Chapter 4
Population-Based CNV Study in Schizophrenia

4.1. Experimental Design and Array Data Quality Control

4.1.1 Case-Control CNV Screen Experimental Design

In this chapter we will present a population based case-control CNV study design to detect schizophrenia-related CNVs. The CNV screen was mainly based on a core CNV dataset generated by hybridizations of 91 schizophrenia samples (SCZ) and 92 ethnically matched Lothian Birth Control Cohort samples (LBC) on WGTP arrays. The CNV data was then analysed for rare copy number variants specific to the SCZ cohort, and frequent copy number variants for disease association study.

A subset of the WGTP CNV data was validated by independent platforms, including quantitative PCR and high resolution oligonucleotide array (Nimblegen, Inc.). Furthermore, 39 patients in the SCZ cohort were hybridized on the Affymetrix SNP array 6.0 platform, and CNV data was included for validation of the WGTP dataset (source: University of Edinburgh, as part of International Schizophrenia Consortium).

WGTP array CGH data was analyzed using the algorithm CNVFinder (Fiegler et al. 2006) (discuss in Chapter 2). Affymetrix SNP array data was analyzed using the Birdseye package (Korn et al. 2008) and PLINK (Purcell et al. 2007). The ascertainment scheme for the Affymetrix CNV data emphasized rare, large CNV events and therefore represents a more conservative calling algorithm than the WGTP data. The SNP platform also provided higher-resolution data and facilitated the mapping of CNV breakpoints with higher precision.

Affymetrix SNP array data on a further 206 SCZ samples (from the same DNA source as our initial cohort) was used to investigate recurrent variants in SCZ which are more likely to be disease-relevant.

Figure 4.1 shows a summary of the case-control DNA samples and the CNV detection platforms used in the current study.

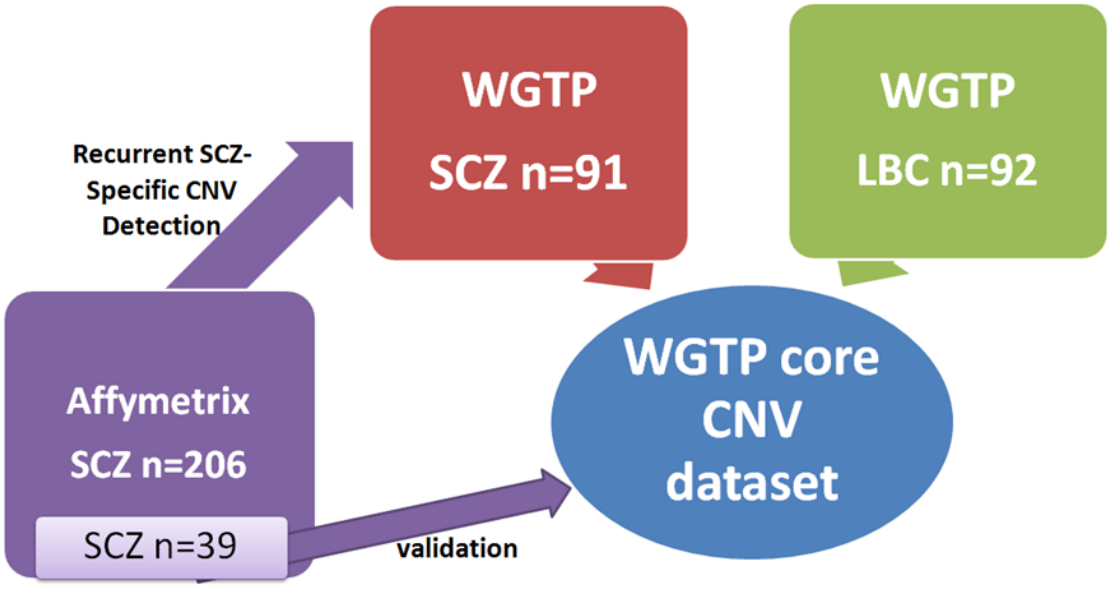


Figure 4.1 Case and control cohorts and CNV detection platforms used in our study

4.1.2 WGTP Array Data Quality Control

BAC array based comparative genomic hybridization has been routinely used for the detection of copy number variations in clinical samples. Normalization and filtering of the array data is critical to enhance the discovery of true biological signals, and to remove experimental artefacts for accurate and reliable results. We performed initial normalization on the WGTP data as previously described (and as documented in section 2.1) (Fiegler et al. 2006). Additional normalization and filtering steps, including correction for GC content and filtering clones of artefacts, were summarized in Figure 4.2, with details included in Appendix C.

In addition to manual inspection of profile qualities, all hybridization results were monitored with two statistical indicators for quality control. The first is global SDe, an estimate of standard deviation of all log2ratio signals in the genome for a particular sample profile. The second is the clone retention rate after fusion of dye swap experimental data. Experiments were accepted for further analysis only if global SDe<0.06, global clone retention rate >90% and clone exclusion rate per individual chromosome <80%. Figure 4.3 presents the quality control indicators for all SCZ and LBC hybridizations included in the CNV analysis.

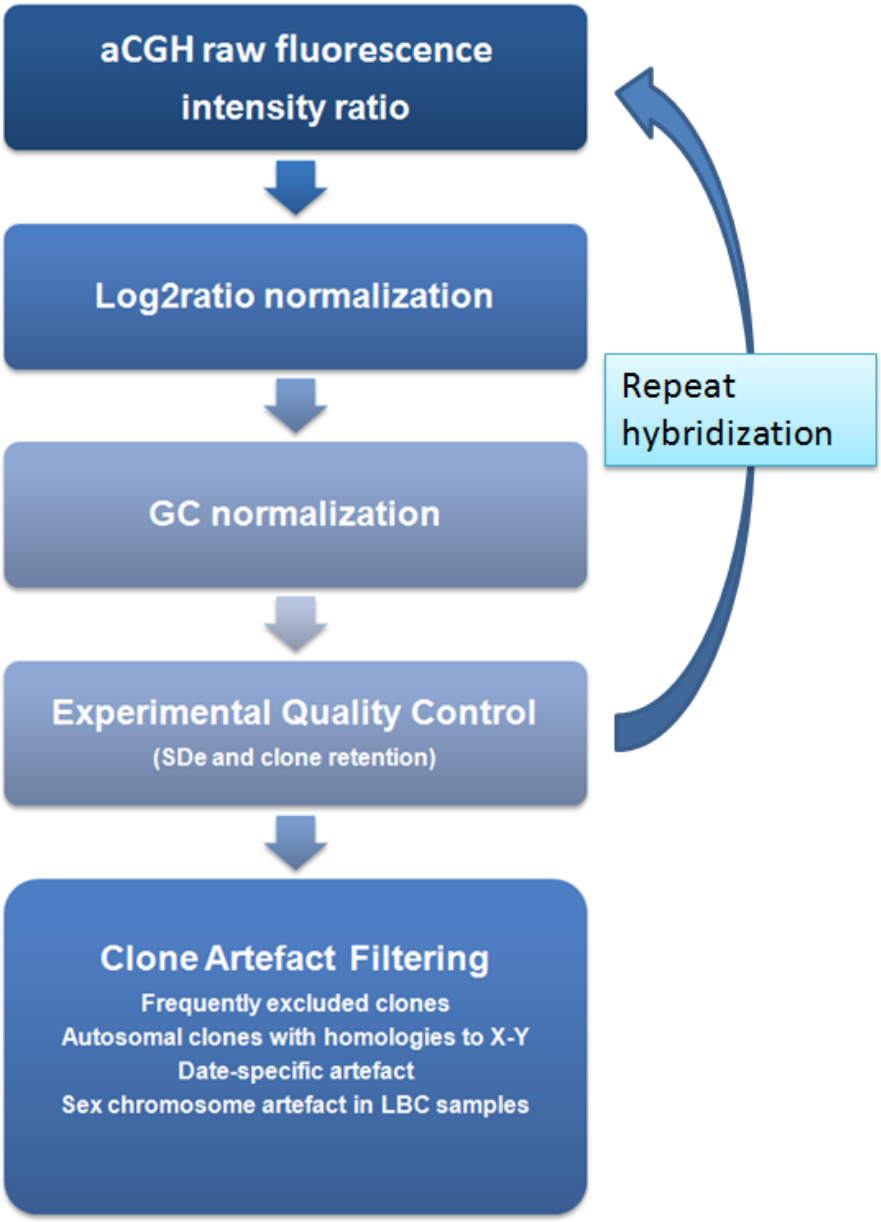


Figure 4.2 Normalization and filtering steps applied to WGTP hybridization data before CNV analysis.

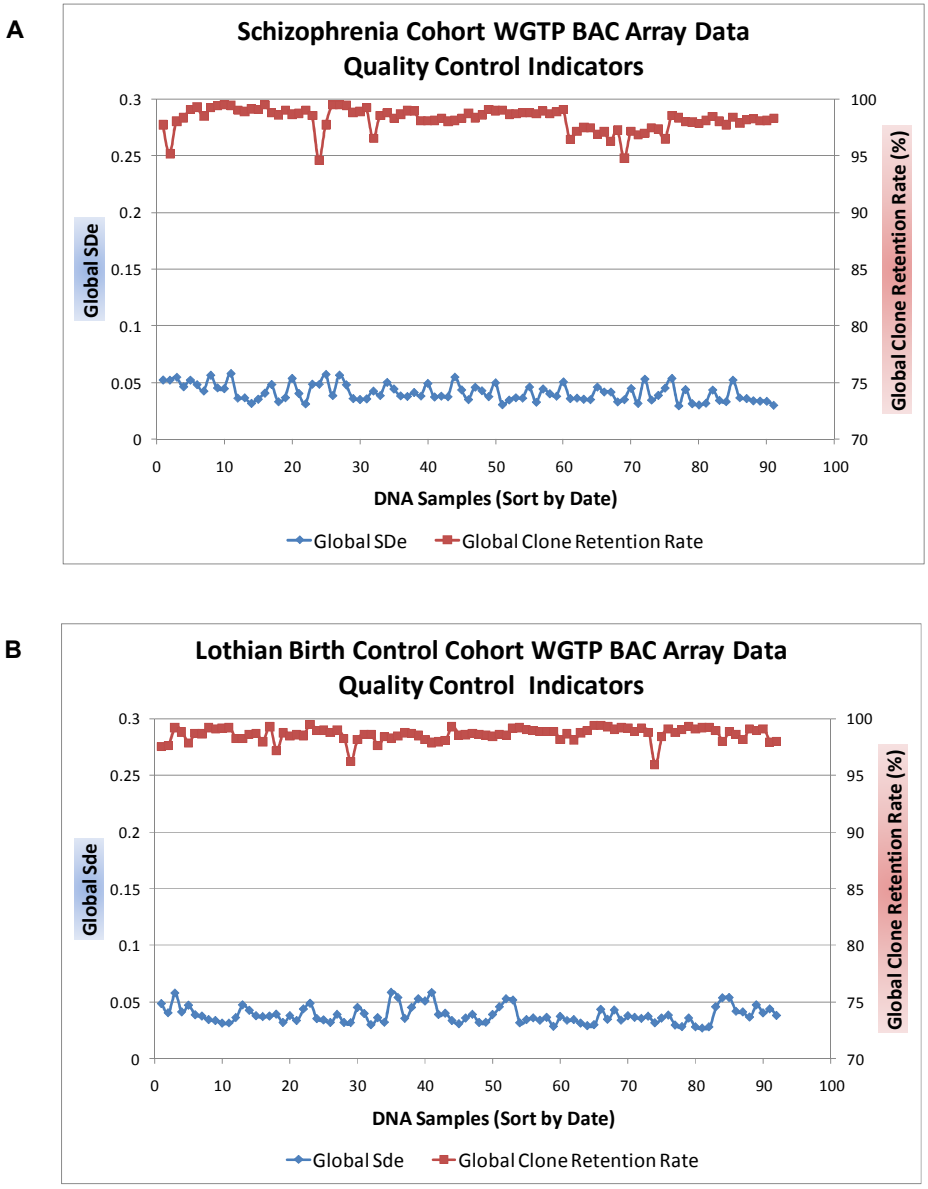


Figure 4.3 WGTP array quality control indicators. Global SDe (blue) and global clone retention rate (red) are plotted against DNA samples (sorted by date) for all samples included in analysis. **a)** SCZ (n=91); **b)** LBC (n=92)
(Global SDe cut-off point was <0.06. Global clone retention rate cut-off point was >90%)

4.2 Copy Number Variation Detection on the WGTP array

4.2.1 Distribution of CNVs and CNV Regions in Case and Control Cohorts

The core copy number variation dataset in the study was generated from DNA samples hybridizing against the WGTP array platform. On average we detected 66 CNVs per SCZ sample (case) and 72 CNVs per LBC sample (control). A list of per-patient CNV information (number and types of CNVs) is given in Appendix D-1.

The total number of CNV calls was 6,036 in the cases (3,650 gains, 2,386 losses) and 6,666 in the controls (3,973 gains 2,689 losses) (Figure 4.4a). CNV size distributions were similar in the two cohorts (Figure 4.5), with the same median length of 237.7 kb. CNV sizes were estimated from the chromosomal coordinates of affected BAC clones. Since average BAC clone size was 164 kb, the reported CNV size may be an overestimation of the true CNV coordinates (McCarroll et al. 2008; Perry et al. 2008).

In each sample cohort, individual CNVs from for multiple individuals may overlap at the same genomic location. To generate sets of non-overlapping CNV genomic locations, we grouped for each cohort all overlapping CNVs into copy number variation regions (CNVRs). The total number of CNVRs was 539 in SCZ (249 gain only; 159 loss only; 131 gain or loss) and 583 in LBC (286 gain only, 127 loss only, 170 gain or loss) (Figure 4.4 b&c).

CNVRs were classified as either “rare” (occurring in only one sample) or “recurrent” (occurring in multiple samples) with respect to the cohort. There were 224 rare and 315 recurrent CNVs in the SCZ cohort, compared with 237 rare and 346 recurrent ones in LBC. The proportion of rare and recurrent CNVs in the cases and controls (Figure 4.4b) showed no statistically significant difference. Distributions of gene content were also

similar for cases and controls (Figure 4.6), with majority of CNVRs contain one or more genes in both SCZ (355 out of 539 CNVRs; 66%) and LBC (390 out of 583; 67%).

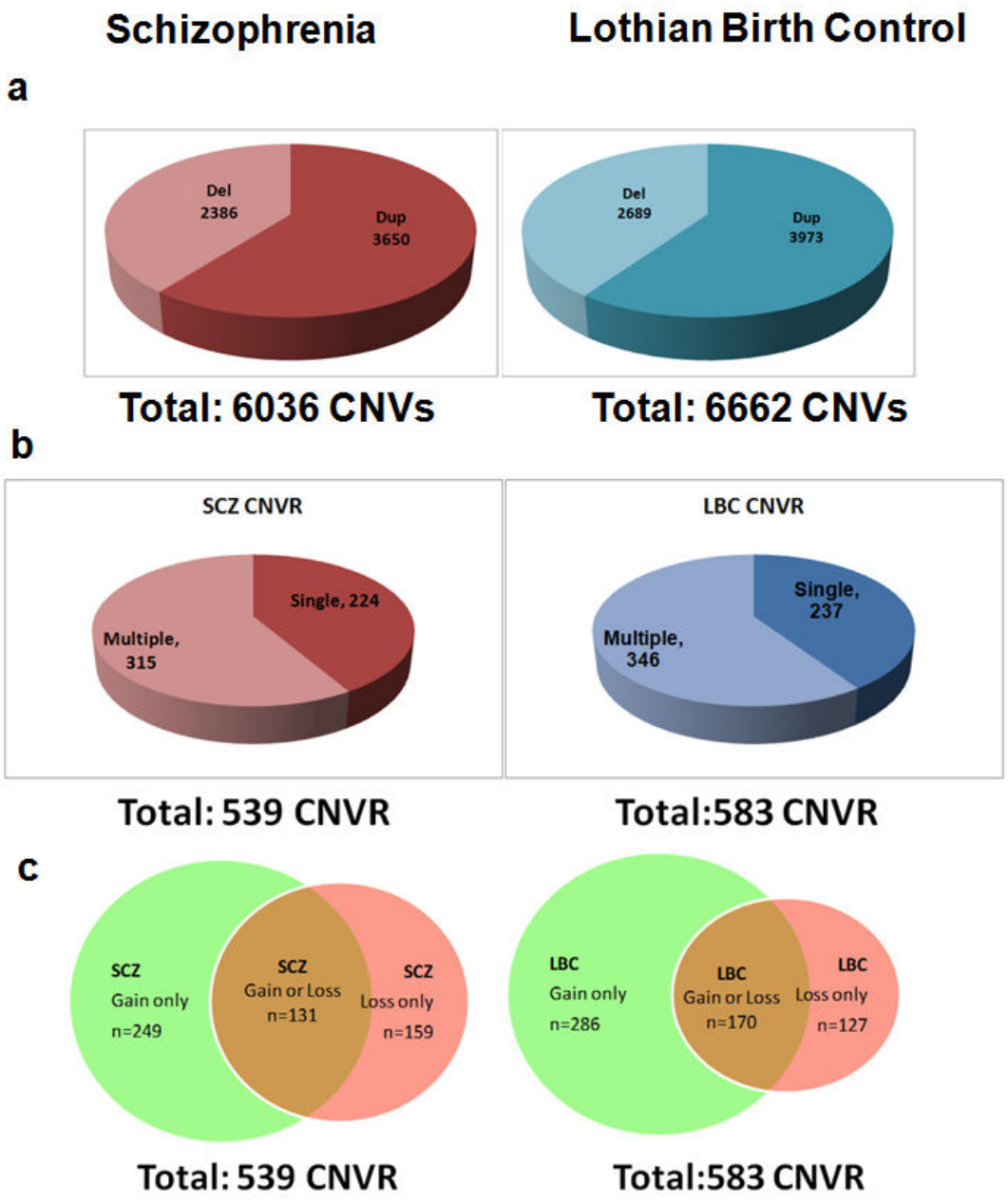


Figure 4.4: Frequency and types of CNVs and CNVRs in SCZ and LBC. a) Number of CNVs (gains/losses) in SCZ and LBC **b)** Number of CNVRs (rare/recurrent) in SCZ and LBC **c)** Number of CNVRs (gains/losses) in SCZ and LBC

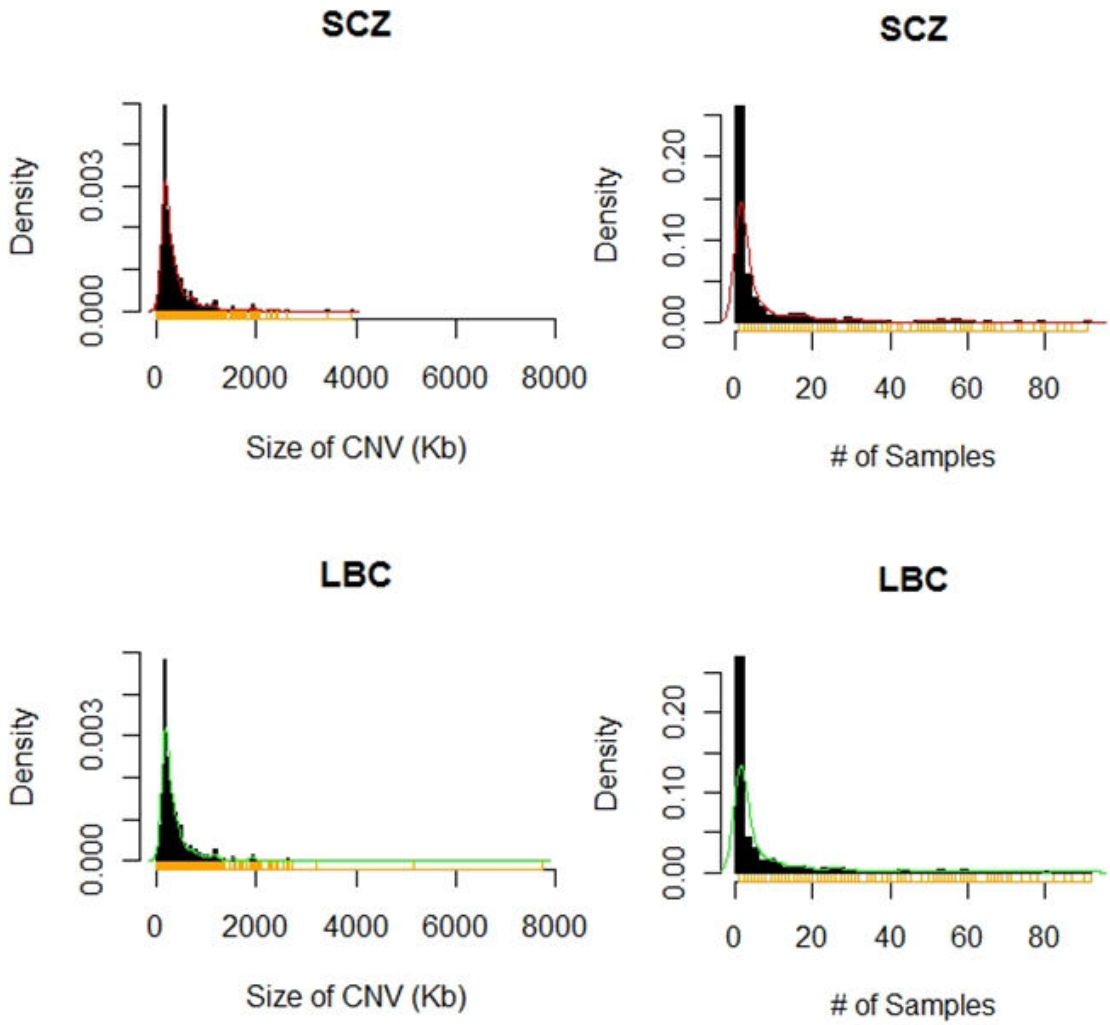


Figure 4.5 Size and frequency distributions of CNVRs in Schizophrenia (SCZ) and Lothian Birth Control (LBC) Cohorts detected using the WGTP platform.
***Top left*:** CNV size distribution in SCZ
***Top right*:** CNVR frequency distribution in SCZ
***Bottom left*:** CNV size distribution in LBC
***Bottom right*:** CNVR frequency distribution in LBC

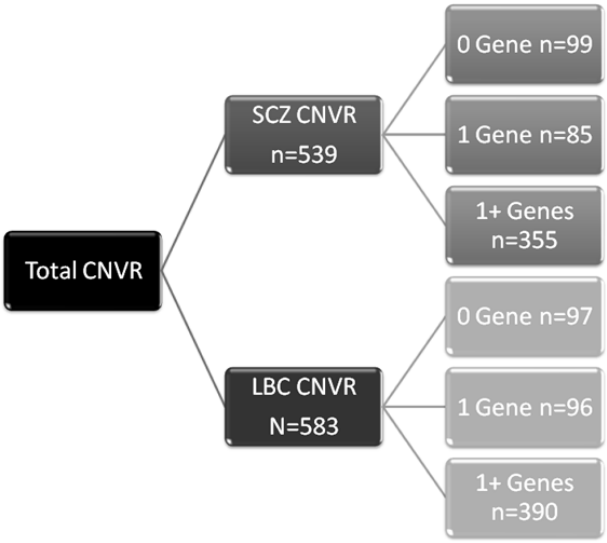


Figure 4.6 CNVR gene content in the SCZ and LBC cohorts.

4.2.2 Bias of CNV Discovery Rate in SCZ Versus LBC

Although CNVs/CNVRs in our cases and control data showed comparable distributions in terms of CNV size, frequency and gene contents, we detected more CNVs/CNVRs in the control cohort than in patients with schizophrenia (10.4% more CNVs in total, 9.10% more CNVs per sample). This bias in CNV discovery rate was mainly due to differences in DNA quality, leading to variability in the quality of array CGH profiles in the two cohorts (with control DNA samples of better quality on average). Lower DNA quality increases the fluorescence intensity log2ratio variance (SDe) and consequently decreases the sensitivity of CNV detection by CNVFinder.

Figure 4.7 shows a negative correlation between number of CNVs detected and estimated standard deviation SDe. The regression (logarithmic) lines fitted on the SCZ and LBC correlation data are comparable (Fig 4.7a). However, because the SDe distribution for LBC samples shows a slight shift towards lower values (mean SDe in:

SCZ = 0.41; LBC= 0.038) (Fig 4.7b), we detected on average a higher CNV number in the control cohort.

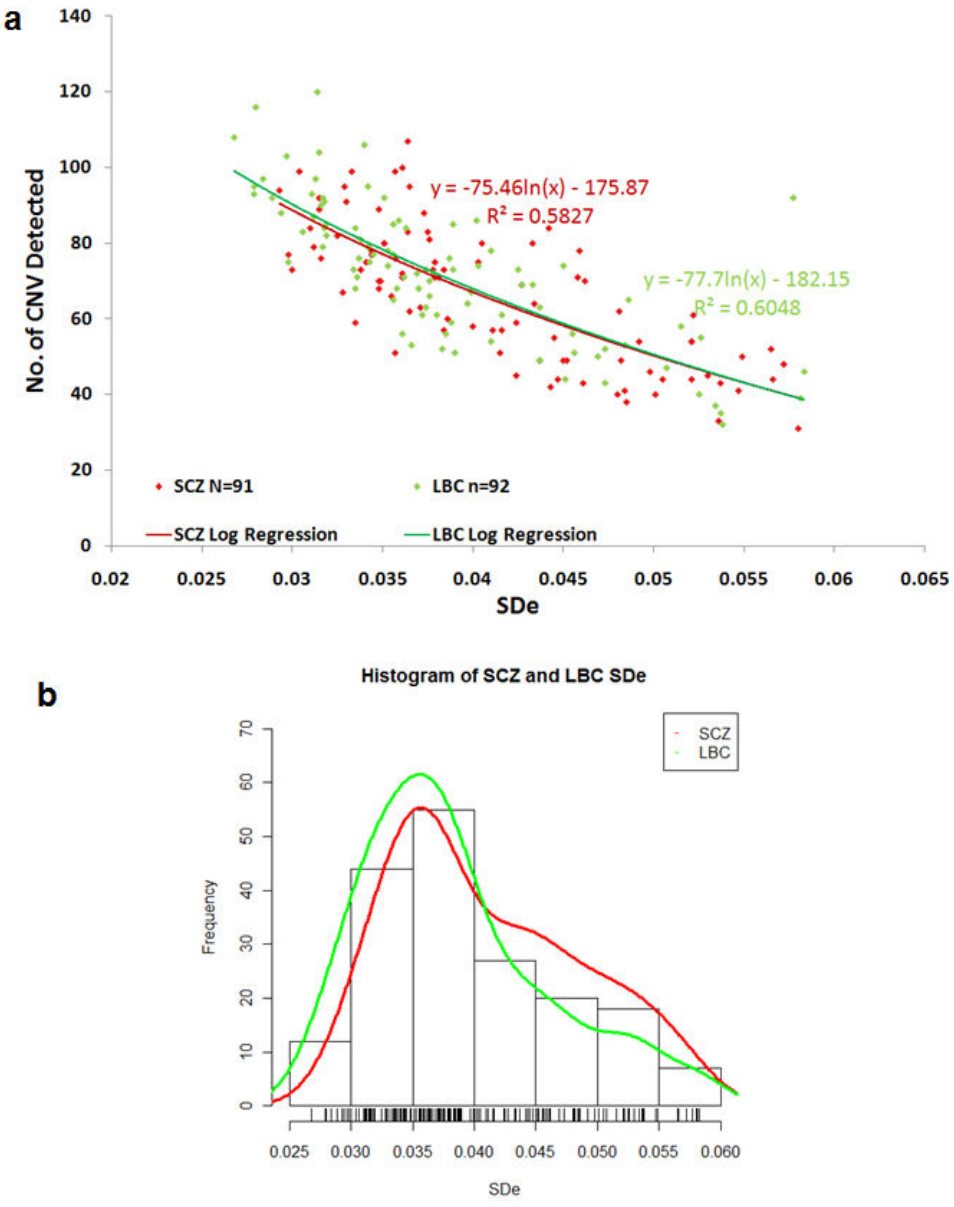


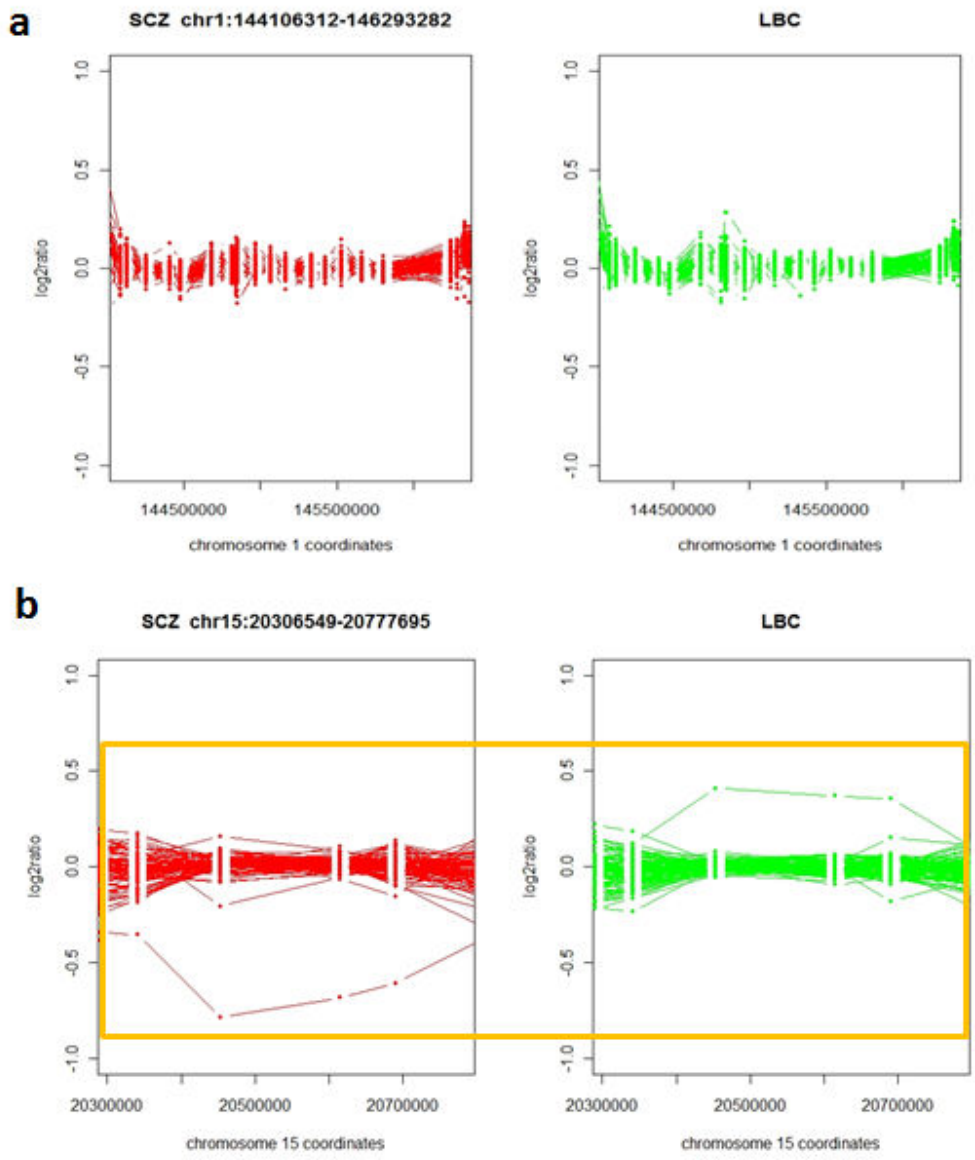
Figure 4.7 Correlation of CNV discovery rate with data quality in SCZ and LBC. a) Number of CNVs detected in each sample against profile variability (SDe). Regression fit showed negative correlation for both SCZ (red) and LBC (green) data. **b)** SDe distribution for SCZ and LBC experiments. Mean SDe is 0.041 in SCZ and 0.038 in LBC, indicating an overall higher quality of the LBC data.

4.3 Comparing WGTP Data with Known Schizophrenia CNV Regions

Novel findings of large-scale genome-wide schizophrenia CNV screen (ISC 2008; Stefansson et al. 2008) revealed 3 recurrent CNV schizophrenia loci at 1q21.1, 15q11.2 and 15q13.3. These 3 recurrent deletions, together with the previously identified 22q11 deletion, represented 4 CNV regions that demonstrated statistical significant association with schizophrenia. We therefore screened the WGTP CNV data (SCZ and LBC) against these 4 genomic loci.

Figure 4.8 illustrates the SCZ (red) and LBC (green) WGTP data at these 4 loci. Of note, we detected the same recurrent deletion at 15q11.2 as Steffanson *et al.* (Stefansson et al. 2008), as shown in one sample in the SCZ cohort. The region was also duplicated in one LBC sample (Figure 4.8b). Given the estimated frequency of 0.55% and 0.19% in cases and controls (Stefansson et al. 2008), we expected the deletion to occur in 0 or 1 sample in our cases, whilst being absent in the controls, which was consistent with our finding. This also confirmed the ability of our platform to detect disease-associated rearrangement of similar size and frequency. As described in section 1.6, the recurrent deletion was flanked by LCRs (Figure 4.9). Within the deletion was a candidate gene *CYFIP1* (cytoplasmic FMR1 interacting protein 1 isoform), which encoded a protein involved in the regulation of translation in neurons, with important functions in synaptic plasticity and brain development (Napoli et al. 2008) (see section 1.6).

At the other three known regions, 1q21.1, 15q13.3 and 22q11, no large deletion was identified in our SCZ cohort. Smaller reciprocal duplications were detected at 1q21 and 22q11, although these have been reported in control cohorts at equal or higher frequency than in corresponding case cohorts (ISC, 2008), and therefore are likely to be benign.



(Figure 4.8 to be continued)

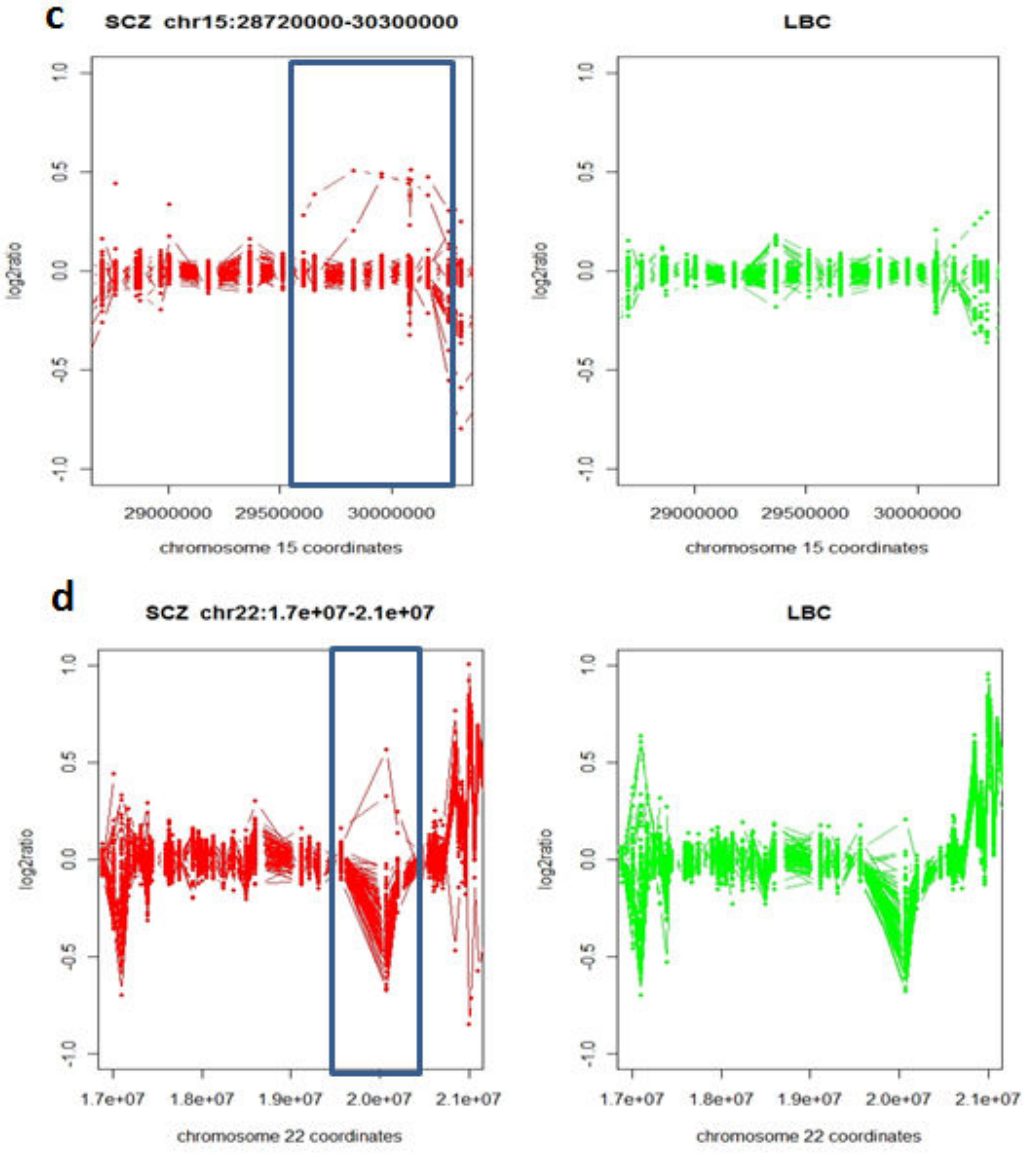


Figure 4.8 WGTP data compared with 4 known schizophrenia CNV loci. SCZ data was presented in red and LBC in green **a)** 1q21.1: no large rearrangement detected **b)** 15q11.2: a large deletion was detected in SCZ, and a reciprocal duplication in LBC (orange box) **c)** 15q13.3: two smaller duplications were detected in SCZ (blue box); **d)** 22q11: two smaller duplications were detected in SCZ (blue box)

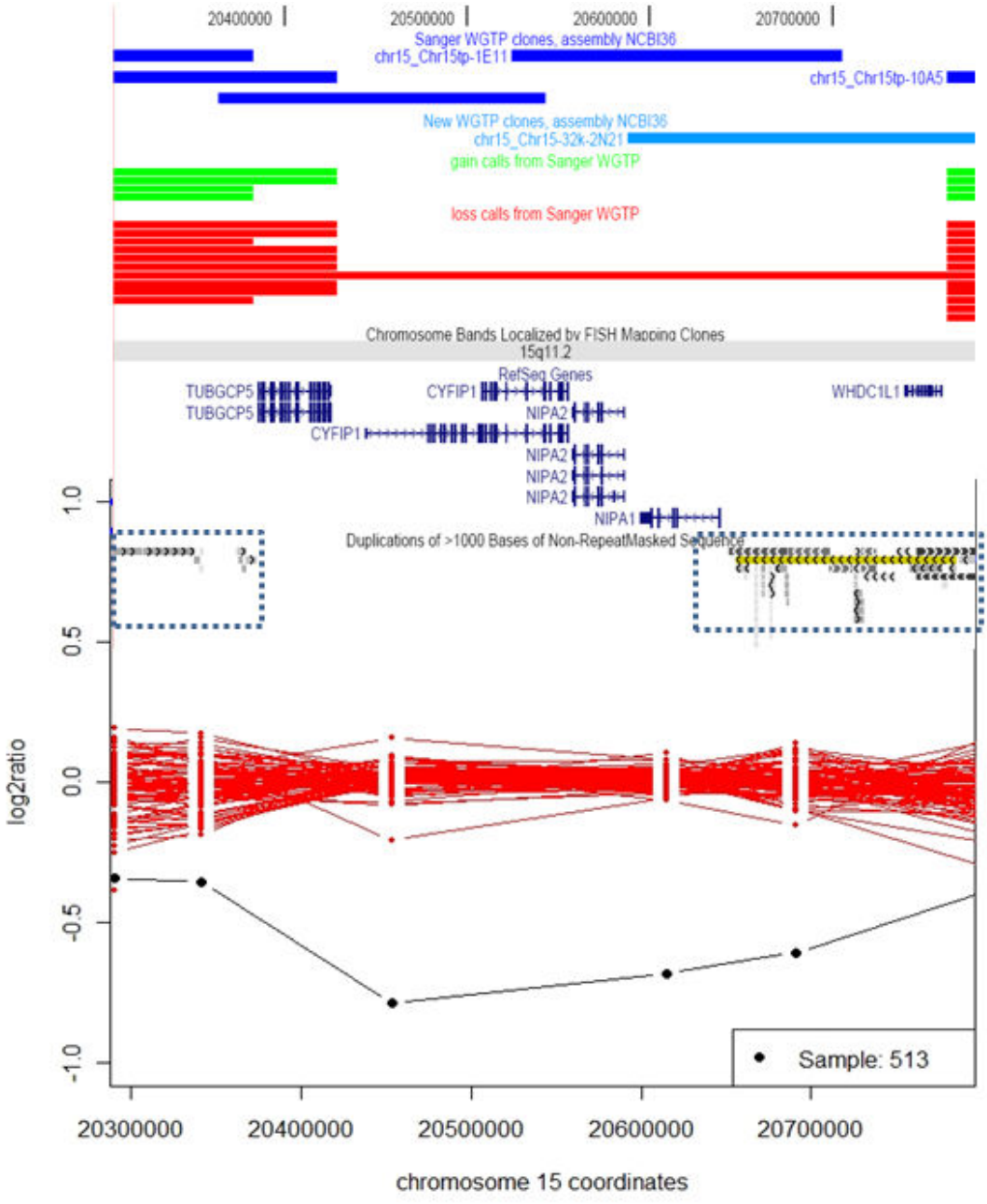


Figure 4.9 WGTP data detected a deletion at 15q11.2 in one patient. The deletion spans gene *CYFIP1*. Regions of low copy repeats (LCR) are highlighted in blue boxes.

4.4 Rare Variants Specific to the Schizophrenia Cohort

With evidence that rare variants significantly associated with schizophrenia were replicated in our WGTP dataset, we concentrated our first approach on rare CNVs that were only detected in schizophrenia samples. This approach was also reinforced by recent CNV findings detecting a higher occurrence of rare, and in some cases *de novo* CNVs in cases compared to controls (ISC 2008; Walsh et al. 2008; Xu et al. 2008). The International Schizophrenia Consortium, for instance, reported a 1.448-fold increase of extremely rare CNVs (singletons in 3000+ cases) in cases versus control. Another report from Walsh *et al.*, suggested the increase of rare genic CNVs was as high as 3-fold in cases versus control. Extrapolating from these data, one could hypothesize that 1/3 to 2/3 of the rare case-specific CNVs would be disease-causing variants.

4.4.1 Rare Variant Detection Using Consecutive Clone Calling Criteria

In our rare variant analysis we included only CNVs involving more than one consecutive clone. This is a common practise in CNV studies to increase data stringency. The corresponding CNVs[2] were grouped into 324 CNVRs in cases and 360 CNVRs in controls.

Less than 1/3 of the CNVRs were cohort-specific (113 in cases and 141 in controls), indicating that majority of variants detected were recurrent copy number polymorphisms in the Scottish population (existing in both Scottish cases and controls).

The 113 SCZ cohort-specific CNVs are listed in Appendix D-2.

[2] This set of CNVs includes a total of 3,980 CNVs in the cases and 4,444 CNVs in the controls.

4.4.2 Validation of SCZ-Specific Rare Variants

A reliable detection platform is crucial in CNV discovery. The WGTP array CGH platform available to our study has previously been tested extensively in terms of accuracy and sensitivity for CNV discovery (Fiegler et al. 2006). We performed additional validation experiments specific to our dataset through qPCR, Nimblegen oligonucleotide array and with Affymetrix SNP array data provided by the ISC.

Validations were performed for two purposes: 1) to confirm the presence of a subset of CNV calls as a measure of platform detection accuracy; and 2) to delineate CNV boundaries and assess the involvement of candidate genes. 29 regions were validated by a combination of these methods (Table 4.1). Details of validation experiments and validation rates were provided in Appendix E.

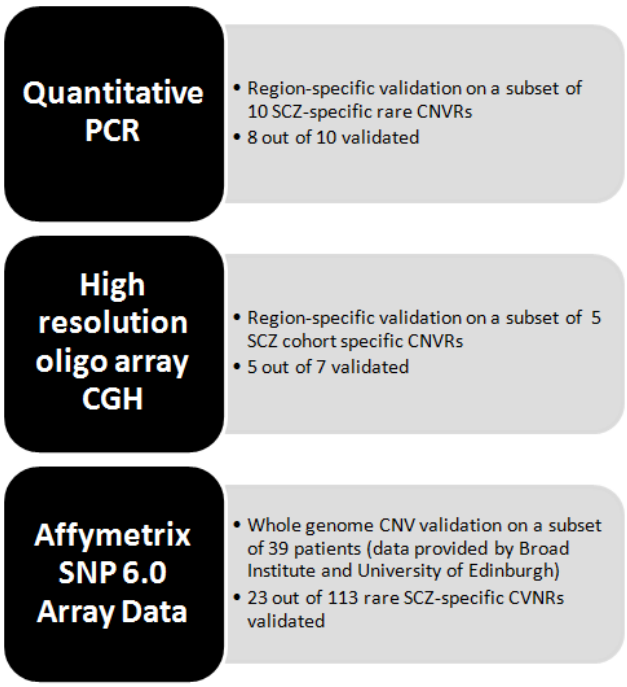


Figure 4.10 WGTP CNV dataset validation strategies.

Table 4.1 Validated SCZ-specific rare variants. 29 regions were validated by a combination of qPCR, Nimblegen oligonucleotide array and Affymetrix SNP array. (***Red***: *DNA samples used in validation.* ***Black***: *genes deleted/duplicated as refined by higher resolution CNV detection platform*)

WGTP Chr coordinates	Genes	Refined Chr Coordinates by higher resolution array	Samples	Validation			CNV Type	
				qPCR	Nimb legen	Affy	gain/ loss	rare/ recurrent
chr5:110212743-110577198	WDR36, TSLP	N/A	**850**	√			loss	rare
chr6:22212251-22607763	PRL	N/A	**5324**	√			loss	rare
chr7:93658283-94064936	SGCE, CASD1, COL1A2	N/A	**4203**	√			loss	rare
chr8:9670799-9988807	TNKS, MSRA	chr8:9830000-9870000	**3789**	√	√		loss	rare
chr9:137112070-137464520	C9orf62, OLFM1	chr9:137280000-137450000	**323**, 4179, 5758	√	√		gain	recurrent
chr3:86998287-87362355	**VGLL3**, CHMP2B	chr3:86898287-87462355	**5386**		√		gain	rare
chr1:68325265-68897278	**DEPDC1,GPR177,RPE65**	chr1:68334011-68755246	**3766**			√	gain	rare
chr1:71346767-71893785	**NEGR1**	chr1:71379226-71687039	**3766**	√		√	gain	rare
chr2:127333465-127814482	**BIN1,CYP27C1**,ERCC3,MAP3K2	chr2:127182184-127688195	**3975**			√	gain	rare
chr3:143326111-143612526	**GK5,TFDP2,XRN1**	chr3:143303643-143554513	**4710**			√	gain	rare
chr3:189546564-189761158	**LPP**	chr3:189564871-189682182	**3945**, 7183			√	loss	recurrent
chr4:11668861-11990799	N/A	chr4:11838410-11948682	**4748**			√	loss	rare
chr4:28258973-28570510	N/A	chr4:28415433-28541944	**3945**			√	loss	rare
chr5:25051483-25333181	N/A	chr5:25121394-25405536	**1295**			√	gain	rare
chr6:1994540-2304998	**GMDS**	chr6:2001510-2293508	**3815**			√	gain	rare
chr6:97469364-98475130	**C6orf167,KLHL32**	chr6:97426034-98410421	**1085**	√		√	loss	rare
chr6:165840858-166673790	**PDE10A,PRR18,SFT2D1,T**	chr6:165893867-166666551	**5758**			√	gain	rare
chr7:17795285-18873121	**HDAC9,SNX13**	chr7:17674099-18634682	**5541**	√		√	loss	rare
chr7:88344723-88962788	**ZNF804B**	chr7:88292092-88731204	**1295**			√	loss	rare
chr8:13581606-13896999		chr8:13684044-13829803	**5541**			√	loss	rare
chr11:31339430-31549584	**DCDC1,DPH4,ELP4,IMMP1L**	chr11:31312028-31554033	**1278**			√	gain	rare
chr12:17761729-18154145	**RERGL**	chr12:17776260-18038013	**3409**	√		√	loss	rare
chr13:112568274-112939870	**ATP11A,F1,F7,MCF2L,PCID2,PROZ**, CUL4A	chr13:112541419-112904310	**385**			√	gain	rare
chr13:113166604-113658050	GRTP1,ADPRHL1,**ATP4B,DCUN1D2,TFDP1,TM CO3**,FAM7B,GAS6,GRK1	chr13:113051168-113366491	**6638**			√	loss	rare
chr14:74945401-75396561	JDP2,**BATF,C14orf1,FLVCR2,TTLL5**	chr14:75046923-75379455	**4100**			√	gain	rare
chr15:99550338-100036184	CHSY1,,SNRPA1,**PCSK6**,TARSL2,TM2D3	chr15:99764132-99956872	**6638**, 385, 7294			√	gain	recurrent
chr20:14561834-14936485	**MACROD2**	chr20:14650902-14813485	**4716**			√	loss	rare
chr21:34576661-34852873	**C21orf51,KCNE1,KCNE2,RCAN1**	chr21:34644865-34827865	**899**			√	gain	rare
chr21:44384864-44825939	AIRE,C21orf2,C21orf33,DNMT3L,ICOSLG,PFKL, TRPM2,LRRC3,**C21orf29,C21orf9,KRTAP1-1,KRTAP1-2,KRTAP1-3,KRTAP1-4,KRTAP1-5**,KRTAP1-6, KRTAP1-7, KRTAP1-8,KRTAP1-9,KRTAP1-10	chr21:44730300-44890569	**5307**			√	gain	rare

4.4.3 Rare Variants in SCZ with Genes Involved in Psychiatric Disorders

One way to assess whether a variant is likely to be disease-causing is to examine its gene content. A query of the Genetic Association Database (GAD) provided by the National Institutes of Health ascertains 8 SCZ cohort-specific CNVRs to contain genes represented in the PSYCH (psychiatric-related) category. These genes are *AHR* (aryl hydrocarbon receptor), *TRPM2* (transient receptor potential cation channel, subfamily m, member 2), *DMPK* (dystrophia myotonica-protein kinase), *PIK3C3* (phosphoinositide-3-kinase, class 3), *GABRG3* (gamma-aminobutyric acid a receptor, gamma 3), *OXTR* (oxytocin receptor), *SGCE* (sarcoglycan, epsilon) and *KIF2A* (kinesin heavy chain member 2) (Table 4.2).

In particular, *PIK3C3*, a gene encoding the phosphoinositide-3-kinase, was detected as duplicated in patient 3857 (ED1013). The PI3K kinase is responsible for receptor-mediated signal transduction and intracellular trafficking. Neuregulin-1, a putative schizophrenia susceptibility gene, was suggested to regulate cell adhesion through a PI3K dependent pathway (Kanakry et al. 2007). *PIK3C3* itself was a putative schizophrenia and bipolar disorder candidate, supported by a number of genetic association studies (Stopkova et al. 2004; Duan et al. 2005; Saito et al. 2005; Tang et al. 2008).

SGCE, a gene detected as deleted in patient 4203, encodes transmembrane components of the dystrophin-glycoprotein complex. *SGCE* is involved in Myoclonus-Dystonia (Marechal et al. 2003; Valente et al. 2005; Misbahuddin et al. 2007), a movement disorder characterized by myoclonic jerks, dystonia and a variety of psychiatric symptoms including anxiety, depression and obsessive-compulsive disorder (Doheny et al. 2002). Furthermore, *SGCE* knockout mice were shown to have anxiety,

depression and altered dopamine levels, all phenotypes associated with schizophrenia (Yokoi et al. 2006).

Oxytocin Receptor (*OXTR*), another gene identified as duplicated in the SCZ cohort, encodes a neurohypophyseal hormone important in brain development (Insel et al. 1999). The role of oxytocin was highlighted in mammalian social behaviours (Israel et al. 2008). Abnormalities of the oxytocin system have been implicated in several neuropsychiatric disorders, including schizophrenia, autism , obsessive-compulsive disorder and post-traumatic stress disorder (Marazziti and Catena Dell'osso 2008). Pharmacologically, oxytocin has been proposed as an antipsychotic drug (Caldwell et al. 2008). Oct -/- mice demonstrated social deficit (Winslow and Insel 2002). Furthermore, mouse models suggested that the oxytocin system is involved in the response to phencyclidine (Lee et al. 2005; Caldwell et al. 2008), a drug when administrated lead to schizophrenia-like behavioural symptoms including hallucination. Schizophrenia endophenotypes such as prepulse inhibition deficits were augmented in the Oct knockout mice (Caldwell et al. 2008).

Table 4.2 SCZ-specific rare variants with genes associated with psychiatric disorders. Genes within CNVRs were compared to the Disease Association Database returning 8 CNV regions with genes related to psychiatric disorders (disease class: PSYCH).

CNVR	GENE	GENE DESCRIPTION	LOCATION	CYT BAND	DISEASE CLASS: PSYCH
3815-gain-14	OXTR	oxytocin receptor	chr3: 8686094-8880344	3p25	attention deficit disorder conduct disorder oppositional defiant disorder, autism
5386-loss-27	KIF2A	kinesin heavy chain member 2	chr5: 61438895-61723574	5q12-q13	schizophrenia
5324-gain-28	AHR	aryl hydrocarbon receptor	chr7: 17320942-17609921	7p15	dementia
4203-loss-28	SGCE	sarcoglycan, epsilon	chr7: 93658283-94064936	7q21-q22	Tourette syndrome; obsessive compulsive disorder, dystonia
5660-gain-61	GABRG3	gamma-aminobutyric acid (gaba) a receptor, gamma 3	chr15: 25060215-25335769	15q12	alcohol dependence, autism
ED1013-gain-73	PIK3C3	phosphoinositide-3-kinase, class 3	chr18: 37767209-38078089	18q12.3	schizophrenia, bipolar disorder
4100-loss-63	DMPK	dystrophia myotonica-protein kinase	chr19: 50553125-51160225	19q13.3	myotonic dystrophy
5307-gain-54	TRPM2	transient receptor potential cation channel, subfamily m, member 2	chr21: 44384865-44825939	21q22.3	bipolar disorder

In the rare variants specific to the SCZ cohort, we also detected several CNVs contain brain- or neuronal-related genes not currently associate with the Genetic Association Database. These CNV loci overlapped with genes *NEGR1, PDE10A, HDAC9 and RCAN1* (Figure 4.11).

The neuronal growth factor, *NEGR1*, was duplicated in one patient. It is brain-expressed and was suggested to be involved in cell adhesion (Figure 4.11a). *PDE10A*, a phosphodiesterase involved in corticostriatal signalling, was duplicated in another patient. *PDE10A* plays a role in cognition, locomotion and behavioural phenotypes in mice (Hebb et al. 2008), and *PDE10* inhibitor has been studied as a novel drug treatment for the cognitive symptoms of schizophrenia (Menniti et al. 2007; Schmidt et al. 2008) (Figure 4.11b).

HDAC9, a histone deacetylase which couples neuron-induced electrical activity to muscle cells (Mejat et al. 2005), was disrupted at the deletion breakpoint of a patient (Figure 4.11c). *RCAN1*, a calcineurin inhibitor involved in calcium-mediated signalling, was disrupted at the duplication breakpoint of another patient (Figure 4.11d). These two variants were further demonstrated to be recurrent in an extended schizophrenia cohort and will be discussed in section 4.4.

Given our modest sample size, we were not able to determine a statistically valid association of any of these rare regions with disease. Rather, our experimental design was aimed at cataloguing SCZ-specific CNVs as a collective dataset with candidate genes of disease relevance. Based on previous studies proposing a large proportion of case-specific rare variants being involved in disease, and the evidence of biologically plausible candidates, this set of genes represents a set of good candidates for future

functional investigation. Accruing CNV data from the literature and collaborative efforts, our dataset can also be extended to confirm disease-causing recurrent CNV loci, as in the case of *CYFIP1* deletion at 15q11.2.

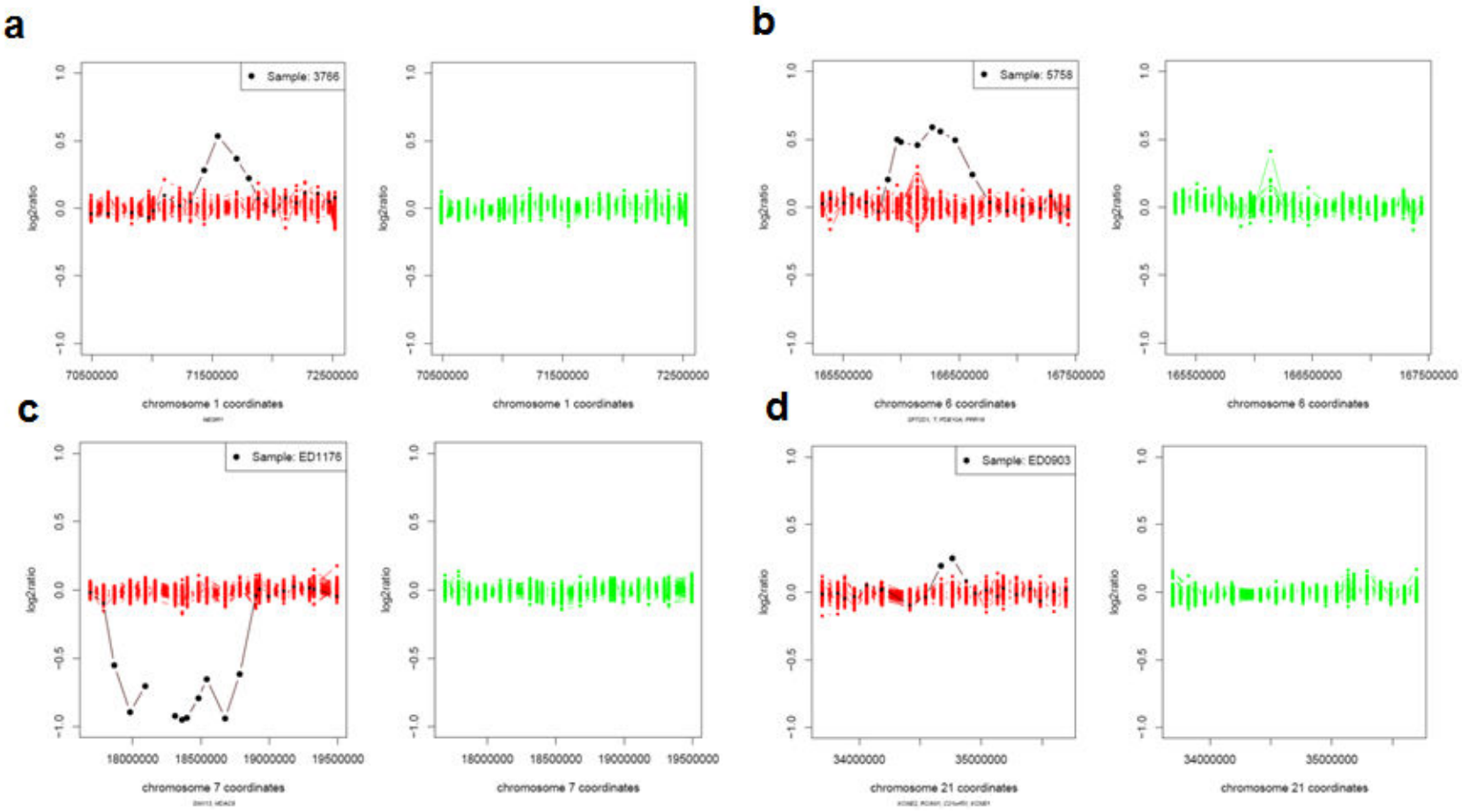


Figure 4.11 Schizophrenia cohort-specific CNVRs containing brain-related or neuronal-related genes. a) Duplication at *NEGR1* (Neuronal growth factor 1) **b)** Duplication at *PDE10A* (Phosphodiesterase 10A) **c)** Deletion disrupting *HDAC9* (histone deacetylase 9) **d)** Duplication spanning *DSCR1/RCAN1* (Down Syndrome Critical Region 1)

4.5 Recurrent SCZ-Specific Variants in Extended Cohort

Among the SCZ-specific variants detected by WGTP, the majority were singletons in the cohort of 91 SCZ samples (see last column in Table 4.1). To detect recurrent SCZ-specific CNVs, we extended our cohort using CNV calls made from the Affymetrix SNP 6.0 array as provided by ISC and University of Edinburgh. The dataset consists of an additional 206 Scottish schizophrenia samples hybridized on the Affymetrix array (SCZ_{Affy}n206). Combined with our original WGTP SCZ cohort of 91 (SCZ_{WGTP}n91), the extended cohort consists of 297 patient samples ($SCZ_{WGP\text{-}Affy}$n297).

A recurrent SCZ CNVR was defined as any CNV region that was detected in the original 91 WGTP dataset (SCZ_{WGTP}n91), and was present in two or more patients in the combined cohort of 297 ($SCZ_{WGP\text{-}Affy}$n297).

A total of 26 recurrent SCZ CNVRs were identified, with a frequency of 2 to 6 patients out of 297 (0.67- 2.02%) (Table 4.3). Although they were all relatively rare events, the observation of these recurrent, putative disease variants is consistent with the genetically heterogeneous model of schizophrenia, and the frequency range is comparable to the rare disease variants observed in previous CNV studies on schizophrenia (Rujescu et al. 2008; Stefansson et al. 2008).

We focussed on 7 variants which were reported in the same patient in both WGTP and Affymetrix platforms (Figure 4.12). Among them, a region at chromosome 21 (CNVR-480) and another at chromosome 9 (CNVR-297) were detected in 4 and 5 patients respectively. Both of these recurrent CNVRs had a frequency of >1% in the disease cohort, and were found to overlap with functionally relevant candidate genes namely *RCAN1* (Down syndrome critical region 1) and *OLFM1* (olfactomedin 1).

Table 4.3 Recurrent SCZ-specific variants as detected in 297 SCZ samples. CNVs were detected by WGTP and/or Affymetrix platforms. Regions overlapping with more than one entry in the Database of Genomic Variant (DGV) for >50% were indicated in the last column. *(**red**: Samples detected by both platforms; blue: samples detected by WGTP only; green: samples detected by Affymetrix only)*

WGTP Identifier	WGTP coordinates	AFFY coordinates	Samples	Genes	DGV
CNVR_97	chr3:189546564-189761158	chr3:189564871-189682182	7183,3945	LPP	
CNVR_123	chr4:91083114-91244292	chr4:91202170-91309914	3443,4255	MMRN1	DGV
CNVR_173	chr5:110212742-110577198	chr5:110337010-110448345	850,4707	TSLP,WDR36	
CNVR_225	chr7:17795285-18873121	chr7:17674099-18707719	5541,5425	HDAC9,SNX13	
CNVR_336	chr12:17168087-17349294	chr12:17172155-18038013	1295,4372,3409		
CNVR_337	chr12:17761729-18154145	chr12:17172155-18038013	3409,4372	RERGL	
CNVR_415	chr16:82947011-83160475	chr16:82977842-83091199	3652,7245	ATP2C2,COTL1,KIAA1609	
CNVR_52	chr2:86130487-86470772	chr2:86237140-86417803	3789,7196	IMMT,MRPL35,POLR1A,PTCD3,REEP1	DGV
CNVR_64	chr2:188772738-189096411	chr2:188736466-189061332	5324,3107	GULP1	
CNVR_91	chr3:143326111-143612526	chr3:143303643-143554513	4710,6704	GK5,TFDP2,XRN1	DGV
CNVR_184	chr6:1994540-2304998	chr6:2001510-2293508	3815,3549	GMDS	
CNVR_185	chr6:7088878-7351773	chr6:7069194-7174697	7294,3511	CAGE1,RIOK1,RREB1,SSR1	
CNVR_188	chr6:29361859-29468501	chr6:29031959-29446150	3071,5367	OR12D3,OR14J1	
CNVR_243	chr7:88344723-88962788	chr7:88495346-89067351	1295,4008	ZNF804B	DGV
CNVR_260	chr8:2303622-2411266	chr8:2318174-2570606	1784,3768,7182,4804,3339		DGV
CNVR_264	chr8:13581606-13896999	chr8:13543770-13664796	5541,4227		DGV
CNVR_267	chr8:47038752-47556339	chr8:47516254-47655712	1784,7403		DGV
CNVR_297	chr9:137112069-137464520	chr9:137200474-137862491	323,4179,5758, 3593,6269	C9orf62,OLFM1	DGV
CNVR_298	chr9:137590463-137867305	chr9:137200474-137862491	3503,3593,6269,5758	CAMSAP1,GLT6D1,KCNT1,LCN9,PAEP,SOHLH1	DGV
CNVR_347	chr12:131115601-131701464	chr12:131210873-131630315	3071,3944	DDX51,EP400,EP400NL,GALNT9,MUC8,NOC4L	
CNVR_362	chr13:113166604-113658050	chr13:113543086-113654850	6638,3890	ATP4B,DCUN1D2,FAM70B,GAS6,GRK1,TFDP1,TMCO3	
CNVR_375	chr14:103592317-103986703	chr14:103860158-104009366	3812,5390,3802,5660,4106,6736	KIF26A	DGV
CNVR_398	chr16:2870891-3098404	chr16:2937256-3042293	3802,1334	CCDC64B,CLDN6,CLDN9,FLYWCH1,FLYWCH2,HCFC1R1,IL32,KREMEN2,MMP25,PAQR4,PKMYT1,THOC6,TNFRSF12A,ZSCAN10	
CNVR_450	chr19:7034753-7162312	chr19:6843781-7059064	6289,4281,3443	INSR,ZNF557	DGV
CNVR_469	chr20:29267569-29480559	chr20:29334284-29732674	3377,1559,3987	DEFB115,DEFB116,DEFB117,DEFB118,DEFB119,DEFB121	
CNVR_480	chr21:34576661-34852873	chr21:34631628-34827865	899,3546,3728,7171	C21orf51,KCNE1,KCNE2,RCAN1	

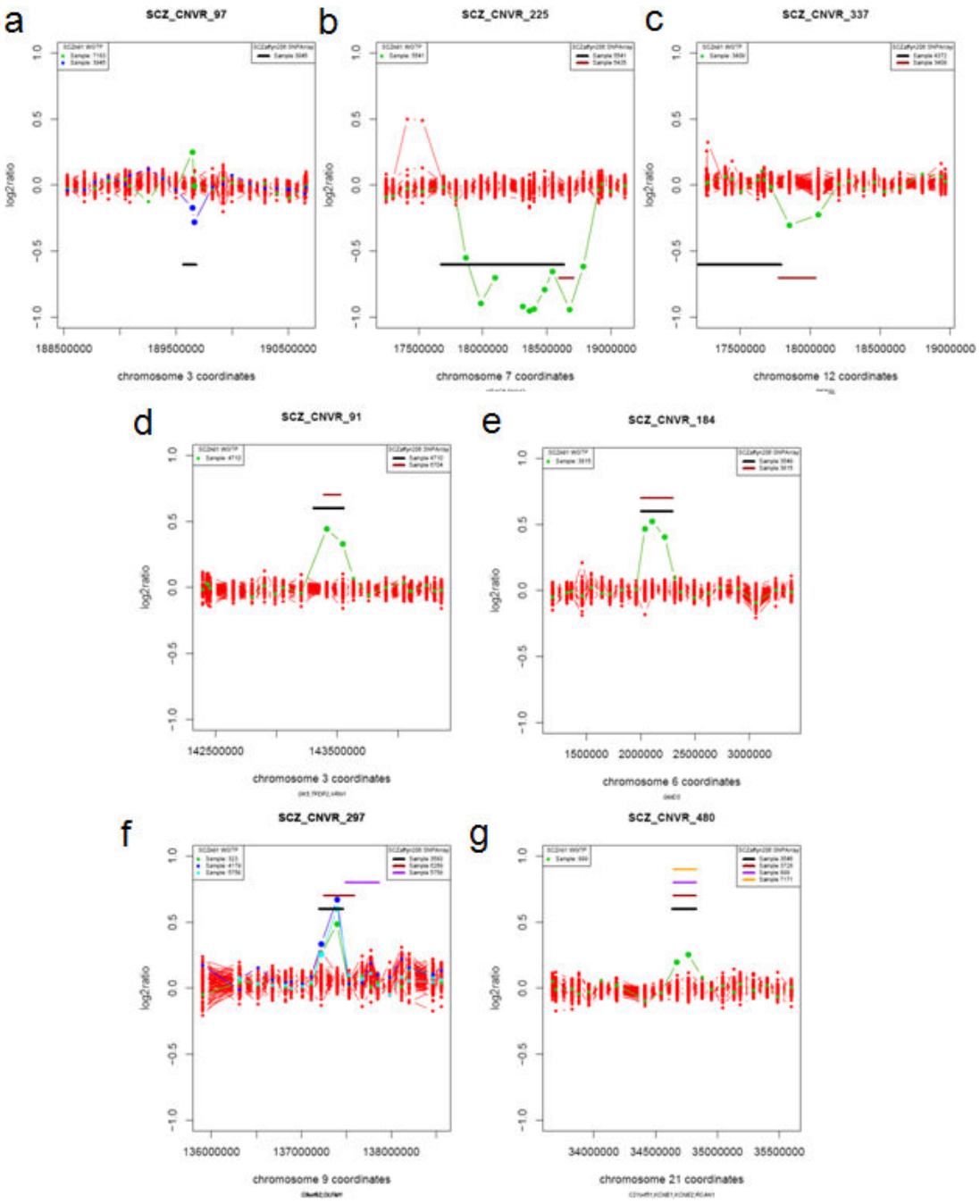


Figure 4.12 Recurrent SCZ-specific CNVR regions detected by both WGTP and Affymetrix platforms. The first 5 regions were identified in 2 SCZ samples: **a)** chr3; 189 Mb (*CNVR_97*) **b)** chr7; 176-187 Mb *(CNVR_225)* **c)** chr12; 171-181Mb ***(****CNVR_337)* **d)** chr3; 143 Mb (*CNVR_91*) **e)** chr6; 1.9-2.3 Mb *(WGTP_CNVR_184).* The final 2 regions occurred in 5 and 4 samples in the SCZ cohort respectively **g)** chr9; 137 Mb *(CNVR_297)*; **h)** chr21; 346Mb *(CNVR_480)*

4.5.1 A CNVR at Down Syndrome Critical Region 1 (RCAN1/DSCR1)

CNVR-480 is a duplicated locus affecting four patients (899, 3546, 3728 & 7171), one detected in SCZ_{WGTP}n91 and another three detected in SCZ_{Affy}n206 (Figure 4.12f). The 3 cases from the Affymetrix platform revealed the same breakpoints at chr9:137200474-137862491. The ~200 kb duplication region harbours two potassium channel genes *KCNE1* and *KCNE2*. Moreover, the gene Down Syndrome Critical Region 1 (*DSCR1*), also known as *RCAN1*, is located at the 3' breakpoint of the duplication according to Affymetrix SNP data. *RCAN1* has been linked to Down Syndrome and Alzheimer's disease (Hoeffer et al. 2007; Porta et al. 2007; Keating et al. 2008).

The gene *RCAN1* encodes a regulator of calcineurin, a calmodulin-dependent protein phosphatase. Both animal models and human studies have established the role of the calcineurin-dependent cascade in neuronal signal transduction and synaptic plasticity, as well as its involvement in psychosis (Eastwood et al. 2005). Genes that encode subunits of calcineurin, for example *PPP3CC*, have been demonstrated as associated with schizophrenia (Eastwood et al. 2005). Furthermore, *RCAN1* was recently shown to affect the expression of GSK-3beta (Ermak et al. 2006), another well-established schizophrenia and bipolar candidate gene (Koros and Dorner-Ciossek 2007). Knockout mouse models of *RCAN1* showed impairment in spatial learning and memory (in particular long-term potentiation), as well as sensorimotor deficits (Hoeffer et al. 2007).

4.5.2 A Variant Near Olfactomedin1 and other Recurrent CNVRs

CNVR-297 is another duplication region detected in three patients (5758, 323 & 4179) SCZ_{WGTP}n91, and in two additional samples (3593 & 6269) in SCZ_{Affy}n206. The duplication at 9q34 (Figure 4.12g) is downstream of the rat neuronal olfactomedin-related ER localized protein *OLFM1*, a gene abundantly expressed in the brain. Characterization of the CNV by oligonucleotide array revealed the same breakpoint in the patients (5758, 323 and 4179). We subsequently genotyped this variant in an extended cohort of 304 cases and 309 controls but failed to show significant association of the CNV with disease.

The other 5 recurrent regions were detected in 2 out of 297 samples. These regions harboured the following genes: LIM-domain containing preferred translocation partner in lipoma *LPP* (Figure 4.12a); histone deacetylase 9 (*HDAC9*), a histone deacetylase, which couples neuronal activities to muscle cells (Mejat et al. 2005); sorting nexin (*SNX13)*, a protein involves in G protein signalling and intracellular trafficking (Zheng et al. 2006) (Figure 4.12b); Ras-related and estrogen-regulated growth inhibitor-like protein *RERGL* (Figure 4.12c); Glycerol kinase 5 *GK5*, Transcription factor *DP-2* (TFDP2) and exoribonuclease *XRN1* (Figure 4.12d) and finally GDP-Mannose 4,6-Dehydratase *GMDS* (Figure 4.12e).

We evaluated the specificity of the recurrent CNVRs in the disease cohort by comparing them to the Database of Genomic Variants (DGV), a public resource cataloguing copy number variants and other structural rearrangements in normal individuals. 11 out of the 26 recurrent regions overlap (>50%) with more than one DGV CNV entries.

Whilst recognizing that some of these recurrent CNVRs overlap with DGV entries, such an overlap does not necessarily refute disease association for two reasons: 1)

Incomplete penetrance of genetic variants in schizophrenia suggests a disease-causing variant could be present in normal individuals although at a lower frequency. For example, known SCZ-associated loci, such as the 15q11.2 deletion, were shown to overlap with DGV entries. 2) Schizophrenia is complex psychiatric illness, occurring in 1% in population, with possibility of late disease onset. Since the apparently normal individuals reported with variants in DGV usually did not undergo rigorous neurological examinations, their disease status and background remain questionable.

In summary, in the extended cohort of $SCZ_{WGP\text{-}Affy}$n297, we detected 26 recurrent regions as putative disease variants. Candidate genes such as *RCAN1* were highlighted due to the relatively high frequency in cases and its biological relevance. Nevertheless, to establish disease association, a substantially larger sample size would be required.

4.6 Frequent Copy Number Variations in SCZ and LBC

4.6.1 Variance-based Clone-by-Clone Cohort Comparison

To investigate the role of frequent CNVs in disease, our first approach was to perform pair-wise comparisons on the SCZ versus LBC array CGH data for each autosomal BAC clones on the array, based on a statistic measure "Vst" (with reference to the F-statistic "Fst" in population genetics) (Tills 1977).

Vst is derived from the individual and combined variance of two sets of quantitative data (i.e. log2ratio in the two cohorts), adjusted by the respective number of observations (i.e. DNA samples), and is summarized as:

$$\mathrm{Vst} = \left\{ \mathrm{V_{total}} - \frac{(\mathrm{V_{P1}} \times \mathrm{N_{P1}} + \mathrm{V_{P2}} \times \mathrm{N_{P2}})}{(\mathrm{N_{P1}} + \mathrm{N_{P2}})} \right\} \div \mathrm{V_{total}}$$

V_{total} = Total variance of the two dataset

V_{P1} = Variance of the first population (e.g. Schizophrenia)

V_{P2} = Variance of the second population (e.g. LBC)

N_{P1} = Number of samples in the first population

N_{P2} = Number of samples in the second population

Vst scores, ranging from 0 to 1, were computed for each autosomal BAC clone spotted on the array. This variance-based approach has previously been used in the WGTP study on 269 HapMap samples (Redon et al. 2006) to examine population differentiation with respect to CNVs (Figure 4.13a). High Vst implies a considerable degree of differentiation among populations and vice versa. Regions previously reported with high

Vst among population, for example the copy number variant at *CCL3L1* associated with HIV-1 susceptibility, were detected with Vst scores up to 0.5 - 0.8 for the affected clones (Redon et al. 2006).

In contrast, Vst scores computed from schizophrenia versus LBC dataset suggests the case and control cohorts have minimal variation difference across the genome (Vst scores < 0.2) (Figure 4.13b). This remarkable similarity can be explained by the fact that schizophrenia and LBC are effectively the same ethnic population. It also implies that no single region of the genome, at the resolution of the WGTP array, shows substantial cohort-wide differentiation between SCZ versus LBC.

To conclusively exclude any Vst-based detection of copy number differentiation between the cases and controls, we examined the top hits of the Vst analysis. First all BAC clones were ranked according to Vst score and a Vst distribution for all clones was plotted. The top 1% clones, each with a Vst score above the threshold of 0.06853, were further filtered for regions with consecutive clones, based on the assumption that single clones with a technical artefact will have high probability of being in the tail of the Vst distribution and therefore should be discarded from the Vst analysis.

19 consecutive-clone regions were included in the top 1% Vst distribution. If these top Vst regions were true (i.e. demonstrating real population-wide differentiation between SCZ and LBC), we should be able to detect CNVs in samples with extreme log2ratio. It is also probable that any such CNV would be small in size compared to the size of a BAC clone, thus giving only a slight change of clone variance with relatively low Vst scores (compared to previous studies). We tested for the presence of cryptic CNVs in five of the top Vst regions using a high-resolution custom designed Nimblegen oligo array, hybridized using 10 available DNA samples from the SCZ cohort. Samples were

hybridized in pairs such that for each test Vst region, there was at least one DNA pair with high versus low log2ratio (Table 4.4).

4 of the 5 regions tested showed no signs of CNV. The Vst region at chr10 55-56Mb was detected with a ~50 kb CNV which overlaps the intron of gene *PCDH13* (protocadherin 13). However, occurrence of this CNV does not correlate with the log2ratio of the two clones at the Vst region, and it is concluded that the CNV was found by chance with no relation to Vst. In summary, none of the Vst regions were detected with CNVs that could be explained by the log2ratio.

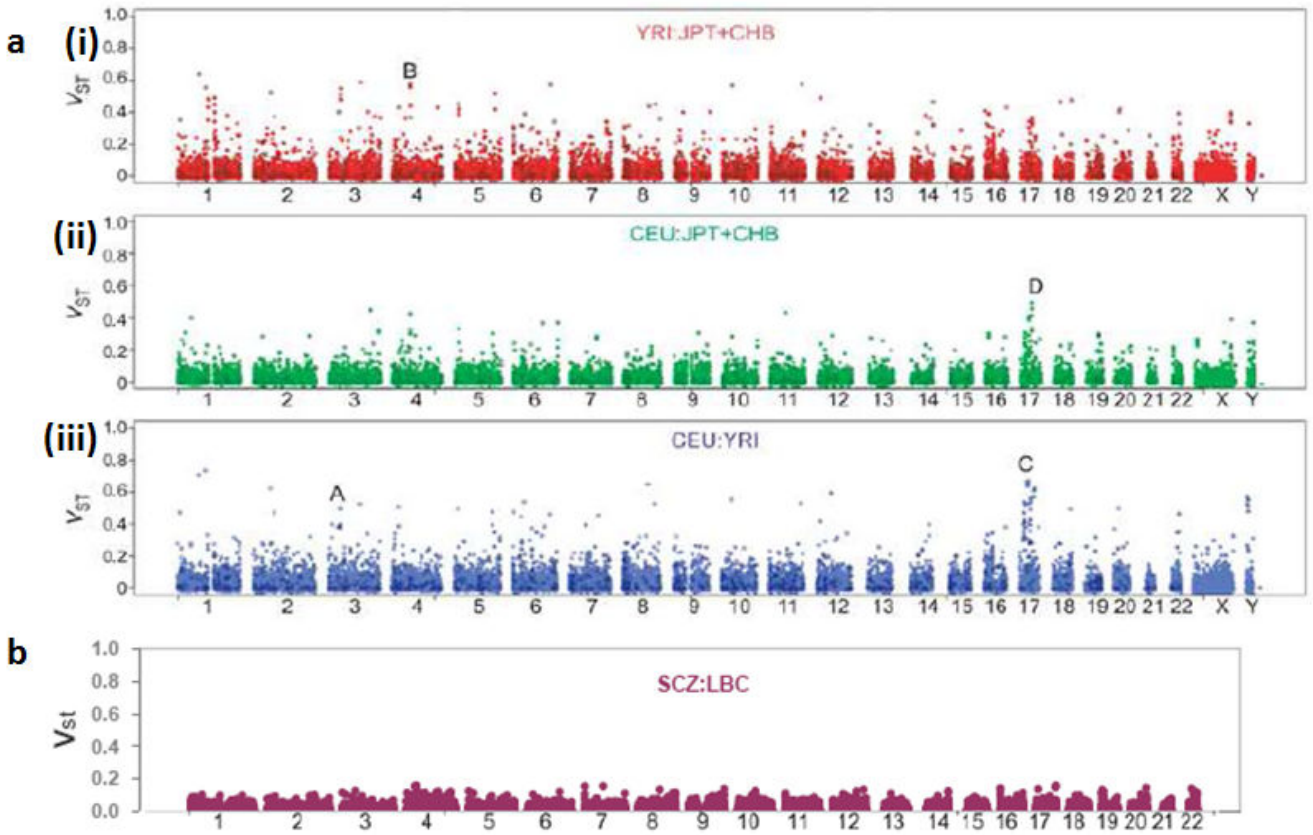


Figure 4.13 Vst scores to identify clones showing SCZ and LBC differentiation in the WGTP analysis. a) Vst scores from previously published HapMap WGTP CNV data (Diagram adopted from Redon *et. al.*). Comparisons were made between populations: **(i)** Yoruba (African) Versus Japanese/Chinese (Asian) **(ii)** Utah (European) Versus Japanese/Chinese (Asian) and **(iii)** Utah (European) Versus Yoruba (African). **b)** Vst scores derived from WGTP log2ratio in the current study, comparing schizophrenia versus control Scottish samples. Vst scores for all clones along the genome showed minimal differentiation (Vst below 0.2).

Table 4.4 Validating 5 top Vst-regions using custom-design oligonucleotide array. Each top Vst region had two consecutive clones with Vst score above the threshold of significance. For each Vst hybridization, a sample with high log2ratio (blue) was hybridized against another sample with low log2ratio (yellow). The final columns showed the minimum/maximum log2ratio of that clone for all samples in the cohort.

chr_corrdinates	signs of CNV?	clones	DNA Pair A		DNA Pair B		DNA Pair C		DNA Pair D		DNA Pair E		min	max
			4398	3341	3584	3789	241	4179	3652	5386	7180	5390		
chr5:84194412-84418390	no	Chr5tp-6E5	-0.0012	0.1146	0.0628	-0.0156	0.0373	0.0667	-0.0247	0.0582	-0.0254	0.0915	-0.0759	0.2033
		Chr5tp-16D5	0.0270	0.0897	0.0176	-0.0216	0.0368	-0.0068	-0.0342	0.0282	-0.0005	0.0684	-0.0754	0.1204
chr10:55963328-56236143	yes, but not in expected sample	Chr10tp-2H6	0.0099	0.0660	0.0935	0.0398	0.0428	0.0490	**-0.0192**	**0.0697**	-0.0430	0.1080	-0.0998	0.1793
		Chr10tp-5H2	-0.0439	-0.0304	-0.0470	-0.0677	0.0122	-0.0230	**0.0159**	**-0.0501**	0.0433	0.0156	-0.1770	0.1567
chr10:69928325-70219374	no	Chr10tp-2E7	0.0002	0.0746	-0.0495	-0.0745	-0.0238	0.0005	-0.0082	0.0435	0.0493	0.1126	-0.0977	0.1137
		Chr10tp-3C7	-0.0138	0.1468	-0.0396	-0.0414	0.0513	0.0695	0.0325	0.0483	-0.0109	0.0896	-0.1194	0.1468
chr16:229695-801844	no	Chr16tp-11B4	0.0946	0.0580	0.2023	0.0378	0.1668	-0.0066	0.0600	0.1321	-0.1565	0.1177	-0.1565	0.2023
		Chr16tp-12E7	0.0526	0.0905	0.1085	-0.0225	0.1995	0.0024	0.0611	0.1185	-0.1484	0.1167	-0.1484	0.1995
		Chr16-32k-3E2	0.1380	0.0333	0.1224	0.0032	0.1899	0.0910	0.0636	0.0352	-0.0798	0.1491	-0.1636	0.2275
		Chr16-32k-2H14	0.1126	0.0627	0.1709	-0.0218	0.2415	0.1336	0.0488	0.1390	-0.0882	0.2213	-0.1198	0.3709
chr22:48767142-49057642	no	Chr22-32k-1J4	0.0224	0.0197	0.1257	0.0024	0.0564	0.0529	-0.0118	0.0540	0.0200	0.0794	-0.0628	0.1558
		Chr22tp-8F6	0.0829	0.0871	0.1937	-0.0182	0.1411	0.0105	0.0158	0.0784	0.0070	NA	-0.1003	0.1937

4.6.2 CNV Genotyping with Bivariate Clustering

A second approach to detect frequency bias in common variants is to genotype CNV based on log2ratio clustering, followed by association tests on the resulting contingency tables of genotype distributions. For data clustering, we employed a model-based, expectation maximization hierarchal clustering algorithm known as Mclust (R package). The algorithm was performed on log2ratio of consecutive BAC clones with high correlation (Pearson correlation >0.5). This method is more effective for estimating genotypes at multi-allelic and complex regions than CNVFinder, which simplifies CNV status to a trichotomous classification of “deletion”, “normal” and “duplication” with respect to the reference DNA. Consistent with our previous analyses, we used a consecutive clone approach (bivariate clustering), decreasing the likelihood of false positives as a result of individual variability of any single clone.

Combined WGTP data on cases (n=91) and controls (n=92) were used for bivariate clustering. A total of 577 clones (out of 26,954 autosomal BAC clones on the WGTP array) were “genotypable” (i.e. resulting in more than one cluster when clustered with its downstream neighbour). The number of clusters was set at a range of 2 to a 9. Examples of bivariate clustering diagrams showing DNA samples with deletion (a), duplication (b) and a multi-allelic region (c) are shown in Figure 4.14.

We genotyped each schizophrenia and control sample for the 577 “genotypable” regions (except samples classified with high uncertainty which were removed in our algorithm). A chi-square test was then performed for the resulting genotype distribution, with the null hypothesis that SCZ and LBC were drawn from identical population.

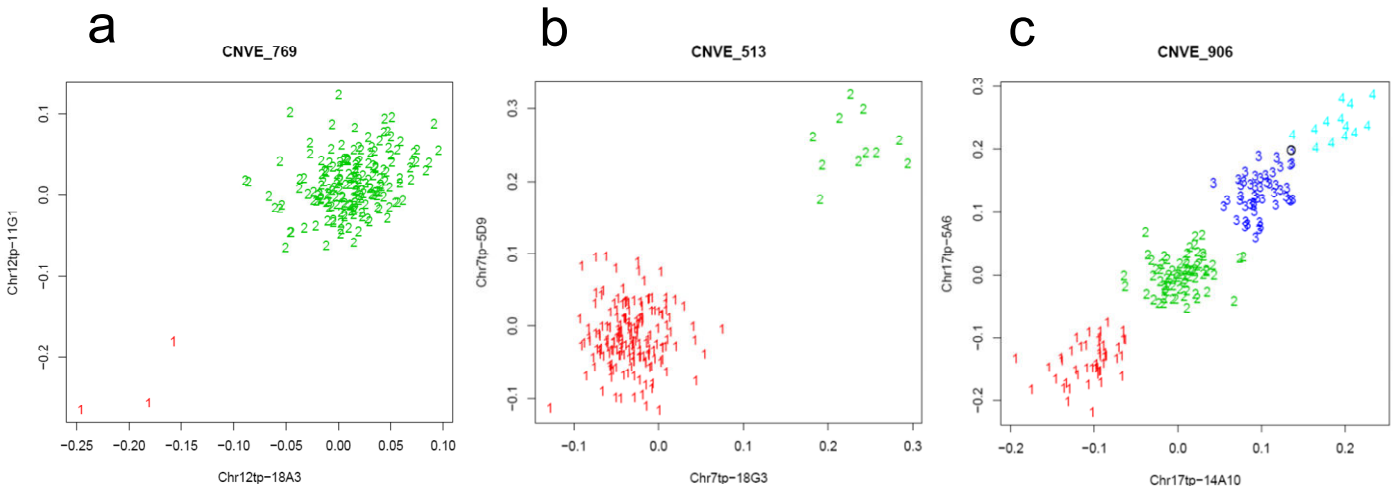


Figure 4.14 Examples of bivariate clustering based on log2ratio of consecutive clones. a) A region with majority of samples centred at log2ratio of 0.0 ("normal", cluster 2) with a minority of samples as deletion ("deletion", cluster 1); **b)** A region with a majority of samples centred at log2ratio of 0.0 ("normal", cluster 1) with a minority of samples as duplication ("duplication", cluster 2); **c)** A multi-allelic region with multiple clusters (cluster 1,2,3 and 4) with different copy numbers. *(x-axis: log2ratio of a particular BAC clone on the WGTP array; y-axis: log2ratio of an adjacent clone)*

31 of the 577 “genotypable” clones demonstrated significant difference in CNV genotype distribution (p-value < 0.05) following bivariate clustering (Appendix D-3). The clustering results for each of the 31 clones were manually inspected. 3 regions were selected for further investigation based on manual clustering and candidate gene involvement (Figure 4.15).

Two of the selected regions (Region 1 & 2) were genotyped in an extended cohort of cases and control DNA samples using PCR-based techniques. The two candidate regions overlap the gene neuronal adhesion molecule Close Homolog of L1 (*CHL1*) at 1p36 and the nicotinic receptor fusion gene *CHRFAM7A* (see Chapter 5) at 15q13 (see Chapter 5).

Region 3 was located near another cluster of highly variable segmental duplication on chromosome 17. The multi-allelic region partially overlapped N-ethylmaleimide-sensitive factor (*NSF*), a gene in the NMDA-receptor signalling complex which is important in membrane vesicle fusion, neurotransmitter release, trafficking of AMPA receptors and other major roles in synaptic plasticity (Haas 1998). Further experiments were performed to explore the genomic structure of the region (see Chapter 6).

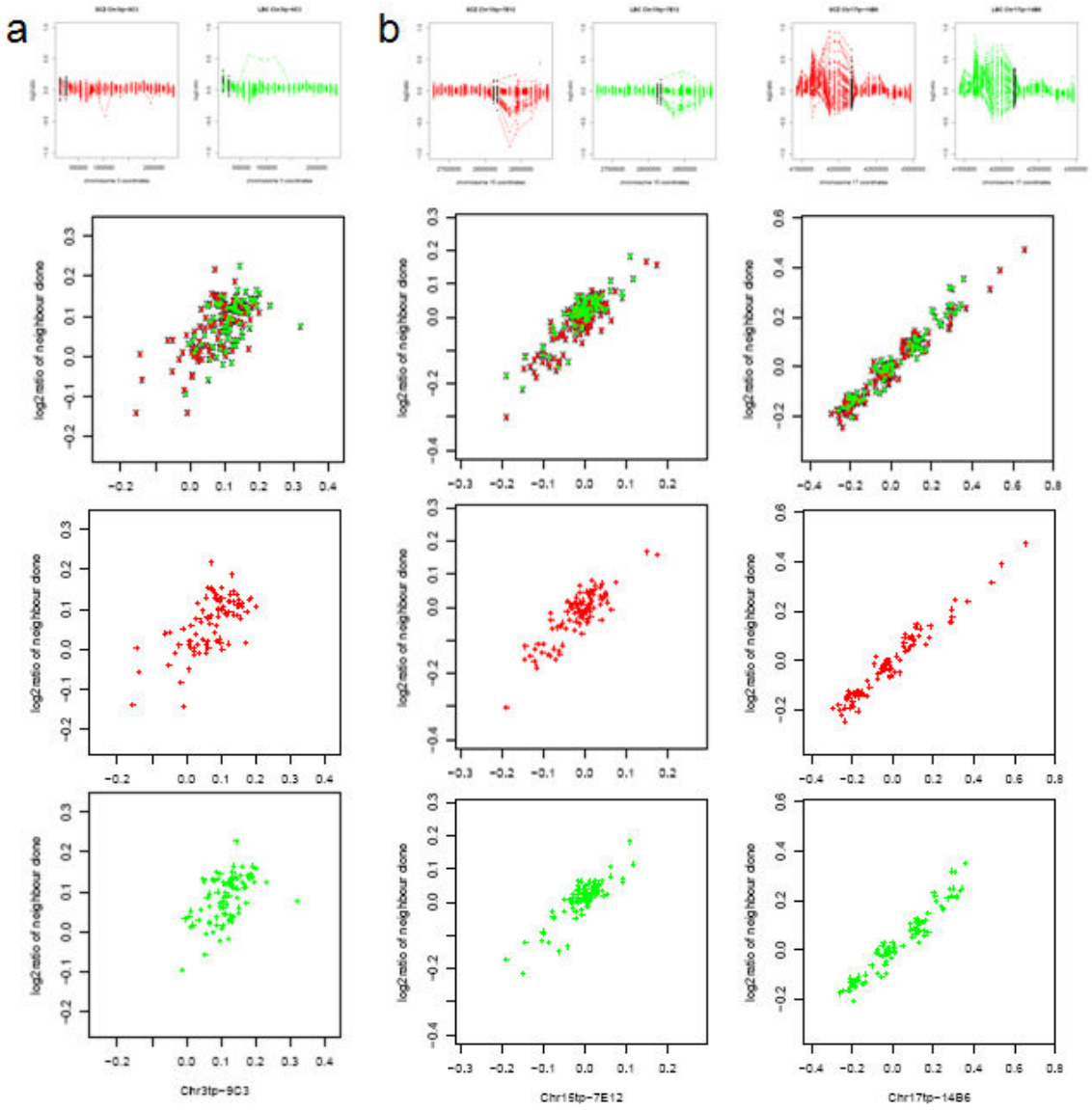


Figure 4.15 Three regions with significant difference of genotype distributions between SCZ and LBC as detected by bivariate clustering. **a)** 3p26 near *CHL1*: 3 CNV genotypes. **b)** 15q13 near *CHRFAM7A:* 2 main CNV genotypes. **c)** 17q21 near *NSF*: multi-allelic region with at least 5 CNV genotypes. *(Red: SCZ; Green: LBC*)

4.7 Chapter Summary and Discussion

In this chapter we described results from a population-based genome-wide CNV screen for schizophrenia patients and matched normal controls. The initial array CGH study was based on 91 patients and 92 controls, with hybridization performed on whole-genome tiling path BAC array. A subset of CNVs was validated by quantitative PCR and Nimblegen oligo-nucleotide array experiments. We also integrated CNV calls on a subset of 39 patients generated from Affymetrix SNP 6.0 array hybridizations, with data provided by the ISC and University of Edinburgh. The analysis was divided into two components: (a) Identification of rare, disease-causing variants in the case cohorts as putative disease candidates; & (b) Assessment of frequent copy number polymorphisms to investigate possible disease association.

A number of recent studies suggested that rare CNVs are responsible for a significant proportion of schizophrenia. In particular, Walsh *et al.* reported that in their case-control cohort, rare genic copy number variants existed as 15% of CNVs in cases but only as 5% in control (Walsh et al. 2008). If this were extrapolated to the general disease-control populations, as many as two thirds of the rare genic variants identified specifically in a case-control study should be implicated in schizophrenia pathogenesis. Accordingly, we focused our analysis of the WGTP data on such rare genic variants in the disease cohort.

We identified 113 rare, schizophrenia specific candidates not detected in the matched control set. We also replicated the finding at the 15q11.2 recurrent deletion region previously demonstrated as schizophrenia-associated in a large-scale study (Stefansson et al. 2008), demonstrating the ability of our approach to reveal potential schizophrenia-associated CNVs. However, to prove disease association of any of these variants, one would need a substantially larger case-control cohort.

Instead, we pursued a candidate CNV discovery approach, and aimed to catalogue SCZ-specific rare variants as a collective set of disease candidates. Studying such rare copy number mutations will lead to a better understanding of the genetics of schizophrenia: first, accumulating information on rare CNVs in SCZ patients will help in discriminating true recurrent disease variants from benign mutations; second, the rare disease variants in many different patients or families may converge in biological pathways, providing insight into the disease phenotype.

An important approach to assess clinical relevance of this set of rare CNVs is to examine the loci with respect to gene content (Lee et al. 2007). For instance, rare case-specific variants with the presence of morbid OMIM gene(s) with brain-related function, or gene(s) previously linked to a neuropsychiatric disorder, would be considered plausible schizophrenia candidates responsible for disease phenotypes. Among the disease-specific variants we detected a number of such candidate genes, including *PIK3C3*, *SGCE, OXTR* and *RCAN1*, all providing functional relevance to psychiatric disorders.

In the second half of the chapter, we described the genotyping of common copy number polymorphisms by a hierarchical model-based clustering algorithm. Genotype distributions in cases and controls were compared using a nonparametric test with the null hypothesis that patients and controls were drawn from the same distribution of copy numbers. A total of 31 CNV regions were identified with a distribution significantly different in cases versus control (chi-square test, p-value <0.05). By applying this case-control association approach to our CNV study, we face the limitation of a) the adequacy of our sample size to confer enough statistical power; and b) the accuracy of converting CNV measurements (as log2 fluorescent ratio) to CNV genotypes. We therefore subsequently genotyped two of these candidate regions, using more precise PCR-based methods for copy number

detection, and in the extended case-control cohorts of ~ 300 + 300 samples (discussed in Chapter 5).

Regarding experimental design, of particular importance to a population-based case-control CNV screen was the appropriate choice of controls (McCarroll and Altshuler 2007). A well-matched control cohort is crucial for a number of reasons: First, a number of studies showed that private rare CNVs associated with specific continental ancestry were common (Khaja et al. 2006; Locke et al. 2006; Redon et al. 2006), perhaps more so than for private SNPs (Jakobsson et al. 2008). To detect rare variants as disease candidates we employed an ethnically well-matched Scottish control cohort for our Scottish patient cohort.

Similarly, well-matched case and control ethnicity is critical for tests of association. The frequency of CNV of interest may differ across populations. Other genetic variants which could alter liability to disease (e.g. the presence of inversion predispose to CNV) may also be population-specific. Moreover, in some cases of CNV disease association such as *CCL3L1* in HIV-1 progression, the copy number *per se* was not a main disease determination factor, rather the gene dose relative to the population average was critical (Gonzalez et al. 2005). All of these again argue for the significance of a well-matched control cohort.

Other practical issues related to case-control study include the differential quality, source and experimental treatment of DNA samples, which could influence CNV data quality in cases versus control. As described in this chapter, our study did face a limitation regarding sample quality. This translates into a bias of CNV detection sensitivity in control versus cases, and has complicated statistical disease association testing, including the calculation of overall disease burden of CNVs.

Finally, we would emphasize on using a control cohort that is free from psychiatric disease for proper schizophrenia CNV study. In this work, the Lothian Birth Control individuals (Deary et al. 2007) have been thoroughly examined for neuropsychiatric functions, with neurological tests performed at the age of ~90 (far beyond the normal age of onset of schizophrenia), and therefore should be free from psychiatric illness. Nevertheless, since schizophrenia is complex and genetic variants could have incomplete penetrance, disease-related variants may still occur in the control cohort.

Despite a few inevitable constraints, our data generated a number of validated, potentially disease-associated rare variants in the schizophrenia cohort. A number of these rare variants harbour psychiatric-disease related genes. Further bioinformatic work to explore these rare variants using a functional biological pathway approach (e.g. the Gene Ontology pathway) could provide more insight into the disease pathways.