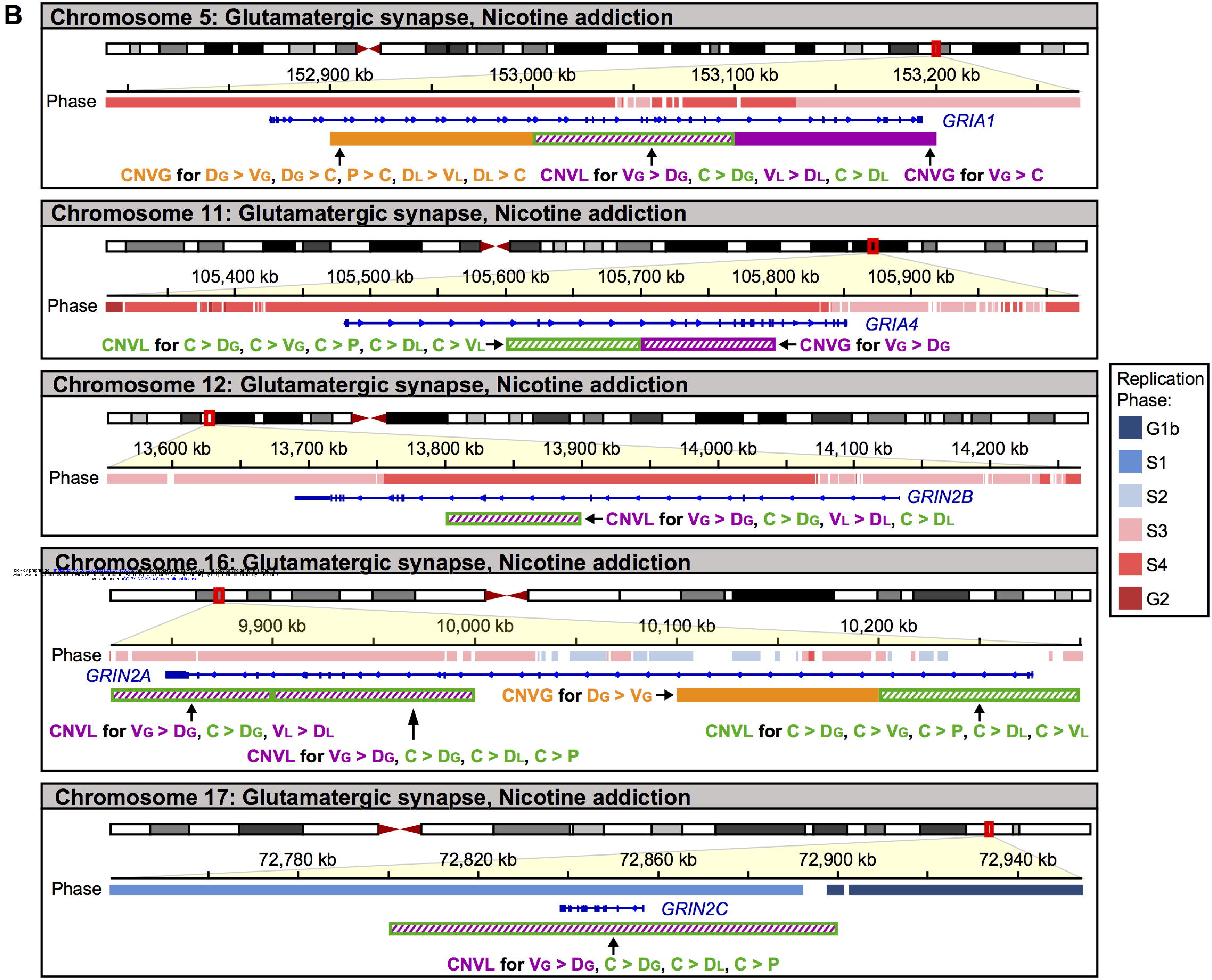
B Frequency-based CNV features

Clustered by CNVG (V_G = 61, D_G = 66)

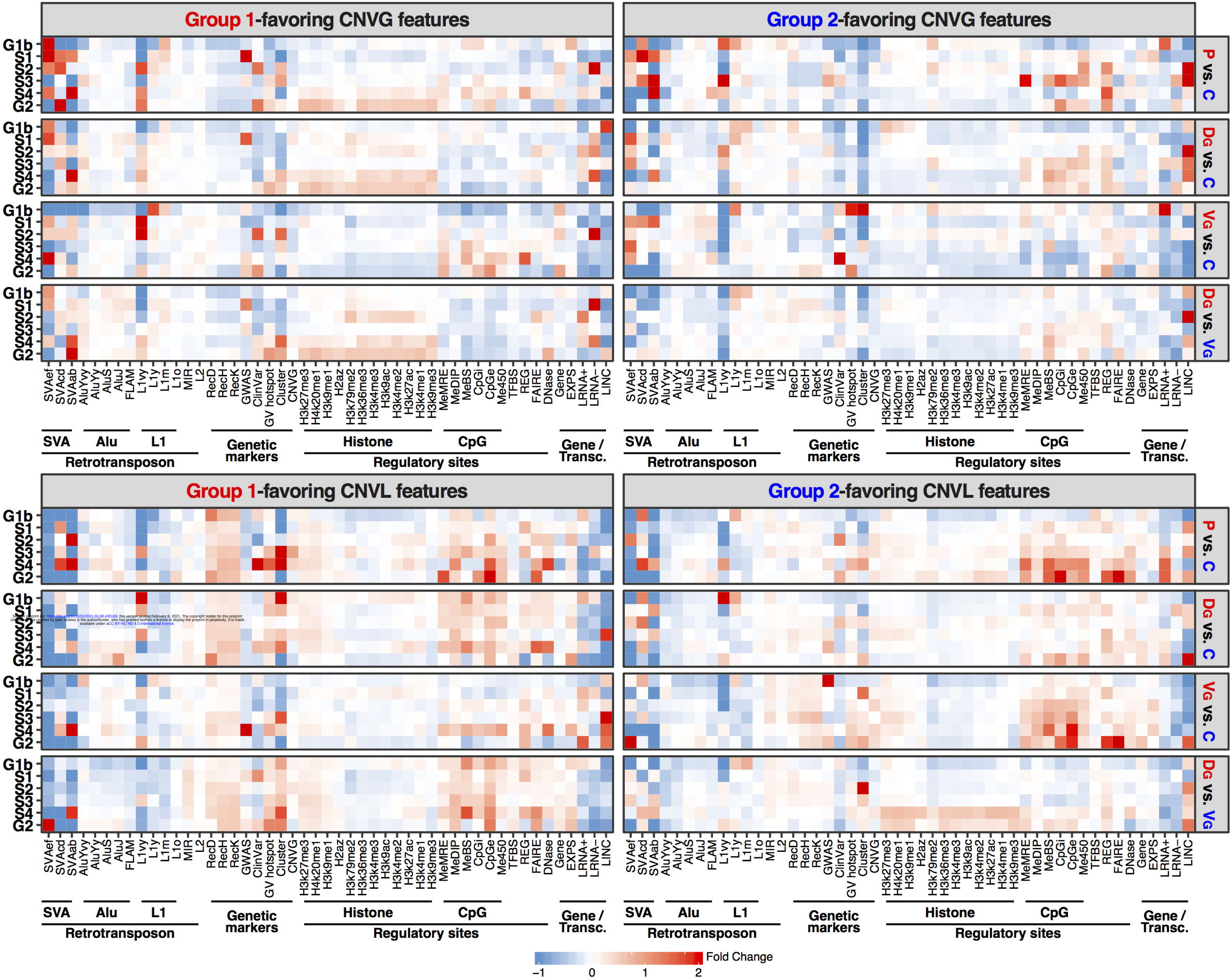
Clustered by CNVL (V_L = 53, D_L = 74)

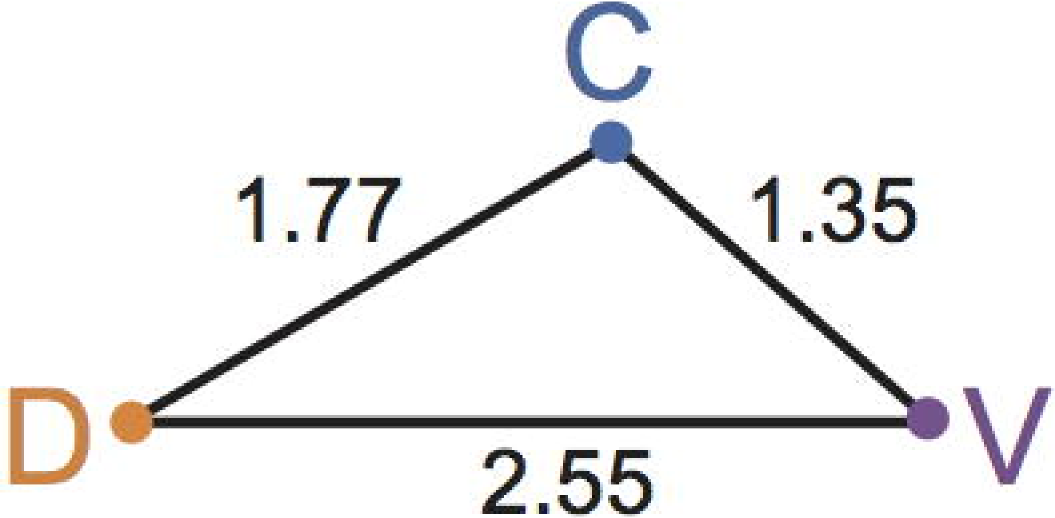
A
P > C features
C > P features
Base pair (x 10^7)
2
1
0
1
2
50-kb
100-kb
450-kb
L/G = 0.75
L/G = 0.51
L/G = 0.64
L/G = 1.32
L/G = 2.13
L/G = 2.05
***
B
DG > C features
C > DG features
Base pair (x 10^7)
50-kb
100-kb
450-kb
L/G = 0.36
L/G = 0.46
L/G = 0.79
L/G = 1.13
L/G = 1.39
L/G = 1.21
***
C
VG > C features
C > VG features
Base pair (x 10^7)
50-kb
100-kb
450-kb
L/G = 2.41
L/G = 1.25
L/G = 1.03
L/G = 1.12
L/G = 2.07
L/G = 1.93
***
D
DG > VG features
VG > DG features
Base pair (x 10^7)
50-kb
100-kb
450-kb
L/G = 0.56
L/G = 0.75
L/G = 1.80
L/G = 1.24
***
CNV type:
CNVG:
solid column
CNVL:
open column
Replication Phase
G1b
S1
S2
S3
S4
G2

A
Chromosome 2: Steroid hormone biosynthesis
234,540 kb
234,580 kb
234,620 kb
234,660 kb
Phase
UGT1A8
UGT1A10
UGT1A9
UGT1A7
UGT1A6
UGT1A5
UGT1A4
UGT1A3
UGT1A1
CNVL for $V_L > D_L$
CNVG for $C > D_G$, $C > V_G$, $C > P$, $C > D_L$, $C > V_L$
Chromosome 7: Steroid hormone biosynthesis
99,200 kb
99,300 kb
99,400 kb
99,500 kb
Phase
CYP3A5
CYP3A4
CYP3A7-CYP3AP1
CYP3A7
CNVG for $V_L > D_L$, $C > D_G$, $C > D_L$
Replication Phase:
G1b
S1
S2
S3
S4
G2


B
Chromosome 5: Glutamatergic synapse, Nicotine addiction
152,900 kb
153,000 kb
153,100 kb
153,200 kb
Phase
GRIA1
CNVG for DG > VG, DG > C, P > C, DL > VL, DL > C
CNVL for VG > DG, C > DG, VL > DL, C > DL
CNVG for VG > C
Chromosome 11: Glutamatergic synapse, Nicotine addiction
105,400 kb
105,500 kb
105,600 kb
105,700 kb
105,800 kb
105,900 kb
Phase
GRIA4
CNVL for C > DG, C > VG, C > P, C > DL, C > VL
CNVG for VG > DG
Chromosome 12: Glutamatergic synapse, Nicotine addiction
13,600 kb
13,700 kb
13,800 kb
13,900 kb
14,000 kb
14,100 kb
14,200 kb
Phase
GRIN2B
CNVL for VG > DG, C > DG, VL > DL, C > DL
Chromosome 16: Glutamatergic synapse, Nicotine addiction
9,900 kb
10,000 kb
10,100 kb
10,200 kb
Phase
GRIN2A
CNVG for DG > VG
CNVL for VG > DG, C > DG, VL > DL
CNVL for VG > DG, C > DG, C > DL, C > P
CNVL for C > DG, C > VG, C > P, C > DL, C > VL
Chromosome 17: Glutamatergic synapse, Nicotine addiction
72,780 kb
72,820 kb
72,860 kb
72,900 kb
72,940 kb
Phase
GRIN2C
CNVL for VG > DG, C > DG, C > DL, C > P
Replication Phase:
G1b
S1
S2
S3
S4
G2

C
Chromosome 6:
41,900 kb
42,100 kb
42,300 kb
42,500 kb
42,700 kb
Phase
TRERF1
CNVL for $V_G > D_G$, $V_L > D_L$, $C > D_G$, $C > D_L$ →
← CNVG for $D_L > V_L$
Chromosome X:
84,200 kb
84,400 kb
84,600 kb
84,900 kb
Phase
POF1B
CNVG for $V_G > D_G$, $C > D_G$, $C > P$ →
Replication Phase:
G1b
S1
S2
S3
S4
G2


Group 1-favoring CNVG features
Group 2-favoring CNVG features
Group 1-favoring CNVL features
Group 2-favoring CNVL features
G1b
S1
S2
S3
S4
G2
P vs. C
Dg vs. C
Vg vs. C
Dg vs. Vg
SVAef
SVAcd
SVAab
AluYvy
AluYy
AluS
AluJ
FLAM
L1vy
L1y
L1m
L1o
MIR
L2
RecD
RecH
RecK
GWAS
ClinVar
GV hotspot
Cluster
CNVG
H3k27me3
H4k20me1
H3k9me1
H2az
H3k79me2
H3k36me3
H3k4me3
H3k9ac
H3k4me2
H3k27ac
H3k4me1
H3k9me3
MeMRE
MeDIP
MeBS
CpGi
CpGe
Me450
TFBS
REG
FAIRE
DNase
Gene
EXPS
LRNA+
LRNA−
LINC
SVA
Alu
L1
Retrotransposon
Genetic markers
Histone
CpG
Regulatory sites
Gene / Transc.
Fold Change
−1
0
1
2


C
1.77
1.35
D
V
2.55