Copy Number Variation in Bipolar Disorder

A thesis submitted for the Degree of Doctor of Philosophy at
Cardiff University
By

Detelina Grozeva

Supervisors: Dr George Kirov
Dr Ian Jones

MRC Centre for Neuropsychiatric Genetics and Genomics
Department of Psychological Medicine and Neurology
School of Medicine
Cardiff University
2010

Copy Number Variation in Bipolar Disorder

A thesis submitted for the Degree of Doctor of Philosophy at
Cardiff University
By

Detelina Grozeva

Supervisors: Dr George Kirov
Dr Ian Jones

MRC Centre for Neuropsychiatric Genetics and Genomics
Department of Psychological Medicine and Neurology
School of Medicine
Cardiff University
2010

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

Acknowledgements

George, I am very grateful as you taught me a lot of invaluable lessons about science and also for being very understanding, always providing support and having time for me even when I developed a habit of just popping in your office.
Ian, thanks for always being very kind and nice; and thanks for the excellent encouragement and for teaching me to think critically when writing, hope that I achieved this in practice.
Nick, even though not my supervisor, gave me the opportunity to work and do a PhD in Cardiff, for which I am ever so grateful.

I wish to also thank all the members of the Bipolar team in Cardiff and Birmingham: John, Ellie, Liz, Katherine, Sian, Lisa, Sally, Christine, Elaine, Charlotte, Bev, Beck and Cat, whose dedication to work has been going very strong through the years. In addition, the participants in our research, even though I know them only by the allocated research ID, are all remarkable people.
Thanks to Nick Bray, Ivan, Varuni, Masashi and Tom for being such nice people.
Thanks to Hywel, Rhodri and Liam for being so lovely and funny, and for making me laugh!
Thanks to Valentina for our regular badminton stress-abolishing sessions.
Many thanks to Elisa, Amy, Sarah, Lucy, Debbie, Evie, Irina, Beccy, Rachel, Denise and Jade for always being so amicable and helping with advice/chocolate/cake or even just with listening through the many years spent together; I feel that you are my extended family. I have to mention Sarah one more time, as I feel that we have been in this together-it was easier as I felt that I am not alone and no, I have not used the words that you have taught me in the PhD.
I am also blessed with having good friends with whom I remained "together" despite the distance (ranging from 140 miles to >9,000 miles)- Gaby, Lynn, Sue, Savina, Joro, Mitko, Niki and Dobri!

Special thanks which I don't think that am able to express with words goes to my parents and grandparents-thank for your unconditional love. Mom and Dad, I appreciate your sacrifices through the years (I know that some of these years have been unbelievably tough for you); thank you for everything that you have done for me!

And Dobril, I would not have been able to do anything without you, thank you for always being by my side and for your absolute, unlimited and categorical help!

Associated publications

1. **Grozeva D**, Kirov G, Ivanov D, Jones IR, Jones L, Green EK, St Clair DM, Young AH, Ferrier N, Farmer AE, McGuffin P, Holmans PA, Owen MJ, O'Donovan MC, Craddock N; Wellcome Trust Case Control Consortium. 2010. Rare copy number variants: a point of rarity in genetic risk for bipolar disorder and schizophrenia. *Arch Gen Psychiatry* 67(4), pp. 318-27;

2. Kirov G, **Grozeva D**, Norton N, Ivanov D, Mantripragada KK, Holmans P; International Schizophrenia Consortium; Wellcome Trust Case Control Consortium, Craddock N, Owen MJ, O'Donovan MC. 2009. Support for the involvement of large copy number variants in the pathogenesis of schizophrenia. *Hum Mol Genet* 18(8), pp. 1497-503;

3. McCarthy SE, Makarov V, Kirov G, Addington AM, McClellan J, Yoon S, Perkins DO, Dickel DE, Kusenda M, Krastoshevsky O, Krause V, Kumar RA, **Grozeva D**, Malhotra D, Walsh T, Zackai EH, Kaplan P, Ganesh J, Krantz ID, Spinner NB, Roccanova P, Bhandari A, Pavon K, Lakshmi B, Leotta A, Kendall J, Lee YH, Vacic V, Gary S, Iakoucheva LM, Crow TJ, Christian SL, Lieberman JA, Stroup TS, Lehtimäki T, Puura K, Haldeman-Englert C, Pearl J, Goodell M, Willour VL, Derosse P, Steele J, Kassem L, Wolff J, Chitkara N, McMahon FJ, Malhotra AK, Potash JB, Schulze TG, Nöthen MM, Cichon S, Rietschel M, Leibenluft E, Kustanovich V, Lajonchere CM, Sutcliffe JS, Skuse D, Gill M, Gallagher L, Mendell NR; Wellcome Trust Case Control Consortium, Craddock N, Owen MJ, O'Donovan MC, Shaikh TH, Susser E, Delisi LE, Sullivan PF, Deutsch CK, Rapoport J, Levy DL, King MC, Sebat J. 2009. Microduplications of 16p11.2 are associated with schizophrenia. *Nat Genet* 41(11), pp. 1223-7.

Table of contents

List of figures

List of tables

Summary

A plethora of studies have suggested that copy number variation in the human genome is extensive and may play an important role in susceptibility to disease, including neuropsychiatric disorders such as schizophrenia and autism. The possible involvement of copy number variants (CNVs) in bipolar disorder has received little attention to date.

This PhD thesis describes work that sought to determine whether large (≥ 100 kb) and rare (with frequency $\leq 1\%$) CNVs are associated with susceptibility to bipolar disorder and to make comparisons with previous findings in schizophrenia.

In order to do that, a genome-wide survey of large and rare CNVs in a case-control sample using a high-density microarray was performed. 1697 cases of bipolar disorder and 2806 non-psychiatric controls were genotyped using Affymetrix 500K array set. Subsequently, the copy number data were inferred and analysed.

The burden of CNVs in bipolar disorder was not increased compared with controls. Furthermore, CNVs > 1 Mb were found with statistically significant lower rate in bipolar disorder as compared to schizophrenia. In addition, CNV loci previously implicated in the aetiology of schizophrenia were not more common in cases with bipolar disorder.

The main observation was that large and rare CNVs do not have a major effect in the susceptibility to bipolar disorder. In addition, it was noted that bipolar disorder and schizophrenia differ with respect to CNV burden in general and specific CNVs in particular. As the same CNVs implicated in schizophrenia, have been observed in autism and mental retardation, and not in bipolar disorder, it is reasonable to postulate that CNVs could be a specific genetic factor for a predisposition to disorders with stronger neurodevelopmental component than bipolar disorder.

This study has provided one of the first glimpses of the possible involvement of copy number variation in the susceptibility to bipolar disorder.

General structure of the PhD thesis

This PhD thesis comprises four main chapters. The first one will provide a general review of bipolar disorder and the putative genetic factors involved in the susceptibility. I will also provide evidence of the involvement of copy number variation in the predisposition to neuropsychiatric disorders. Then, I will list the aims and the objectives of the PhD project. Chapter three, the materials and methods section, provides a detailed description of the cases and the controls analysed in the thesis and the laboratory, statistical and bioinformatical methods applied in the copy number variation analyses. Then, I will move on to present the results from the performed analyses of copy number variation in bipolar disorder. In the last chapter (i.e. the discussion) I will discuss the main findings and will put them into perspective of what is known in the field of neuropsychiatric genetics in general with respect to copy number variation. Finally, given the new work reported here, I will bring to a close with general conclusions and possible future avenues for research.

1. Introduction

In the introduction section of the PhD I will summarise what is known about bipolar disorder in terms of classification, diagnostic criteria, mode of inheritance and aetiology. Then I will move on to discuss why it is important to study genetic factors and what is known with respect to the involvement of such genetic variants in the susceptibility to the disorder. I will then move on to consider the contribution of a relatively recently discovered type of variation in the human genome, namely copy number variation. I will briefly review the mechanisms for copy number variation occurrence, how CNVs could influence the phenotype and the current methods for their detection. Given the existing evidence for the involvement in susceptibility to conditions such as mental retardation, autism and schizophrenia, I will consider the potential involvement of copy number variation in liability to bipolar disorder.

1.1. Bipolar disorder

Bipolar disorder causes unusual intense emotional states with extreme changes in mood, energy levels, activity and the ability to carry out day-to-day tasks. The symptoms are severe and cannot be compared with the normal ups and downs that everyone experiences once in a while.

Bipolar disorder is not a disorder that has been recognized recently. The ancient Greeks and Romans were the first to describe conditions related to mood. At that time the terms melancholia and mania were coined.

Aretaeus of Cappadocia (ca.150) is credited to be the first who made the connection between melancholia and mania (Akiskal 2009). His description of mania closely corresponds to the contemporary one:

> *"There are infinite forms of mania, but the disease is one of them. If mania is associated with joy, the patient may laugh, play, dance all night and day, and go to the market crowned as if a victor in some contest of skill. The ideas the patients have are infinite. They believe they are experts in astronomy, philosophy, or poetry."*

Nowadays it is recognized that bipolar disorder is a chronic disease with periods of remission and subsequent relapse. It is a disabling condition, which in tragic cases can lead to suicide.

Bipolar disorder is a condition which along with unipolar depression is referred to with the umbrella term of "Mood disorder" or "Major affective disorder". The subdivision to bipolar and unipolar disorder is based on the course of illness. Major affective disorder may have manic-depressive (bipolar) or exclusively depressive course of illness (American Psychiatric Association 1994).

Bipolar affective disorder is characterised by episodes of mania (bipolar I disorder) or hypomania (bipolar II disorder) interspersed with episodes of depression (American Psychiatric Association 1994). According to the World Health Organisation (WHO), 29.5 million people were affected by bipolar affective disorder in 2004. In addition it was estimated that bipolar disorder is among the top 20 leading causes of disability, with around seven years of healthy life lost through time spent in states of less than full health (World Health Organisation 2008). Moreover, now it is thought that mood disorders are among the most frequent psychiatric illnesses in community (Rihmer and Angst 2009).

These figures and the severity of the disease warrant further investigation of the symptoms, causes, effective treatment and prevention of bipolar disorder.

Emil Kraepelin established the manic-depressive illness as a nosological entity and differentiated it from *dementia praecox* (which came to be known as schizophrenia). Kraepelin's work was based on extensive clinical observation (i.e. longitudinal study of 1000 cases); he described the episodic course and the different family history, which separate manic-depressive "insanity" from *dementia praecox*. Kraepelin noted that the melancholic and manic states are recurrent, whereas *dementia praecox* leads to degeneration of cognition and the personality of the affected individuals (reviewed in (Kendler 1986)). This delineation is still prevalent nowadays. However, observations of overlapping boundaries between bipolar disorder and schizophrenia put this division to the test.

Kraepelin did not distinguish between what nowadays is termed as unipolar disorder and manic-depressive illness. In his work, people suffering from depression were still classified as manic-depressives (reviewed in (Kendler 1986)).

After Kraepelin's work, the classification of manic-depression was still evolving. Based on clinical observation and the different family history data in patients with depression, Leonhard proposed to distinguish between bipolar and unipolar disorders (Goodwin and Jamison 2007). The face validity of the separation of unipolar disorder from bipolar disorder was later supported by independent research carried out by Angst, Perris and others (reviewed in (McGuffin and Katz 1986)).

In contrast to bipolar disorder, unipolar disorder is markedly more common in terms of frequency of occurrence. The lifetime prevalence estimates substantially vary from 4.4% to 17.1% according to the applied diagnostic criteria, methodology and sample (Jones et al. 2002). In addition, it has been observed that the two disorders differ with respect to greater familial burden in bipolar disorder as compared to unipolar disorder (Kelsoe 2009). The separation between the two types of affective disorders was later implemented in the World Health Organisation International Classification of Diseases and in the American Psychiatric Association Diagnostic and Statistical Manual of Mental disorders (American Psychiatric Association 1994; World Health Organisation 1992).

1.1.1. Classification and diagnostic criteria

Bipolar disorder cannot be detected with a laboratory test, which makes diagnosing challenging. Great efforts have been undertaken in the psychiatric field to establish diagnostic criteria. One of the most important tasks was to introduce diagnostic criteria which are "reliable", i.e. two independent investigators will agree if a certain diagnostic label applies to a certain patient (Strachan and Read 2004). To standardise the diagnosis, classification systems like the Diagnostic and Statistical Manual of Mental Disorders, 4th edition (DSM-IV) (American Psychiatric Association 1994) and International Statistical Classification of Diseases and Related Health Problems, 10th revision (ICD-10) have been introduced (World Health Organisation 1992).

The DSM-IV criteria for manic and major depressive episodes are summarised in Table 1.

Table 1 DSM-IV diagnostic criteria for major depressive episode and manic episode

Major depressive episode	Manic episode
A. Five or more of the following symptoms present during the same two week period and represent a change from normal functioning 1. Depressed mood 2. Loss of interest or pleasure 3. Change in appetite and weight 4. Sleep change 5. Psychomotor agitation or retardation 6. Loss of energy 7. Feeling of worthlessness or excessive or inappropriate guilt 8. Diminished concentration 9. Recurrent thoughts of death/suicidal ideation B. The symptoms do not meet the criteria for a mixed episode C. The symptoms cause clinically significant distress or impairment in social, occupational or other important areas of functioning D. The symptoms are not due to direct physiological effects of a substance or a general medical condition E. The symptoms are not better accounted for by bereavement	A. Abnormally and persistently elevated, expansive or irritable mood lasting at least one week (if hospitalised- any duration) B. Three or more of the following symptoms (four if mood is only irritable) present to a significant degree 1. Increased self-esteem or grandiosity 2. Decreased need for sleep 3. More talkative than usual or pressure to keep talking 4. Flight of ideas or racing thoughts 5. Distractibility 6. Increase in goal oriented activity or psychomotor agitation 7. Excessive involvement in pleasurable activities that have a high potential for painful consequences C. The symptoms do not meet the criteria for a mixed episode D. The mood disturbance is sufficiently severe to cause impairment in occupational functioning or in usual social activities or relationship with others or to necessitate hospitalisation, or there are psychotic features E. The symptoms are not due to direct physiological effects of a substance or a general medical condition **Hypomanic episode:** *milder* form of a manic episode with less duration. A distinct period of persistently elevated, expansive or irritable mood, lasting at least four days. Like a manic episode, pathological change in functioning is observed, but the episode is not severe enough to cause marked impairment in functioning or to necessitate hospitalisation and there are no psychotic features
Mixed episode The criteria are met for both manic and for a depressive episode except the duration. It is still sufficiently severe to cause significant impairment in functioning. An example could be the following: *May feel sad and hopeless while feeling energised*	

The presence of all of A, B, C, D and E is required to diagnose illness. Modified from Psychiatric Genetics & Genomics (Jones et al. 2002)

According to DSM-IV, bipolar disorder spectrum is further subclassified into bipolar I disorder (BDI), bipolar II disorder (BDII), cyclothymia and bipolar disorder not otherwise specified (BD NOS) (American Psychiatric Association 1994).

BDI is diagnosed if one or more manic episodes or mixed episodes are present while BDII is diagnosed if the individual experiences milder forms of mania (termed hypomania) (Jones et al. 2002). Cyclothymia is a mild form of bipolar disorder, characterised with episodes of hypomania that interchange with mild depression for at least two years. When the symptoms of the illness do not meet the criteria for BDI and BDII (e.g. the symptoms do not last long enough or not all the required symptoms are present), BD NOS is diagnosed. Even though the BDII, cyclothymia and BD NOS are characterised with somewhat milder presentation of the disease, they are still out of the normal range of behaviour.

Due to the fact that the current diagnostic criteria are not based on evidence coming from the biology of psychiatric diseases, the most appropriate boundaries between the bipolar spectrum subphenotypes and the other psychiatric illnesses are still unclear. At the severe psychotic end of the spectrum, some patients exhibit features which do not belong strictly to bipolar I disorder or schizophrenia but belong to both of the disorders (Jones et al. 2002). More about the overlap between bipolar disorder and schizophrenia will be discussed further in section 1.1.6 (page 27).

1.1.2. Epidemiology

The lifetime rates of developing bipolar disorder are comparable between the different human populations and range between 0.8 and 1% (American Psychiatric Association 1994; Weissman et al. 1996). This figure has been challenged and there are now convincing data that bipolar disorder may affect ~5% of the population. This is largely due to the better detection of bipolar II disorder. Results from the National Comorbidity Survey Replication Study showed a lifetime prevalence of 1% for Bipolar I disorder, 1.1 % for bipolar II disorder and 2.5% for clinically manifested subtreshold bipolar disorder (Rihmer and Angst 2009). The lifetime prevalence rates for the separate disorders comprising the bipolar spectrum disorders are presented in Table 2.

Table 2 Lifetime prevalence rates of bipolar I disorder, bipolar II disorder, cyclothymia and hypomania

	Lifetime Prevalence [%]
Bipolar I disorder	0.0-2.4
Bipolar II disorder	0.3-4.8
Cyclothymia	0.5-6.3
Hypomania[a]	2.6-7.8
Full bipolar spectrum	2.6-7.8

[a] Including recurrent brief hypomania (lasting one to three days, not fulfilling the DSM-IV criteria for hypomania) (adapted from (Rihmer and Angst 2009))

When the gender distribution is studied across cases, it has been observed that both genders are equally distributed (Weissman et al. 1996). The most frequent age of onset of bipolar disorder is ~ 20 years of age. Earlier onset in men has been noted, four to five years earlier than women (Rihmer and Angst 2009). It has also been observed that individuals may develop the first episode of bipolar disorder at any time of their life (Jones et al. 2002). When rates of bipolar I disorder have been compared in Caucasians, African American and Hispanic individuals, it has been observed that the prevalence is not different across the different race groups (Rihmer and Angst 2009). It has been suggested that some seasonality is associated with the illness, with depressive episodes found to be more frequent in spring and autumn, whereas mania has been found more frequently in summer (Goodwin and Jamison 2007).

It has been noted that bipolar disorder is most frequent among divorced, separated or widowed individuals and that the rate of family breakdown is increased. Bipolar I disorder can lead to unemployment and low income resulting in regression in the socioeconomic status. On the other hand, a lack of significant downward social drift has also been noted. This has been attributed to the natural history of the illness, i.e. frequently the onset is after university, the prognosis of the illness is good and in addition, individuals often have compensatory energy, ambition and cognitive changes (Goodwin and Jamison 2007). For example, bipolar II individuals tend to belong to higher social classes, are educated above average and are relatively overrepresented among creative people (Rihmer and Angst 2009).

1.1.3. Aetiology

1.1.3.1. Genetic factors

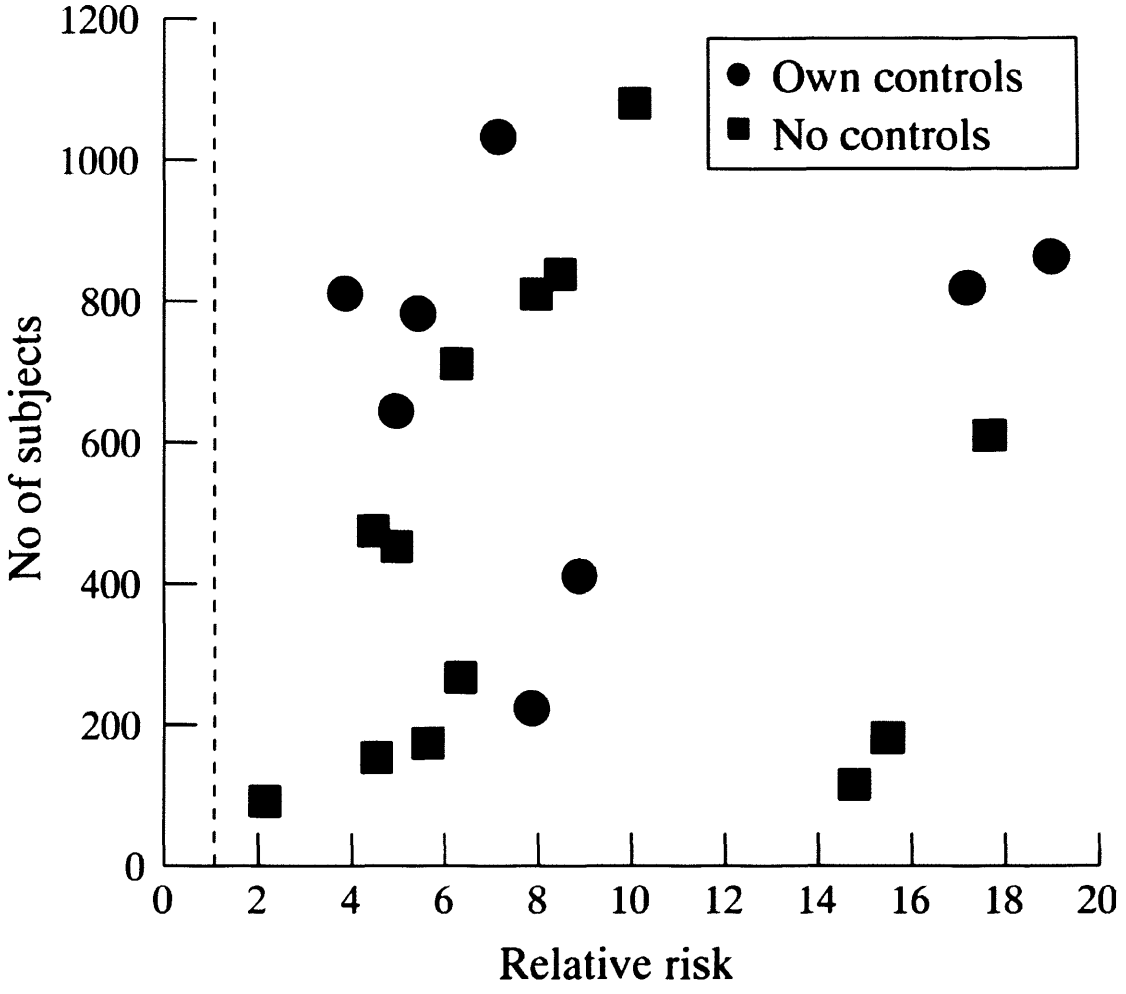
To determine if a non-Mendelian disorder like bipolar disorder is caused by genetic factors, family, twin and adoption studies have been performed. Such studies, conducted in the last 30-40 years provided consistent body of evidence supporting the hypothesis that bipolar disorder is indeed a genetic disorder (Craddock and Sklar 2009; Kelsoe 2009).

1.1.3.1.1. Family studies

Family studies have clearly shown that there is a reduction of the risk to develop bipolar disorder when the genetic relatedness to the affected proband diminishes (Craddock and Jones 1999). This has been shown by measuring a genetic parameter called relative risk, which is generally used to answer the question if a given illness runs in families. The relative risk is defined as the ratio of the rates of developing bipolar disorder in 1st degree relatives of bipolar disorder probands compared to the rates of illness observed in the 1st degree relatives of healthy controls, or where control subjects are not available, to an assumed risk in the general population of 1% (Kelsoe 2009). Figure 1 shows the relative risk of bipolar I disorder in 1st degree relatives of bipolar disorder cases as a function of the number of analysed samples in the relevant data that the graph has been based on. The graphical summary has been based on family studies published up to 1999, which have met the following criteria: the probands suffer from bipolar I disorder, the lifetime risk of bipolar disorder has been measured in 1st degree relatives and at least some of the relatives have been interviewed directly. The data for producing the graph come from 21 studies, eight of which included control sample set. All of the studies have observed a relative risk of > 1, which reflects an increased risk of developing bipolar disorder in relatives of a bipolar proband compared with the risk in the general population (Craddock and Jones 1999). Based on the eight studies that included their own control subjects, an overall relative risk has been measured in 1st degree relatives

of bipolar I cases- seven, indicating strong familial risk of developing bipolar disorder (reviewed in (Jones et al. 2002)).

Figure 1 Family studies of bipolar disorder



(Craddock and Jones 1999)

A recent study based on a large sample set of bipolar disorder affected individuals (n=40,487) along with their families has provided support of the hypothesis that 1st degree relatives of individuals with bipolar disorder exhibit an elevated risk for developing the condition than the populational risk. Having a mother or a father with bipolar disorder inreases the risk of an offspring for developing the illness 6.4 times ($95\%CI$: 5.9 – 7.1), compared with a person whose mother or father do not have bipolar disorder (Lichtenstein et al. 2009).

In conclusion, the family studies support the familial nature of bipolar disorder. It is of note, that not only bipolar disorder is observed in the families of bipolar probands. Other psychiatric illnesses, namely unipolar depression, bipolar II

disorder, schizoaffective disorder and schizophrenia, have been also detected suggesting some degree of common genetic underpinnings between the separate disorders (Jones et al. 2002; Kelsoe 2009; Lichtenstein et al. 2009).

1.1.3.1.2. Twin and adoption studies

Data from family studies have indicated that bipolar disorder tends to run in families. As individuals in families share both genetic and environmental factors, family studies are unable to show whether a disease is due to inheritance of specific genetic factors or to shared environment. To this end, twin and adoption studies provide a relatively more powerful approach to separate genetic from environmental factors, than family studies (Hayden and Nurnberger 2006; Kelsoe 2009).

The most common way of performing twin studies is to identify monozygotic (MZ) and same-sex dizygotic (DZ) twin pairs where one of the twins has bipolar disorder. Subsequently, the other twin from the pair is examined to determine the proportion of twin pairs both affected with bipolar disorder (concordance rate) (Kelsoe 2009).

The basic principle of the twin studies is that monozygotic (identical) and dizygotic (fraternal) twins share common environment, whereas only the monozygotic pairs share ~100% genetic factors (DZ twins share on average 50% common genetic factors). Any greater similarity in MZ as compared to DZ twins should be attributable to heritable genetic factors (McGuffin et al. 1994).

Table 3 (page 12) summarises several twin studies (reviewed in (Jones et al. 2002)).

Table 3 Lifetime rates of affective disorder in co-twins of bipolar twin probands

Reference	Sample			Lifetime rates of illness in co-twin of bipolar twin probands (proband-wise concordance rate) [%]			Comment
	Source	N MZ pairs	N DZ pairs		MZ	DZ	
(Kringlen 1967)	Norway twin and psychosis register	6	0	BD-BD	67		~ small sample size
(Allen et al. 1974)	USA Veteran twin register (15,909 twin pairs in register)	5	15	BD-BD	20	0	~ low rate of BD detected in the twin sample-0.07%
(Bertelsen et al. 1977)	Denmark twin and psychiatry registers	34	37	BD-BD BD-BD/UP	62 79	8 19	
(Torgersen 1986)	Norway twin register	4	37	BD-BD BD-BD/UZ	75 100	0 0	~ small sample which may overlap partially with that of Kringlen
(Kendler et al. 1993)	Sweden twin and psychiatric registers	13	22	BD-BD BD-BD/UZ	39 62	5 14	~ large sample, used questionnaire assessment, likely to have underestimated concordance
(Cardno et al. 1999)	UK psychiatric hospital twin register	22	27	BD-BD	36	7	Diagnoses based on hospital notes, likely to have underestimated concordance

Based on Jones et al. (Jones et al. 2002); BD-BD- refers to twin pairs in which both have BDI; BD-BD/UP- refers to twin pairs where one twin have BDI and the other has broadly defined BD phenotype (including unipolar depression)

Overall, an increase in the concordance rate is observed for bipolar disorder in the MZ twins compared to those in the DZ twins. Pooling the data together from these studies provides an estimate of MZ concordance for bipolar I disorder of 50% ($95\%CI: 40-60$) (Jones et al. 2002).

It is of note, that the concordance rate associated with monozygotic twins is not 100%. This suggests that factors which are not heritable (i.e. environmental factors) could also play part in developing bipolar disorder (Kelsoe 2009).

A genetic parameter called heritability provides an estimate of the proportion of causation of a specific character that is due to genetic causes (Strachan and Read 2004). Based on twin studies, the heritability for bipolar disorder has been estimated to be 93% ($95\%CI: 0.69-1.00$) (Kieseppa et al. 2004). A similar heritability estimate has been calculated by McGuffin et al.- 89% (McGuffin et al. 2003). A lower figure than these previous observations has been estimated by Lichtenstein et al.- 59% (Lichtenstein et al. 2009).

Another approach to separate genetic from environmental factors is using adoption studies. The most common approach is to study probands who have bipolar

disorder and who have been adopted at birth. Subsequently, the rates of psychiatric illness are determined in the biological and the adoptive parents (Kelsoe 2009).

Due to the inherent difficulties in making such observations, there are a limited number of studies. The results from the adoption studies have been insofar inconsistent. A study by Knorring et al. investigated the biological and adoptive parents of probands with affective disorder and the biological and adoptive parents of healthy controls. No statistically significant differences between the groups were observed. In contrast, a study by Mendlewicz and Rainer based on 29 bipolar and 22 control adoptees found almost a 3-fold increase in the rate of bipolar disorder in the biological parents of bipolar probands as compared to the adoptive parents (18% risk versus 7% risk respectively). Furthermore, in a Danish sample, a 3-fold increase in the rate of unipolar illness and a 6-fold increase in the rate of suicide in the biological relatives of probands with affective disorders have been noted as compared to the adoptive parents and the biological/adoptive parents of healthy probands (reviewed in (Jones et al. 2002)).

Adoption studies are characterised with difficulties in obtaining large number of subjects. Therefore the examined sample sets are quite small, which could be the reason why these data are less consistent than the data obtained using twin studies (Kelsoe 2009).

The recent large epidemiological study by Lichtenstein at al. screened adopted children whose biological parents had bipolar disorder. The results unequivocally supported the role of genetic variants in the susceptibility to bipolar disorder. Adopted children with an affected biological parent were > 4 times more likely to develop the disorder than the general population (Lichtenstein et al. 2009).

In conclusion, data from family, twin and adoption studies provide compelling evidence of the involvement of genes in developing bipolar disorder. In addition, these studies also show a gradation of the risk for developing bipolar disorder between the various classes of relatives with monozygotic co-twin showing the highest risk, when the lowest risk is observed in an unrelated person from the general population. The approximate risk for developing bipolar disorder in the lifetime of relatives of a bipolar proband is presented in Table 4.

Table 4 Approximate lifetime rates for developing bipolar disorder in various classes of relative to individual affected with bipolar disorder

Degree of relation to affected individual	Risk of bipolar disorder [%]
Identical co-twin	40-70
First degree relative	5-10
Unrelated individual from the general population	0.5-1.5

Adapted from (Jones et al. 2002)

1.1.3.2. Environmental factors

As noted previously, the concordance between monozygotic twins is not 100%. This suggests that other than genetic factors also play an essential role in the susceptibility to bipolar disorder.

The data with respect to the possible involvement of environmental factors are not conclusive. Sleep deprivation, drug abuse, hormonal alterations, pregnancy, childbirth, winter-spring birth, traumatic brain injuries, stressful life events have been found to be involved in bipolar disorder (Goodwin and Jamison 2007; Tsuchiya et al. 2003). Stressful life events do appear to trigger episodes of bipolar depression, while life events involving goal attainment appear to elicit manic symptoms (Johnson 2005).

1.1.4. Mode of inheritance

Despite the consistent body of data that bipolar disorder is a genetic disorder, the mode of inheritance is still not clear. This is likely due to the fact that no gene involved in the susceptibility to the disorder causes a 100% of the people who have the genotype to develop the disorder. In addition, there is a decrease in the observed rate of illness when the genetic distance from the affected proband increases (Table 4), which is not consistent with a single gene disorder (Craddock and Jones 1999). Therefore, bipolar disorder is not recognised as a disorder which is inherited in a Mendelian manner (that is, one to one relationship between genes and trait). However, there are families with multiple affected individuals, where the inheritance is likely to be in agreement with Mendel's laws (Blackwood et al. 1996; Freimer et al. 1996). Segregation analyses on large pedigrees have generated mixed results, with some studies showing consistency with a single gene model (Pauls et al. 1992; Spence et al.

1995), while others have not observed a transmission of a major locus (Bucher et al. 1981; Goldin et al. 1983). X-linked causation has been also suggested (Hayden and Nurnberger 2006). Nevertheless, the involvement of genes with major effect in certain families does not explain the majority of bipolar cases.

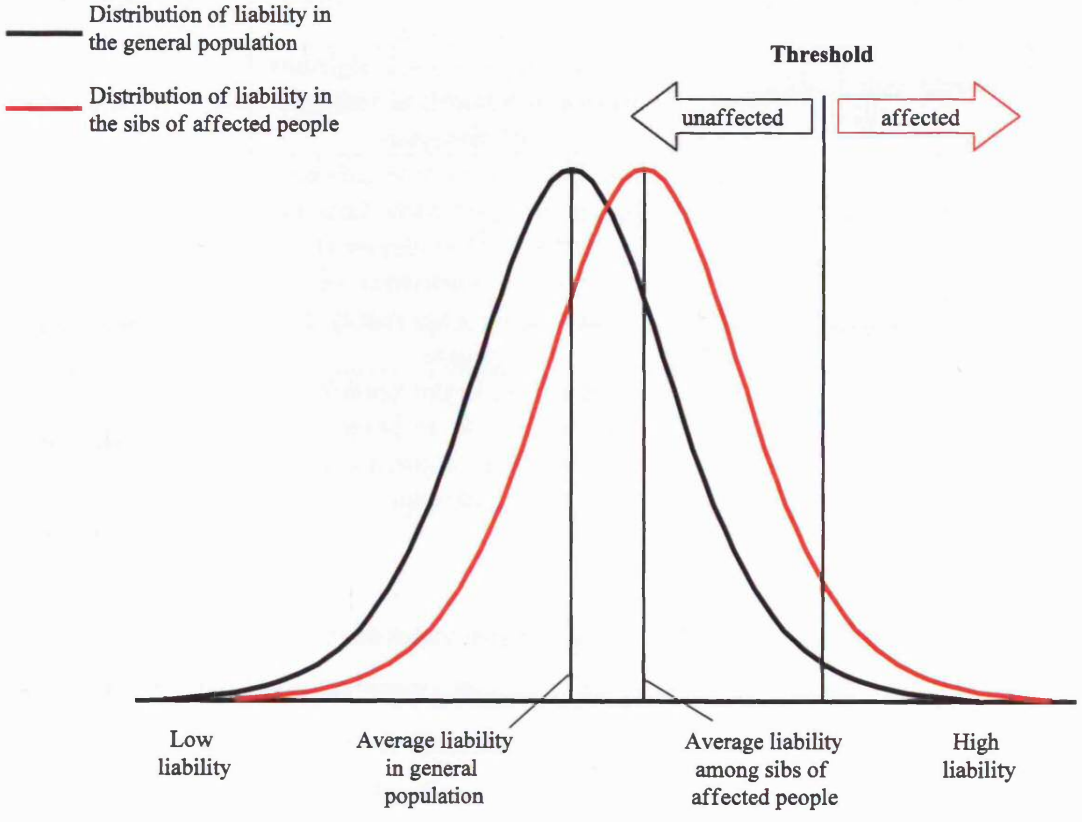
Based on these data, it has been suggested that bipolar disorder is a complex disorder, involving more than one locus (Jones et al. 2002). Complex non-Mendelian conditions like bipolar disorder are suggested to depend on two, three or many genetic loci, with contribution from environmental factors. Such conditions have been combined under the umbrella term- multifactorial disorders. The genetic determination could involve a small number of loci (oligogenic model) or many loci, each with a weak effect (polygenic model) (Strachan and Read 2004).

It has been suggested that the liability to develop bipolar disorder is a continuous trait and the disorder is expressed when a certain threshold along this continuum has been reached. Based on this hypothesis, when describing factors which determine developing conditions like bipolar disorder, often the term susceptibility is used. The reason stems from the hypothesis that such disorders are caused by a combination of many different genetic and environmental factors. Adding or removing one additional factor, could potentially determine disease or healthy status, without this specific factor being the cause of the disorder. In addition, Falconer and Mackay have also proposed the existence of a critical threshold for the underlying factors; the threshold can be regarded as the neutral point of the balance between the factors, and altering the balance of the factors can move the phenotype either way- disease or health (Falconer and Mackay 1996).

This model also provides explanation of the observed rate of illness in different classes of relatives, which exponentially declines as the genetic distance from the affected proband increases (presented in Table 4, page 14). Affected people have an unfortunate constellation of factors exceeding a certain threshold with each gene contributing a small amount towards the overall genetic effect. As probands's relatives share genes with them, they will also possess some of these factors and will have an increased susceptibility to the condition without surpassing the critical threshold. The liability to the disorder is defined by multiple genes acting together to determine a quantitative trait and follows a normal distribution in the population. People whose liability is above a certain threshold are affected. The distribution of liability in the relatives of affected people also follows a Gaussian distribution, but is

characterised with mean shifted to the right as on average they will have more susceptibility factors than the people sampled from the general population. Therefore a greater proportion of the relatives will have a liability over and above the critical threshold (Figure 2) (Falconer and Mackay 1996).

Figure 2 Falconer's threshold model for non-Mendelian characters



Adapted from (Strachan and Read 2004)

As previously suggested, multiple genes are likely to be involved in the susceptibility to bipolar disorder which leads to the following question: How do these genes interact to produce the phenotype? Several genetic mechanisms of how the underlying factors interact have been suggested in bipolar disorder. The proposed mechanisms are presented in Table 5, page 17.

Table 5 Genetic mechanisms suggested in bipolar disorder

Genetic mechanism	Description based on (Strachan and Read 2004)	Reference
Allelic heterogeneity	many different mutations within a given gene are causing the specific condition in different patients	(Sandkuyl and Ott 1989)
Locus heterogeneity	the clinical phenotype results from mutations at any one of several loci	(Hodgkinson et al. 1987)
Epistasis	multiple disease genes interact together to determine disease susceptibility	(Craddock and Jones 1999)
Dynamic mutation	unstable portions of DNA that expand when they are passed from parent to the child	(McInnis et al. 1993)
Imprinting	the expression of a gene depends upon its parental origin	(McMahon et al. 1995)
Mitochondrial inheritance	the disease mutation lies in the mitochondrial genome leading to a maternal pattern of inheritance	(McMahon et al. 1995)

adapted from (Jones et al. 2002)

Overall, even though there are many mechanisms proposed to explain the possible relations between the contributory factors, none of them have been unambiguously shown to be associated with bipolar disorder (Kelsoe 2009).

With a prevalence of 1% bipolar disorder is regarded as a common disorder in comparison to Mendelian monogenic diseases, which are very rare and relatively uncommon (Antonarakis and Beckmann 2006). Two hypotheses have been proposed about the genetic background of common disorders. Namely these are the common disease/common variant (CD/CV) and common disease/rare variant (CD/RV, multiple rare variant hypothesis) hypotheses (Alaerts and Del-Favero 2009). Pritchard and Cox suggested that less penetrant susceptibility variants (that is, variants that not always manifest themselves in a given phenotype), involved in complex diseases could be subject to less strong natural selection and therefore able to reach intermediate frequencies in the population (Pritchard and Cox 2002). According to the CD/CV hypothesis, the genetic risk is dependent on a few, common disease predisposing factors (present in $> 1\%$ in the general population) with a small predisposing effect (odds ratio < 2), such as the effect of several genes would be combined together to

produce the observed frequency of the disorder in the population (Alaerts and Del-Favero 2009; Bodmer and Bonilla 2008; Chakravarti 1999; Gershon 2000; Yang et al. 2005).

Although persistence in the population at this rate (i.e. 1%) is explicable by a polygenic model comprising many weak effects (that is, CD/CV model), it is also reasonable to postulate that variants of large effect may be replenished by new mutations. In addition, the involvement of common single nucleotide polymorphisms (SNP, any variation at a single nucleotide with a frequency > 1% in a population) has not yet explained a significant proportion of the genetic risk for developing common disorder like bipolar disorder (Cichon et al. 2009). Therefore, it has been suggested that a large number of rare variants (possibly hundreds of different genes) with relatively large effects (odds ratio > 2), linked with locus and allelic heterogeneity, could explain a substantial proportion of the susceptibility to neuropsychiatric disorders (Alaerts and Del-Favero 2009; Bodmer and Bonilla 2008; Gorlov et al. 2008; McClellan et al. 2007; Smith and Lusis 2002). The common disease-rare variant model allows for a heterogeneous collection of rare recent mutations that could account for some of the susceptibility to bipolar disorder (similar to Mendelian diseases) (Pritchard 2001; Reich and Lander 2001). With respect to CD/RV hypothesis, it has been argued that causal alleles will not reach common frequencies due to reduced fecundity, which will lead to negative natural selection. If "slightly deleterious" alleles account for modest changes in gene function; and if the combined action of several alleles result in an increased risk, then the selection may not be strong enough to purify new mutations, but will keep them at low frequencies (Alaerts and Del-Favero 2009).

In summary, it has been previously shown that genes influence the susceptibility to bipolar disorder. However, not 100% of the variance is explained by genetic factors. This suggests that other factors such as environmental factors, unknown genetic factors, gene-gene and gene-environment interactions, epigenetic effects and some factors with purely stochastic origin may influence the development of disease (Cichon et al. 2009). It is also important to acknowledge that the genetic complexity has to be fully appreciated when searching for the genetic factors predisposing to common disorders. Genetic mechanisms like dynamic mutations (e.g. expanding trinucleotide repeats), imprinting (e.g. Prader-Willi syndrome), copy number variation and mitochondrial inheritance may also influence and contribute to

the risk of developing neuropsychiatric illness. To add to the complexity that stems from gene-environment interplay, different genes and different genetic mechanisms may be involved in different pedigrees (that is genetic heterogeneity) with one or more affected individuals.

1.1.5. Identifying complex disease genes: linkage and candidate genes studies

Since it has been established that episodes of severe mood tend to run in families, the genetic factors influencing the susceptibility ought to be studied. Recent advances in molecular genetics provide the tools needed to identify such susceptibility variants. Strategies for disentangling the genetic component to bipolar disorder susceptibility comprise of linkage and association methods (Hayden and Nurnberger 2006).

1.1.5.1. Linkage analyses

In essence, linkage analysis involves genotyping several hundred markers that cover the human genome in families segregating the disease of interest. Markers inherited along with the disease are indicative of the chromosomal regions that could harbour disease genes (Kelsoe 2009).

Linkage analysis could be hypothesis driven (testing particular candidate genes) or genome-wide analysis without the requirement of a prior biologically driven hypothesis. It is applied in order to identify broad genomic regions on which the disease gene may reside. Linkage analysis makes use of the following characteristic of the genetic loci- if two loci are linked they will be transmitted together from a parent to the offspring more often than expected under independent inheritance. More specifically, two loci are linked if during meiosis, recombination between them occurs with $< 50\%$ probability. Linkage analysis is based on studying the chromosomal segments that are inherited together in a family, relying on the fact that the likelihood of sharing of chromosomal segments by the affected individuals is relatively high (Teare and Barrett 2005). Linkage is usually reported as a logarithm of the odds that loci are linked (LOD scores). Higher LOD scores are sign of a greater probability of linkage between the studied and disease variants (Hayden and Nurnberger 2006).

Linkage analysis has proved a very successful method for mapping loci causing rare Mendelian conditions. Unlike Mendelian disorders, complex diseases proved to be much more difficult with respect to using linkage as a method for identifying contributory variants (Jones et al. 2002). The difficulties arise as complex diseases pose some inherent challenges: the exact mode of inheritance is difficult to establish as incomplete penetrance, gene and locus heterogeneity have been observed. In addition, several/many genes and environmental factors may contribute to developing the disease. Therefore, for studying multifactorial disorders like bipolar disorder, model-free (non-parametric) linkage analysis is used. The basic principle is that, between affected relatives an excess sharing of alleles or chromosomal segments identical by descent (IBD) would be observed in the region of a disease-causing gene irrespective of the mode of inheritance. Several methods are applied to test if the IBD sharing at a locus is greater than expected under the null hypothesis of no linkage. The methods include testing sibling pairs, both of whom are affected with the disease and testing families with larger numbers of affected individuals of different relationship (Teare and Barrett 2005).

To date, a large number of linkage studies in bipolar disorder have been conducted. A large meta-analysis of 18 data sets was performed in an attempt to identify susceptibility regions for bipolar disorder in the combined data. No region reached genome-wide statistical significance. It was concluded that larger data sets are required (especially of cases with bipolar I disorder and schizoaffective disorder, bipolar type). It is of interest that no correspondence was observed between the highest ranked regions from linkage studies in bipolar disorder and schizophrenia (Segurado et al. 2003).

Hayden and Nurnberger have reviewed linkage studies after 1999, taking into account only linkage studies passing the Lander and Kruglyak's cut off for significant linkage result (i.e. LOD score of 3.3 when studying pedigrees and 3.6 when studying affected sibling pairs) (Lander and Kruglyak 1995) and instances where a region has been implicated by independent studies (Hayden and Nurnberger 2006). Based on these stringent criteria Hayden and Nurnberger concluded that the following loci are the most supported: 2p, 4p, 4q, 6q, 8q, 11p, 12q, 13q, 16p, 16q, 18p, 18q, 21q, 22q and Xq (Hayden and Nurnberger 2006).

In summary, the application of linkage analysis to complex conditions like psychiatric disorders has not been as successful as with rare Mendelian disorders. It

has been largely marred by unreplicated findings (Burmeister et al. 2008; Risch and Botstein 1996). In addition, the genes driving these statistically significant linkage results have not been clearly identified due to the relatively large size of the implicated regions. Such regions could comprise hundreds of genes, some of which may be biologically relevant candidates (Barnett and Smoller 2009; Teare and Barrett 2005).

A reason why the linkage studies in neuropsychiatric conditions were not as successful as in discovering genes for monogenic diseases, perhaps in part is because when genes of small effect are sought, the required sample sizes become very large. Based on power calculations, it has been shown that linkage analysis can provide evidence for the location of a disease locus with high genotype relative risk (≥ 4) and intermediate allele frequencies between 5 and 50%. The sample size of detecting linkage to genes with a genotype relative risk of < 2 (which is typical for genes with a small overall effect), could be prohibitively large (Risch and Merikangas 1996; Risch 2000).

Based on the results from linkage studies, one could conclude that no gene of major effect underlies susceptibility in most of the affected individuals. Furthermore as linkage studies have not been able to accurately and reproducibly identify loci involved in bipolar disorder, it has become clear that there are no common loci of large effect involved in developing the disease (Craddock and Sklar 2009). However, genes with large effect could still exist; they may be specific to particular families which could explain the large extent of non-replicated linkage studies. The results from linkage studies have generated evidence that bipolar disorder is a genetically complex disorder. In order to allow for this intrinsic complexity, the focus of the genetic research in bipolar disorder has shifted towards association studies. Such studies are more suitable for the inherent challenges in detecting susceptibility variants which could be relatively common in the population and comprise a modest increase in the risk for developing bipolar disorder (Barnett and Smoller 2009).

1.1.5.2. Association studies

In a seminal work published in 1996, Risch and Merikangas suggested that association studies are better equipped than linkage studies when genetic variants have weak effect on risk for developing of disease (Risch and Merikangas 1996). Furthermore, association studies are relatively easier to conduct as it is difficult to recruit extended families. Association studies are based on testing the association between a disease and a genetic locus using either family-based or case-control design (Burmeister et al. 2008). Initially, the design of the studies was based on a candidate gene approach, although these were surpassed by genome-wide based search.

1.1.5.2.1. Candidate genes approach

Before the advent of whole-genome arrays, earlier association studies of bipolar disorder have used relatively small sample sizes and tested polymorphisms in functional candidate genes, or genes suggested by linkage and cytogenetic studies. Some of the obvious candidate genes have been the ones involved in neurotransmitter systems targeted by medication. Namely, the antidepressant drugs block the re-uptake of serotonin and noradrenaline, when antipsychotic block the dopamine receptors. Such systems have been the "usual suspects"- dopamine, serotonin, glutamate and the noradrenalin systems. Genes from these systems which have received a lot of attention over the years include the genes encoding monoamine oxidase A (*MAOA*), catechol-O-methyltransferase (*COMT*), the serotonin transporter (*5-HTT*) and the NMDA glutamate receptor, subunit 2 B (*GRIN2B*) (Barnett and Smoller 2009; Craddock et al. 2001; Jones and Craddock 2001). In addition, genetic variants in biological candidate genes in a metabolic pathway that is either known or hypothesised to be associated with bipolar disorder have also been investigated (Burmeister et al. 2008). One such pathway has been the circadian system given the episodic nature of bipolar disorder (Craddock and Sklar 2009). It is of note, that lithium and valproate, used for treatment of bipolar disorder, have an effect on the circadian rhythm. Association studies have found some evidence for correlation between SNPs in genes that control circadian rhythms and predisposition to bipolar disorder- *CLOCK* and *BMAL1* genes (reviewed in (Barnett and Smoller 2009)).

Other genes investigated in bipolar disorder have been suggested on the basis that statistically significant associations have been observed in schizophrenia. To date, a number of genes already implicated in schizophrenia have been investigated in bipolar disorder. Namely, these are: disrupted in schizophrenia 1 (*DISC1*), G72/G30 locus, neuregulin 1 (*NRG1*), brain-derived neurotrophic factor (*BDNF*), dystrobrevin binding protein 1 (*DTNBP1*), tryptophan hydroxilase 2 (*TPH2*), dopamine receptor D4 (*DRD4*) and solute carrier family 6 (neurotransmitter transporter, dopamine), member 3 (*DAT1*) (Craddock and Sklar 2009; Serretti and Mandelli 2008).

For all of the aforementioned genes, studied for association with bipolar disorder, some statistically significant and some non-significant findings have been observed. The replication of association signals is essential for association studies, but as yet, no unequivocal support of the above-mentioned genes in developing bipolar disorder has been shown. One reason for this could be the inherent limitation of the candidate genes approach. It relies heavily on our current understanding of the disease pathology and hence depends on our prior knowledge. Studies which are not based on a certain biologically driven hypothesis and therefore explore new unbiased avenues of research are the genome-wide association studies.

1.1.5.2.2. Genome-wide association (GWA) studies

GWA studies have become possible due to the recent advances in methods for genotyping, statistical analysis tools and cost. They are based on the advent of high-throughput genotyping arrays and the extensive SNP information, provided by the HapMap Consortium (http://hapmap.ncbi.nlm.nih.gov/) (Burmeister et al. 2008; The International Hapmap Project 2003; Wollstein et al. 2007). These recent advances permit peppering the whole human genome with a dense collection of markers (hundreds of thousands, even millions). These SNPs are then genotyped simultaneously and rapidly in large sample sets and subsequently the data are interrogated for possible associations with disease without the need for a biologically-driven prior hypothesis about the studied variants (Alaerts and Del-Favero 2009; Barrett and Cardon 2006; Hirschhorn and Daly 2005).

Several genome-wide surveys in bipolar disorder have been published, with the first being the Wellcome Trust Case-Control Consortium study (WTCCC) on

1868 cases and 2938 controls (Wellcome Trust Case Control Consortium 2007). In Table 6 are summarised the recent whole-genome studies in bipolar disorder along with the main findings.

Table 6 Whole-genome association studies in bipolar disorder

Study (reference)	Sample		Method	Top hit (SNP, nearest gene, chromosome, significance)
	N cases	N controls		
WTCCC (Wellcome Trust Case Control Consortium 2007)	1868	2938	Affymetrix 500K array, individual genotyping	rs420259, *PALB2*, chr 16, $p = 6.3\times10^{-8}$
STEP-UCL (Sklar et al. 2008)	1461	2008	Affymetrix 500K/5.0 array, individual genotyping	rs4939921, *MYO5B*, chr 18, $p = 1.7\times10^{-7}$
ED-DUB-STEP2 (Ferreira et al. 2008)	1098	1267	Affymetrix 5.0/6.0 array, individual genotyping	rs7221510, *SKAP1*, chr 17, $p = 1.4\times10^{-6}$
NIMH/Bonn (Baum et al. 2008a)	1233	1439	Illumina 550 array, pooled and individual genotyping	rs1012053, *DGKH*, chr 13, $p = 1.5\times10^{-8}$
WTCCC, STEP-UCL and ED-DUB-STEP2 (Ferreira et al. 2008)	4387	6209	Meta-analysis	rs10994336, *ANK3*, chr 10, $p = 9.1\times10^{-9}$; rs1006737, *CACNA1C*, chr12, $p = 7.0\times10^{-8}$
NIMH/Bonn and WTCCC (Baum et al. 2008b)	3101	4377	Meta-analysis	rs10791345, *JAM3*, chr 11, random effect $p = 1\times10^{-6}$; rs4806874, *SLC39A3/ZIP3*, chr 19, $p = 5\times10^{-6}$
EA/AA study (Smith et al. 2009)	1000 EA 345 AA	1033 EA 670 AA	Affymetrix 6.0 array, individual genotyping	EA: rs5907577, chr X, $p = 1.6\times10^{-6}$; rs10193871, *NAP5*, chr 2, $p = 9.8\times10^{-6}$; AA: rs2111504, *DPY19L3*, chr 19 $p = 1.5\times10^{-6}$; rs2769605, *NTRK2*, chr 9 $p = 4.5\times10^{-5}$

EA=European ancestry; AA=African ancestry

Most of the GWA studies in bipolar disorder did not reach the generally accepted genome-wide statistical significance evidence level for association for GWA studies, i.e. applying a Bonferroni correction for 1 million independent tests (a

threshold of $p < 5.0 \times 10^{-8}$). A better approach, which has been proven successful in another complex disorder, type 2 diabetes, is pooling raw genotypes together (Zeggini et al. 2008). Such meta-analysis combining the WTCCC study and the NIMH/Bonn, suggested the involvement of additional two genes in the predisposition to bipolar disorder- *JAM3* and *SLC39A3/ZIP* (Baum et al. 2008a; Baum et al. 2008b; Wellcome Trust Case Control Consortium 2007). In addition, the most robust statistical evidence has come from a meta-analysis of WTCCC, STEP-UCL and ED-DUB-STEP2 (Ferreira et al. 2008). Two signals met the genome-wide statistical significance threshold of $p < 5.0 \times 10^{-8}$ in the large sample set comprising 4387 cases and 6209 controls- *CACNA1C* and *ANK3*. *CACNA1C* encodes an alpha-1 subunit of a voltage-dependent calcium channel (http://www.ncbi.nlm.nih.gov/gene). The *ANK3* gene is a member of a family of proteins that link the integral membrane proteins to the underlying spectrin-actin cytoskeleton and have key roles in activities such as cell movement, activation, contact, and maintenance of specialized membrane domains (http://www.ncbi.nlm.nih.gov/gene). Additionally, *ANK3* is known to modulate the activity of neuronal sodium channels. Interestingly, both genes implicated by the meta-analysis, code for proteins that are known to influence neuronal excitability through ion channel function. Markedly, dysfunctions in ion channel function have been implicated in diseases with episodic course like epilepsy, ataxia and migraine. The findings from the Ferreira et al. study have provided compelling evidence that ion channelopaties may also be associated with bipolar disorder (Craddock and Sklar 2009). Even though further evidence is emerging for the involvement of polymorphisms in *CACNA1C* and *ANK3* in the liability to bipolar disorder (Bigos et al. 2010; Erk et al. 2010; Schulze et al. 2009), additional work is required to build up the knowledge of the exact function and ultimately potentially targeting these genes by drugs.

Genome-wide association studies in bipolar disorder revealed several novel loci which have not been implicated before. Interestingly, evidence has not been observed for any of the previously associated candidate genes. This finding is in contrast to findings in type 2 diabetes, where the most significant GWA studies replicated previous findings (Scott et al. 2007). It has to be noted that the study on type 2 diabetes has been based on > 38,000 cases and controls, whereas the studies on bipolar disorder have examined fewer individuals. A reason for the little agreement of

association studies pre-WGA studies advent and since could be the inherent complexity of bipolar disorder. In addition, there is little correspondence between the most significant results coming from separate whole-genome studies in bipolar disorder. Possible explanations for such non-replications are that GWA studies are characterised with some inherent limitations which could result in detecting false-positive findings (type I errors). This could be caused by population stratification or lack of suitable multiple hypothesis testing. Another limitation of GWA studies is due to genotyping error, a true susceptibility variant could have been missed (that is, type II errors, β). Alternatively, the lack of replicated associations in bipolar disorder could be due to a non-sufficient number of studied individuals, i.e. genuine positive results are detected in one study but not replicated in another similarly sized sample set. A solution would be to analyse much larger sample group, i.e. increase in power to detect certain effect size ($1-\beta$).

In summary, WGA studies performed in bipolar disorder have resulted in mixed success. They have not been as successful as WGA studies in non-psychiatric common diseases, including diabetes, cardiovascular disease, Crohn's disease, etc. Nevertheless, new candidate genes have been suggested (i.e *ANK3*, *MYO5B*, *CACNA1C*, *DGKH* and others).

GWA studies are equipped to detect particular type of genetic variation (i.e. common SNPs) and a particular type of genetic effect (i.e. weak effect) (Newton-Cheh and Hirschhorn 2005). The detected polymorphisms explain a small proportion of the genetic variance and the majority of the genetic risk for developing bipolar disorder remains to be elucidated. There is a possibility that other types of genetic variation may account for some of the susceptibility to bipolar disorder.

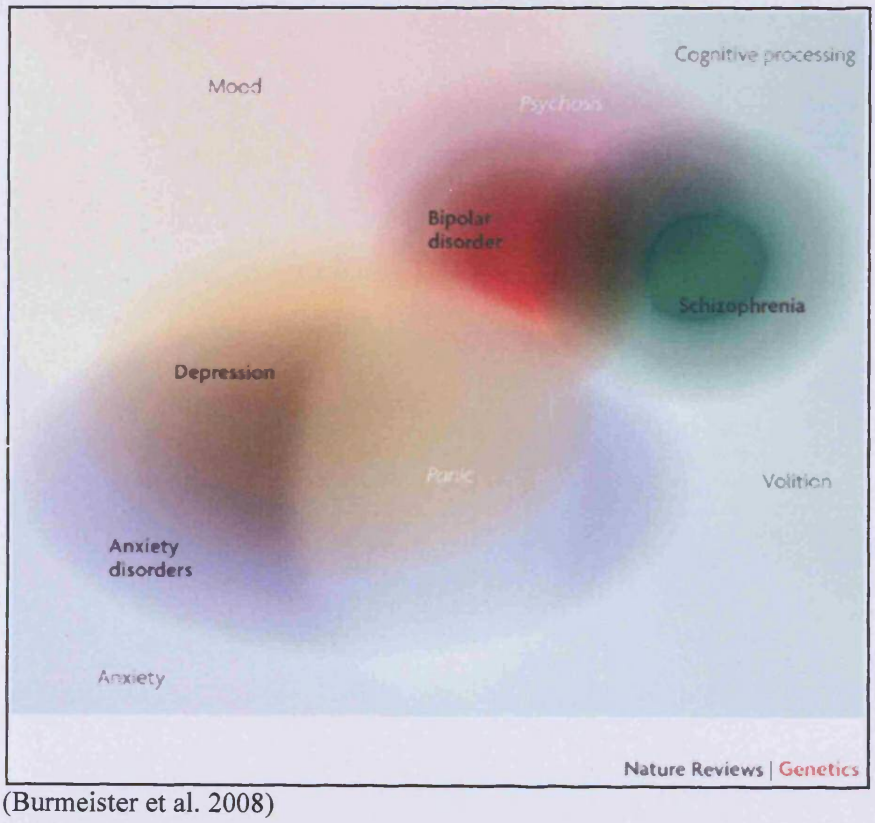
In conclusion, the phenotype and genetics of psychiatric disorders including bipolar disorder have proved to be complex. As a result, most genetic risk alleles have not been found, despite the high heritability observed in bipolar disorder. Potential reasons could be the involvement in the susceptibility of multiple genes, allelic heterogeneity, epistatic gene interactions and non-analysed environmental factors among other factors. In addition, genetic factors like *de novo* mutations, copy number variation, epigenetic changes, gene-gene interactions or a complex combination of all the factors have not received much attention as factors potentially accounting for liability to developing bipolar disorder.

1.1.6. Overlap between bipolar disorder and schizophrenia

It has to be noted that a psychiatric diagnosis is made on the basis of diagnostic criteria through clinical examination. Thus, one potential problem could be a partial overlap in the separate psychiatric categories. For example, psychosis (e.g. hallucinations or delusions) can be part of schizophrenia, bipolar disorder and psychotic depression. Furthermore, mood disturbances are common in schizophrenia (Craddock et al. 2009). Also the "intermediate" phenotype of schizoaffective disorder could suggest the existence of clinical continuum rather than a distinct separation between bipolar disorder and schizophrenia (Alaerts and Del-Favero 2009). Therefore, one could speculate that patients do not cluster neatly into separate diagnostic groups but rather intermediate forms are seen frequently.

In addition, if symptoms are not distinct, the boundaries between the separate diagnoses could be somewhat arbitrary (Figure 3, page 27) (Burmeister et al. 2008).

Figure 3 Graphical representation of the "boundaries" between the separate diagnoses



(Burmeister et al. 2008)

Therefore, Burmeister et al. concluded that due to the somewhat "blurred" boundaries between the separate diagnoses, it is possible that these diagnostic

categories are quite heterogeneous and that the diagnoses may not reflect the true genetic aetiology (Burmeister et al. 2008).

It has been observed that relatives of bipolar probands are at an increased risk of developing other psychiatric phenotypes, viz. unipolar depression, schizoaffective disorder and schizophrenia (Craddock et al. 2005; Mortensen et al. 2003; Valles et al. 2000). This suggests that certain, currently unknown genetic factors could be shared between bipolar disorder and schizophrenia. Rzhetsky et al. analysed 1.5 million patient records from a clinical database at the Colombia University Medical Center and 161 disorders to examine if there was a correlation between some of the disorders. Based on these extensive data, it was suggested that autism, schizophrenia and bipolar disorder share significant genetic component (Rzhetsky et al. 2007).

There is accumulating body of experimental evidence that bipolar disorder and schizophrenia could share some of the genetic susceptibility factors (Bramon and Sham 2001; Craddock et al. 2005). A balanced translocation between chromosomes 1 and 11 ((1;11)(q42;q14.3)), involving *DISC1* (disrupted in schizophrenia 1) and *DISC2* (disrupted in schizophrenia 2) genes has been observed in both schizophrenia and bipolar disorder probands in the same Scottish pedigree (Blackwood et al. 2001). It has been shown that bipolar disorder and schizophrenia have shared linkage signals at identical loci in the human genome, i.e. 6q21-25, 10p14, 13q32-34, 18p11 and 22q11-13 (Barnett and Smoller 2009; Berrettini 2000; Bramon and Sham 2001; Smoller and Gardner-Schuster 2007). In addition, some genetic loci have been proposed to be susceptibility factors for both schizophrenia and bipolar disorder: *BDNF*, *COMT*, *NRG1*, *G72/G30* (*DAOA*) and *DISC1* (Kelsoe 2009). Furthermore, Moskvina et al. analysed genome-wide association data from schizophrenia and bipolar disorder and observed an excess of genes showing evidence for association for both of the disorders (Moskvina et al. 2009). Based on these data, one could speculate that the genetic associations are not unique to one of the traditional diagnostic categories of schizophrenia and bipolar disorder.

Lichtenstein et al. studied a population-based group of more than two million families. In this group, large sample sets of patients suffering from bipolar disorder (~40,000 cases) and schizophrenia (~35,000 cases) were screened. When relatives of bipolar disorder probands were analysed, an increased risk not only for developing bipolar disorder, but also for developing schizophrenia were found to be elevated (Table 7).

Table 7 Relative risk for schizophrenia and bipolar disorder

Relation to proband		Risk for SZ when proband has BD		Risk for BD when proband has SZ	
		RR	95%*CI*	RR	95%*CI*
Biological relationship					
Parent	Offspring	2.4	2.1-2.6	5.2	4.4-6.2
Sibling	Sibling	3.9	3.4-4.4	3.7	3.2-4.2
Sibling	Maternal half-sibling	1.4	0.7-2.6	1.2	0.6-2.4
Sibling	Paternal half-sibling	1.6	1.0-2.7	2.2	1.3-3.8
Adoptive relationships*					
Biological parent	Adopted away offspring	4.5	1.8-10.9	6.0	2.3-15.2

* the biological parent in the adoptions was the one with the illness, not the proband
(Lichtenstein et al. 2009)

The increased risks for schizophrenia existed for all relationships, including adopted children to biological parents with bipolar disorder. Furthermore, the comorbidity between the two disorders was found to be mainly ascribable (63%) to additive genetic effects common to both of the disorders thus providing evidence that bipolar disorder and schizophrenia partly share common genetic cause. In addition, it has been also observed that a proportion of the genetic variance is not attributable to common genetic factors shared between the disorders suggesting that some genes may be associated with risk for both of the conditions while some- with the risk for one of the disorders (Lichtenstein et al. 2009). The study of Lichtenstein et al. is of great importance in the elucidation of a possible common genetic cause between bipolar disorder and schizophrenia, as it makes use of extremely large meticulously analysed sample and is likely to yield robust results. Another large cohort of 2.6 million persons and their biological parents from Denmark have provided evidence for a possible overlap in the determinants between the two disorders (Gottesman et al. 2010). In couples where both individuals have been diagnosed with schizophrenia, the incidence of the offspring to have bipolar disorder was ~ 10 times higher than in the general population (10.8%; $95\%CI: 2.6-19.0$).

The separation between schizophrenia and bipolar disorder initially originated from Emil Kraepelin (reviewed in (Kendler 1986)). However, the recent evidence challenges the Kraepelinian dichotomy of *dementia praecox* and manic-depressive illness. The hypothesis of a continuum between the disorders, rather than separate conditions, has been supported by a large body of data (Craddock and Owen 2005).

In summary, it has been shown that there is an overlap in the risk across bipolar disorder and schizophrenia. This suggests that some of the susceptibility factors could be shared between the disorders. Therefore it is hoped that the research coming from genetic studies will elucidate the relationship between these diagnostic entities which were initially thought to be separate disorders.

1.2. Copy number variation

1.2.1. Description

Large duplications and deletions have been detected in the human genome for quite some time (Miller and Therman 2001). Until recently, their frequency was presumed to be low and directly related to a specific rare genetic disorder (Freeman et al. 2006). However, in recent years, it has become clear that deletions and duplications are widely-spread in the human genome and that this type of variation does not always lead to a disease (Iafrate et al. 2004; Redon et al. 2006; Sebat et al. 2004). Apart from being part of the normal genetic variation, structural changes in the human genome like deletions and duplications are being recognised as a cause for developing diseases, which have been termed genomic disorders (Lupski 2009).

Historically, the first differences observed between the genomes of two individuals, have been mainly rare changes in the quantity and structure of the chromosomes. Namely, these have been aneuploidies, structural rearrangements, heteromorphisms and fragile sites (Miller and Therman 2001). Detecting these changes is facilitated by their large size ($\geq$ 3 Mb) as they are visible with a microscope (Feuk et al. 2006a).

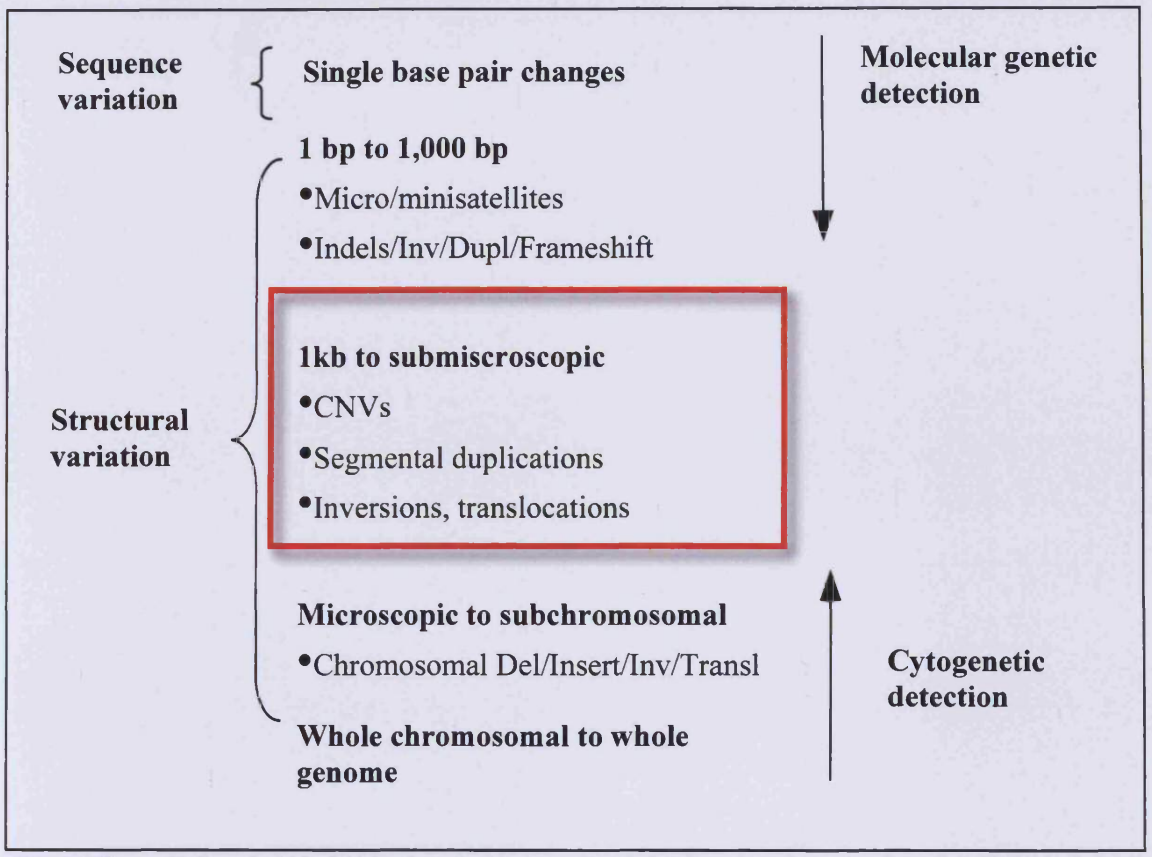
Consequently with the development of new molecular biology methods, the detection of smaller ($\leq$ 3 Mb) variations has become possible (Figure 4, page 32). These include single nucleotide polymorphisms (SNPs), repetitive elements that comprise short DNA sequences (i.e. microsatellites, minisatellites, small (< 1 kb) insertions, deletions, duplications, inversions and frameshift mutations) (Feuk et al. 2006a). These 1 – 1000 bp variants were thought to be the most widely-spread type of genetic variation in the human genome, with SNPs being the most abundant, numbering at least 10 million (Kruglyak and Nickerson 2001). The variation at single nucleotide level has been studied extensively (Frazer et al. 2007; International

HapMap Consortium 2005; The International Hapmap Project 2003). The analysis of the human genome has suggested that single nucleotide polymorphisms are the main source for genetic and phenotypic variation (Feuk et al. 2006a; Freeman et al. 2006; Scherer et al. 2007; Sharp et al. 2006a). In addition, the microscopic variation (that is ≥ 3 Mb) in the human genome has been examined comprehensively too (Jacobs et al. 1992; Warburton 1991). Until recently, the variation between these two categories (i.e. 1kb < variation < 3 Mb) has been largely unknown.

However, due to the discovery of new methods and tools for genotyping, knowledge with regards to DNA variation involving segments > 1 kb and < 3 Mb, has started to emerge (Iafrate et al. 2004; Redon et al. 2006; Sebat et al. 2004; Tuzun et al. 2005).

Variants within this range and present at a variable copy number in comparison with a reference genome, are defined as submicroscopic structural variants and are referred to as copy number variants (CNV) (Feuk et al. 2006a).

Figure 4 Types of genomic variation



adapted from (Scherer et al. 2007); Inv-inversion; Dupl-duplication; Insert-insertion; Del-deletion; Transl-translocation

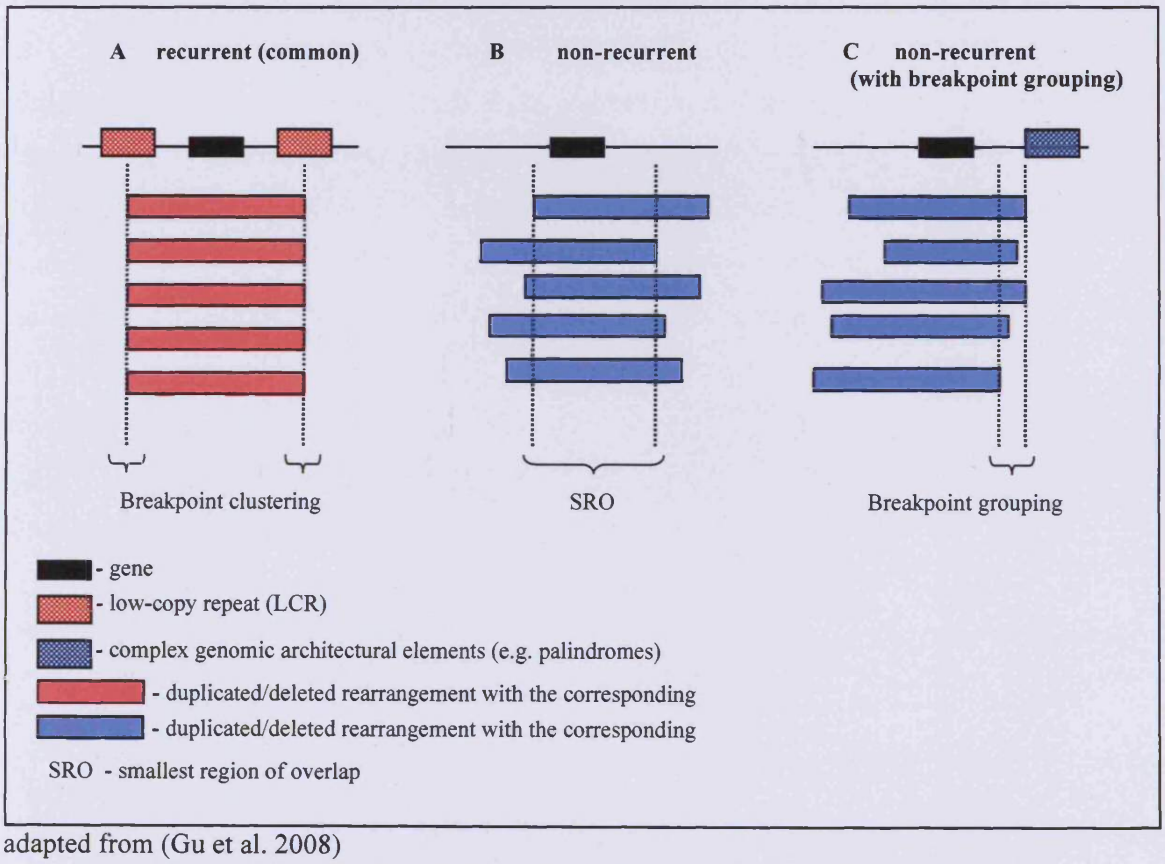
Iafrate et al. and Sebat et al. were the initial two studies that pointed to the intrinsic dynamic structure of the human genome (Iafrate et al. 2004; Sebat et al. 2004). Furthermore, these two breakthrough studies demonstrated that copy number variants are widely distributed throughout the human genome and could underlie normal human phenotypic variation and susceptibility to disease (Iafrate et al. 2004; Sebat et al. 2004).

1.2.2. Mechanisms for the occurrence of copy number variants

Three mechanisms have been proposed for the occurrence of copy number variants in the human genome (Lupski and Stankiewicz 2005). Specifically, these comprise non-allelic homologous recombination (NAHR), non-homologous end-joining (NHEJ) and fork stalling and template switching (FoSTeS) (Lee et al. 2007; Lupski and Stankiewicz 2005). These mechanisms could arise in both germ and

somatic cells. If they occur in germ cells, the genomic rearrangements could lead to a genomic disorder, and if they occur in somatic cells, the rearrangements can cause disorders like cancer (Gu et al. 2008). If the disease-causing genomic rearrangements share a common size and fixed breakpoints, they are said to be recurrent and they could share similar clinical manifestations in multiple affected individuals. Non-recurrent rearrangements are of different sizes in affected individuals. Even though, the rearrangements do not affect exactly the same genetic region, they could share a region of overlap, which could explain shared clinical features in different patients (Figure 5, page 33) (Gu et al. 2008).

Figure 5 Recurrent and non-recurrent genomic rearrangements associated with genomic disorders



adapted from (Gu et al. 2008)
A. Recurrent copy number variants, sharing a common size and common breakpoints in different patients;
B. Non-recurrent rearrangements. They do not involve the same breakpoints, but they share a region of overlap, which could explain common phenotypic features in different patients;
C. Non-recurrent rearrangements with grouping of one breakpoint. The grouping of the breakpoint is dissimilar to the one observed in breakpoint clustering as these non-recurrent rearrangements have one of their breakpoints localized in one small genetic region. Like clustering where low copy repeats are responsible for the genetic rearrangements, here the underlying genetic culprit for the rearrangement could be a DNA palindrome

1.2.2.1. Non-allelic homologous recombination (NAHR)

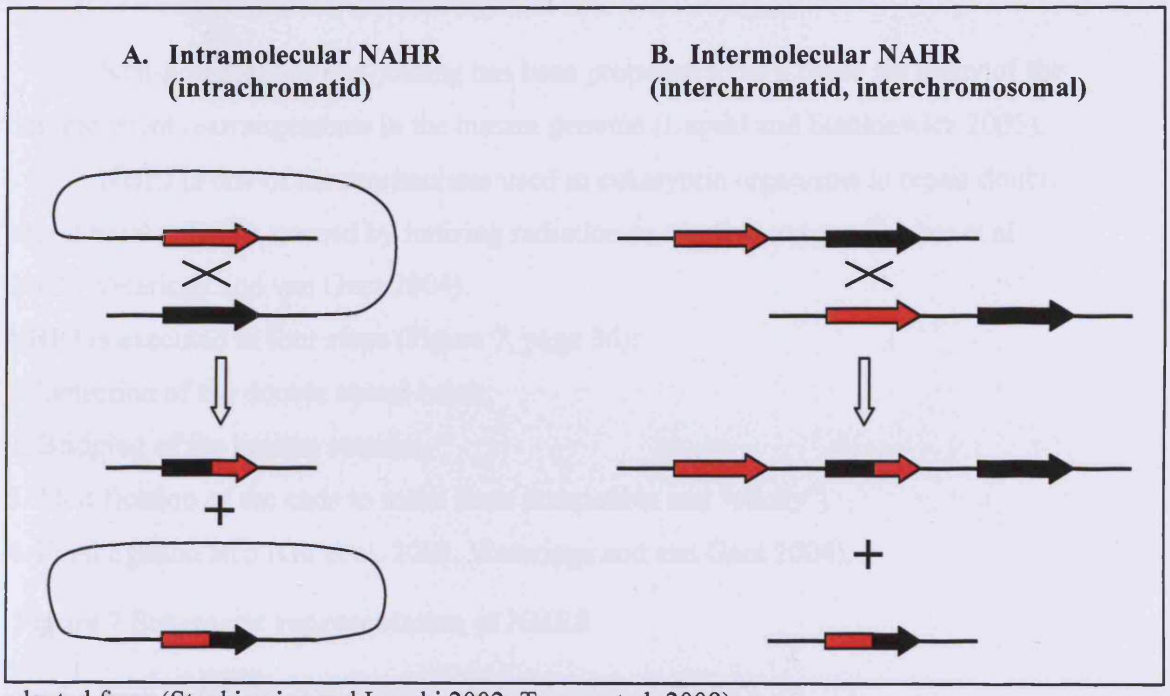
Non-allelic homologous recombination has been proposed to be the major mechanism for occurrence of recurrent rearrangements in the human genome (Lupski and Stankiewicz 2005).

Most recurrent genomic rearrangements originate by NAHR between two low copy repeats (LCR, also called segmental duplications, SD) (Stankiewicz and Lupski 2002). Low copy repeats are chromosome-specific DNA sequences, normally with size variation between 10 and 300 kb and of **> 95%** similarity to each other (Stankiewicz and Lupski 2002). As a result of the high degree of sequence similarity, in meiosis and mitosis, non-allelic copies of LCRs can be aligned, as a substitute of the normal allelic positions. The outcome of this "misalignment" between non-allelic sequences and the following crossover between them could lead to a genomic rearrangement in the daughter cells (Gu et al. 2008). Thus, the non-allelic LCRs play a role of "mediators" for the homologous recombination. In addition, they are the cause for the observed breakpoint clustering between different individuals (Figure 5, page 33) (Gu et al. 2008).

A plethora of data has shown that segmental duplications are associated with NAHR (Redon et al. 2006; Sebat et al. 2004; Sharp et al. 2005). Itsara et al. replicated this finding and demonstrated that NAHR is a major contributor to copy number variation. Specifically, a 25-fold enrichment of CNVs between pairs of homologous LCRs was observed (Itsara et al. 2009).

Depending on the location of the NAHR, there are different classes (Figure 6, page 35) (Lupski 1998). Recombination between low copy repeats can occur in one of three ways: between paralogs on the same chromatid (i.e. intrachromatid), between LCRs on sister chromatids (i.e. interchromatid or intrachromosomal) and between those on homologous chromosomes (i.e. interchromosomal) (Lupski 1998; Stankiewicz and Lupski 2002; Turner et al. 2008).

Figure 6 Different classes of NAHR



adapted from (Stankiewicz and Lupski 2002; Turner et al. 2008)
A. Intrachromatid NAHR generates a deletion and a circular DNA molecule;
B. Interchromatid and interchromosomal NAHR generates deletion and reciprocal duplication;
Low copy repeat
The thin line represents the genomic region affected by NAHR

Examples of genomic diseases caused by NAHR include Charcot-Marie-Tooth disease type 1, Hereditary Neuropathy with Liability to Pressure Palsies, Potocki-Lupski syndrome, Smith-Magenis syndrome, DiGeorge/Velocardiofacial syndrome and *inter alia* many cancers are also thought to originate from somatic NAHR (Gu et al. 2008).

1.2.2.2. Non-homologous end-joining (NHEJ)

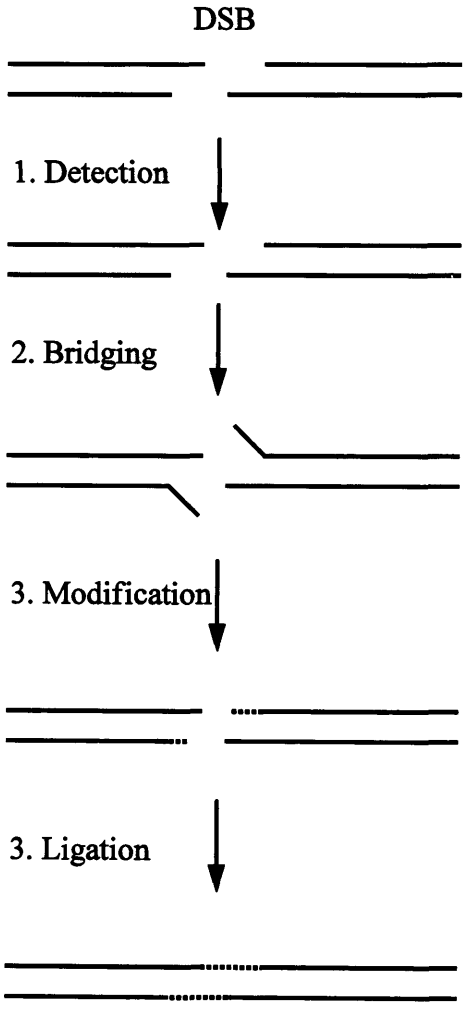
Non-homologous end-joining has been proposed to be a cause for many of the non-recurrent rearrangements in the human genome (Lupski and Stankiewicz 2005).

NHEJ is one of the mechanisms used in eukaryotic organisms to repair double strand breaks (DSB), caused by ionizing radiation or reactive oxygen (Lieber et al. 2003; Weterings and van Gent 2004).

NHEJ is executed in four steps (Figure 7, page 36):

1. Detection of the double strand break;
2. Bridging of the broken strands;
3. Modification of the ends to make them compatible and "sticky";
4. Final ligation step (Gu et al. 2008; Weterings and van Gent 2004).

Figure 7 Schematic representation of NHEJ



adapted from (Gu et al. 2008)
DSB- double strand break

When compared with NAHR it is obvious that low copy repeats are not required for the occurrence of NHEJ. A characteristic of the NHEJ mechanism is that it leaves an "information scar" at the rejoining site as the pre- joining editing of the ends (during step 3) involves cutting out or adding nucleotides from the ends or to the ends. This potentially could lead to a deletion of some genetic material (Gu et al. 2008).

The proposed mechanism by which the non-homologous end-joining could cause the occurrence of duplication is the following: if a strand break has occurred on one of the DNA strands, one of the broken ends could potentially invade and copy using the sister chromatid as a template. After DNA synthesis, the chromosome is repaired by rejoining of ends by the NHEJ mechanism (Lee et al. 2006a).

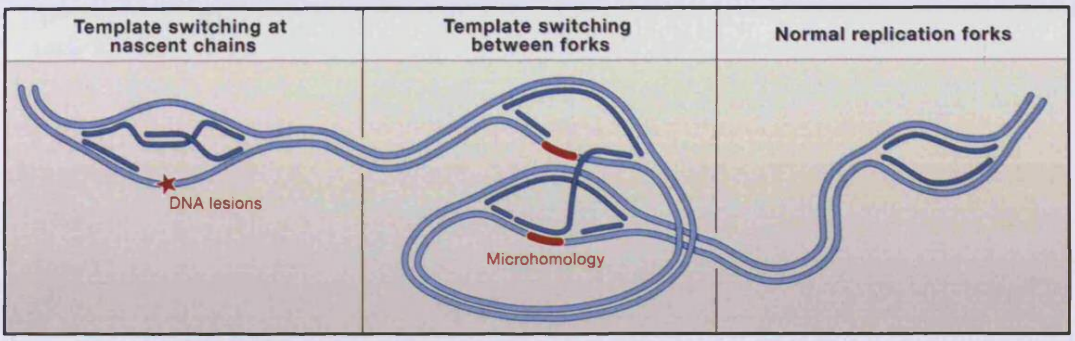
As yet, the extent of the involvement of non-homologous end-joining as a mechanism for occurrence of copy number variation has not been fully characterised (Lupski and Stankiewicz 2005).

1.2.2.3. Fork stalling and template switching (FoSTeS)

FoSTeS can lead to non-recurrent rearrangements and is a replication-based mechanism (Lee et al. 2007).

During DNA replication, the active replication fork can stall and pause near DNA lesions or near low copy repeats. The lagging strand disengages from the original chain and due to sequence similarity can switch to another replication fork and start DNA replication. Some unusual genetic architecture such as microhomology due to LCR, palindromes, cruciforms or non-B DNA structure could facilitate this faulty replication. The fork may be in physical proximity, but separated by some distance. The result of this switch could be joining of different sequences from separate genomic positions. If the switching was to a downstream fork, the end result would be a deletion, whereas if the switching was to a fork located upstream, the result would be a duplication (Gu et al. 2008; Lee et al. 2007). Schematic representation of the FoSTeS mechanism is given in Figure 8.

Figure 8 Schematic representation of FoSTeS mechanism



(Branzei and Foiani 2007)

Lee at al. hypothesised that this could be the mechanism by which complex genomic rearrangements associated with a dysmelinating central nervous system disorder occurs (Pelizaeus-Merzbacher disease) (Lee et al. 2007). It has also been suggested that duplications at amyloid precursor protein gene (*APP*), which have been found to be associated with Alzheimer's disease, could have occurred by this mechanism (Lee et al. 2007). Furthermore, it has been hypothesised that FoSTeS is potentially the main mechanism for the origin of duplications and the occurrence of LCRs in the human genome (Gu et al. 2008).

1.2.3. Frequency of copy number variants in the general population

"Each human's genome is distinguished by extra, and sometimes missing, DNA that can powerfully impact everything from development to disease" (Cohen 2007).

The advent of the genome-scanning technology has revealed an unexpected magnitude of deletions and duplications in the human genome. Therefore, the extent of copy number variation in healthy populations and disease had to be determined.

In 2004, two papers presented analysis of CNVs in healthy human populations. The main observation was that on average each person has ~12 CNVs across the genome (Iafrate et al. 2004; Sebat et al. 2004). When the total number of nucleotides affected by these CNVs was taken into account, it was hypothesised that CNVs contribute to the overall variation between two people at the same rate as SNP. Another study suggested that 12% (~360 Mb) of the human genome is covered by CNVs and that they are affecting many genes and functional elements (Redon et al. 2006).

Evidence from sequencing of the genome of one individual has revealed that insertions and deletions comprise a minority of the observed variations (22%), while they account for ~ 74% of the affected nucleotides (Levy et al. 2007). Taking CNVs into account has showed that two haploid genomes differ by 0.5% which was a substantial increase over the previously estimated difference of 0.1% (Levy et al. 2007; Venter 2010). It was later discovered that the genomes of different individuals may differ by between 1% and 3% (Venter 2010).

Itsara et al. studied 2500 individuals for CNV by mining data from an Illumina array and observed that 65% to 80% of the individuals have a CNV > 100 kb, 5 – 10% of the individuals have CNVs > 500 kb, while 1 – 2% have CNVs > 1 Mb. The average amount of CNVs per person is estimated to be between three and seven variants (Itsara et al. 2009). Another important observation has been that the majority of the genomic variations are present at ~ 0.02 to 1% frequency and spanning 6% of the human genome, whereas polymorphic CNVs encompass 0.09% of the genome. Another study also reported that large CNVs affect much less of the genome than previously thought (i.e. the estimation by Redon et al. of 12%). Furthermore, the detected CNVs were smaller than the ones found by Redon et al., with which a direct comparison was performed (McCarroll et al. 2008). This overestimation has been explained by the usage of large insert BAC clones, which are characterised by decreased sensitivity. In conclusion, these data have not replicated the initial estimation of 12% of the genome is encompassed by CNVs, but suggested rates closer to 5%. This rate was also further supported by Pinto et al. who studied healthy control population and observed that 160 Mb of the genome (~ 5%) is covered by CNVs. Of these CNVs, 96% are rare with frequency of < 2%, and the rest are common (Pinto et al. 2007).

Currently, it is recognised that CNVs are surprisingly omnipresent in the human genome. They can be common (frequency > 1%) or rare (frequency < 1%), inherited or *de novo*, biallelic or multiallelic (Alaerts and Del-Favero 2009). With respect to their size, a negative correlation has been observed, i.e. decrease in size results in increase in frequency (Estivill and Armengol 2007).

Like other mutations, it is thought that copy number variation is a mechanism that contributes to the plasticity of the human genome and ultimately to evolution. To this end, it is not surprising that CNVs often affect genes that are involved in sensory

perception and immune response functions (Alaerts and Del-Favero 2009; Cooper et al. 2007). It has been suggested that the observed enrichment of CNVs in genes associated with olfaction and immunity is a sign of adaptive benefit of an increased gene dosage (Nguyen et al. 2006). An example is the observed copy number variation that affects the *CCL3L1* gene (human immunodeficiency virus-1-supressive chemokine and ligand for the HIV coreceptor CCR5). Interestingly, a lower *CCL3L1* copy number than the populational average, is associated with susceptibility to HIV (Gonzalez et al. 2005). The more copies are carried of the *CCL3L1* gene, the less likely is a person susceptible to HIV. The gene codes for a protein which binds to a receptor, used by the virus to penetrate the white blood cells; hence the more protein produced, the fewer receptors are available for annealing of HIV.

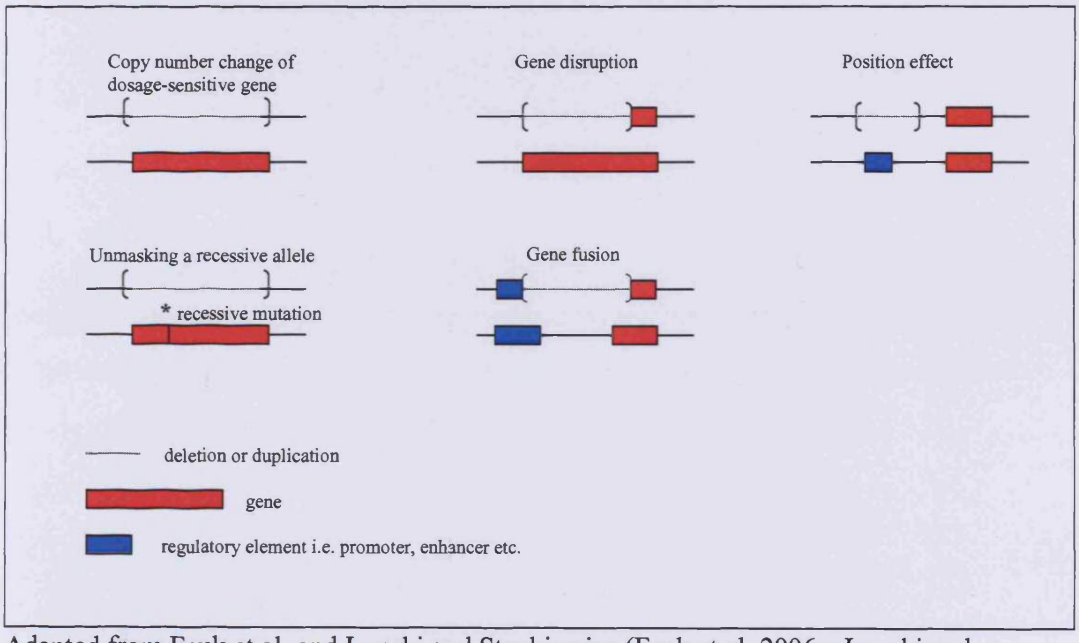
Another example suggesting that copy number variation could be a substrate for natural selection, resulting in differences between individuals and populations, is a duplication affecting the salivary amylase (*AMY1*) gene. *AMY1* gene participates in the digestion of starch in food. Individuals can have between 1 and 10 copies of this gene with a corresponding level of amylase in the saliva (Perry et al. 2007). There is a clear difference between the populations in the number of copies depending on the amount of starch consumed, with the higher number of copies observed in populations with a high-starch diet (Perry et al. 2007). It has been argued that this example of positive selection may improve the digestion of starchy food and possibly prevent intestinal disease (Perry et al. 2007).

In conclusion, there is evidence that copy number variation is one of the factors accounting for normal genetic differences between people. In addition, it is reasonable to speculate that if a disorder has a large genetic component accounting for the susceptibility, copy number variation can be involved in the aetiology (based on the fact that CNVs are genetic variants). CNVs can influence the expressed phenotype by several mechanisms which are discussed in the next paragraph.

1.2.4. Influence of copy number variants on phenotype

It is logical to hypothesise that changes in copy number may exert an effect on the phenotype by changes in gene expression levels (Henrichsen et al. 2009a). Schematic representation of the manner by which the phenotype could be influenced by changes in copy number is presented in Figure 9.

Figure 9 Schematic presentation of the influence of copy number variation on phenotype



Adapted from Feuk et al. and Lupski and Stankiewicz (Feuk et al. 2006a; Lupski and Stankiewicz 2005)

Rearrangements could influence the phenotype by the following mechanisms:

A. Gene dosage: altering the copy number of a gene sensitive to dosage effects, may lead to a change in the amount of mRNA leading to a change in the amount of the affected protein. For certain genes and pathways for which the amount of the product is crucial, this could lead to a disease (Shianna and Willard 2006). An example of such mechanism is the duplication of the gene encoding peripheral myelin protein-22 (*PMP22)*, which leads to an extra dose of the gene, causing the sheaths that protect nerve cells to disintegrate and thus obstructing the signal between brain and hands and feet (Cohen 2007). The resulting disease is called Charcot-Marie-Tooth disease type 1A. It is of note that a deletion at the same locus causes another neuropathy- hereditary neuropathy with liability to pressure palsies;

B. Gene disruption: genes that partially overlap with a copy number variant will be directly disrupted. This could lead to reduced expression of dosage-sensitive genes (Feuk et al. 2006a);

C. Position effect: could be manifested due to either deletion or duplication (Kleinjan and van Heyningen 2005; Lee et al. 2006b). There is evidence that regulatory elements can be a million base pairs away from a gene, hence even if the CNV does not directly affect a gene, it can still affect it in an indirect manner by affecting the regulatory elements (e.g. enhancer, repressor elements, etc.) (Feuk et al. 2006b). Merla et al. studied a deletion associated with Williams-Beuren syndrome and observed that not only is the deletion correlated with low relative levels of expression of genes within the deletion, but also with a decreased expression of genes up to several megabases from the copy number rearrangement (Merla et al. 2006);

D. Unmasking of recessive mutation by deletion: in dosage-sensitive genes a deletion could also lead to disease if it reveals a recessive mutation on the homologous chromosome (Feuk et al. 2006a);

E. Fusion genes at the breakpoint generating gain-of-function mutation: copy number variants in which breakpoints coincide with a gene, could also lead to formation of new transcripts through gene fusion or exon shuffling (Feuk et al. 2006a).

Stranger et al. ascertained the relative contribution of CNVs to phenotypic variation. They analysed the gene expression levels of ~14,000 genes in individuals that participated in the HapMap project and estimated the association of the expression levels with CNVs. The main observation was that CNVs account for ~ 20% of the detected genetic variation; the rest was explained by SNPs (Stranger et al. 2007).

Evidence that CNVs have a phenotypic effect also comes from studies in mice and rats. Henrichsen et al. studied whole genome expression data from six major organs in mice. Their primary finding was that not only the expression of genes that map within CNVs was associated with changes in copy number, but CNVs were also found to influence the expression of genes which were not directly affected by CNVs, but were up to 500 kb away from the copy number change (Henrichsen et al. 2009b). This study was one of the first to demonstrate that CNVs have direct influence on

tissue transcriptomes and the findings are consistent with Merla et al. study presented above.

As yet, not many studies have investigated the functional impact of CNVs at the cellular and organism level, further studies are warranted to assess the global influence of CNVs on the gene expression, phenotypic variation and the correlation with disease.

1.2.5. Methods for copy number detection

The methods for copy number detection are constantly evolving with a continuing improvement in resolution. The whole genome or a specific sequence could be targeted. The whole-genome scans include karyotyping, SNP arrays used to identify CNV structure, array-based comparative genome hybridisation and computational approaches based on sequence-assembly comparison. Targeted scans include multiplex amplification and probe hybridisation, quantitative PCR, fluorescence in situ hybridisation and multiplex ligation-dependent probe amplification. The list presented above does not intend to be fully comprehensive.

The method for inferring copy number variants used in this PhD work was based on whole-genome SNP array data. The details will be presented further in the Methods section of the thesis.

Usually, when reporting genomic imbalances observed with one of the above-mentioned methods, the detected variants are subsequently confirmed using an alternative method for detection. Compared to SNPs which can be detected with high confidence, CNVs are much more difficult to identify. With respect to SNPs the identification involves identifying which nucleotide is present at a certain locus, whereas CNVs involve deletion/duplication of a certain sequence which is more challenging to infer. Therefore, a validation with a different than the initial discovery method is required in order to ensure that the observed results are robust.

With the improvement of methods for CNV detection, studies of the possible involvement of this type of variation in various disorders are beginning to emerge. In the next section I will present evidence with respect to the potential association between copy number variation and susceptibility to psychiatric disorders.

1.2.6. CNV involvement in neuropsychiatric disorders

Structural variation was found to be a potentially important factor for developing neuropsychiatric conditions such as mental retardation, autism and schizophrenia.

Herein I will predominantly focus on studies published before 2008 as this was the start of the work described in this PhD thesis. In this way I will provide a snapshot of the field as it was at the beginning of my PhD. More recent findings are discussed along with the main findings from this study in the Discussion chapter (page 132).

1.2.6.1. Mental retardation

The cause of mental retardation (MR) is currently unknown in 1/3 to 1/2 of the cases. It is largely assumed that chromosomal abnormalities (visible with microscope) are the most frequent cause and may account for up to 10% of cases (Friedman et al. 2006).

The development of methods for CNV detection has prompted the discovery of chromosomal imbalances which are smaller than the ones, detected using light microscopy karyotyping (Friedman et al. 2006; Vissers et al. 2003). Table 8 (page 45) summarises the main findings (until 2008) with respect to the involvement of structural variants in developing mental retardation.

Table 8 Pathogenic CNV in mental retardation

Reference	Diagnosis	N cases	Locus	% of cases with CNV
(de Vries et al. 2005)	MR	100		10 (*de novo*)
(Friedman et al. 2006)	Idiopathic MR	100		11 (*de novo*)
(Shevell et al. 2008)	Global developmental delay	94		6.4
(Shaffer et al. 2006)	Variety of developmental problems	1500		5.6
(Sharp et al. 2006b)	MR	290	17q21.31,1q21.1,15q13, 15q24, 17q12	5.5
(Koolen et al. 2006)	MR, dysmorphic features	1200	17q21.31	0.3
(Koolen et al. 2008)	Development delay, dysmorphic features	22	17q21.31	
(Shaw-Smith et al. 2006)	Learning disability	3	17q21.31	*de novo*
(Ullmann et al. 2007)	Autism	182	16p13.1 (duplication)	
	MR	95	16p13.1 (deletion)	
(Sharp et al. 2008)	MR, epilepsy	2082	15q13.3 (deletion)	
(Brunetti-Pierri et al. 2008)	Autism, MR and/or congenital anomalies and dysmorphic features	16,557	1q21.1 (deletion/duplications)	
(Mefford et al. 2008)	MR, autism, congenital anomalies	5218	1q21.1 (deletion/duplications)	del-0.5 dup-0.2

literature review until 2008

The pathogenicity of these CNVs has been suggested on the basis of the following criteria: not observed in control populations, *de novo* events or phenotypic similarities between carriers which could eventually be classified as syndromes.

It has been observed that submicroscopic chromosomal abnormalities are a frequent cause for developing idiopathic mental retardation. It is thought that ~10% of cases with MR/intellectual disability carry a pathogenic copy number variant and this observation has been independently confirmed in a number of studies. Some specific variants have been found to be involved in liability to mental retardation. It is of note that the evidence for the involvement for some of the implicated variants comes from multiple studies and these independently replicate the original findings. Many different variants have been observed associated with the disorder, suggesting that brain dysfunction could be a consequence of disruption of many different genes. With respect to some of the observed variants (i.e. 16p13.1, 1q21.1), incomplete penetrance has been noted. This suggests that these variants are predisposing factors and not sufficient for developing of the disorder.

1.2.6.2. Autism

Autism spectrum disorders are neuropsychiatric disorders where genetic factors are known to play a role in the susceptibility. Cytogenetic evidence suggests that 3 to 5% of autism cases could be due to chromosomal abnormalities (Freitag 2007). In addition, several loci of potentially pathogenic CNVs have been also detected (Abrahams and Geschwind 2008).

Sebat et al. assessed the possible involvement of *de novo* events in developing autism, by screening 188 simplex families, 77 multiplex families and 196 control subjects. *De novo* events were found to be correlated with autism ($p = 0.0005$) and were present with a frequency of 10% in simplex families, 3% of patients in families with another affected first-degree family member, whereas amongst controls the frequency was 1% (Sebat et al. 2007). Similarly to this finding, Christian et al. observed that ~11% of the screened autism cases could be due to CNVs (Christian et al. 2008).

Some of the submicroscopic variants implicated in the causality of ASD are presented in Table 9.

Table 9 CNV studies in autism

Reference	Diagnosis	N cases	Locus	% of cases with the CNV
(Brunetti-Pierri et al. 2008)	Autism, MR and/or congenital abnormalities and dysmorphic features	16,557	1q21.1 (deletion/duplication)	Deletion-0.16 Duplication-0.1
(Mefford et al. 2008)	MR, autism or congenital abnormalities	5218		Deletion-0.5 Duplication-0.2
(Kim et al. 2008; Szatmari et al. 2007)	autism		*NRXN1*	
(Falk and Casas 2007)	MR, autism		2q37	
(Miller et al. 2009)	Autism, MR	2886	15q13.3	Deletion -0.17 Duplication-0.17
(Pagnamenta et al. 2008)	Autism			
(Christian et al. 2008)	Autism	372	15q11-15q13 (maternal duplication)	0.8
(Kumar et al. 2008; Marshall et al. 2008; Weiss et al. 2008)	Autism		16p11.2 (duplication/deletion)	1
(Ullmann et al. 2007)	Autism	182	16p13.1 (duplication)	1.6
	MR	95	16p13.1 (deletion)	2
(Durand et al. 2007)	Autism	324	22q13/*SHANK3*	0.9

literature review until 2008

A region, subject to genomic imprinting, at **15q11-13** has been observed to be associated with autism spectrum disorders (Hogart et al. 2008). It has been shown that duplications at this region account for ~1% of autism cases and that these duplications are predominantly of maternal origin (Freitag 2007).

The deletion at **16p11.2** locus has been investigated in a large sample consisting of a plethora of phenotypes including schizophrenia, bipolar disorder, attention deficit hyperactivity disorder, panic disorder, language disorder and a vast sample set of controls (n=18,834). The observed frequencies were: 1% in autism, 0.1% in the other psychiatric phenotypes and 0.01% in general population (Weiss et al. 2008). The reciprocal duplication has been detected in schizophrenia and bipolar

disorder cases (Walsh et al. 2008; Weiss et al. 2008). This locus will be discussed in more detail in the Discussion section (5.10.2, page 153) and in order to avoid reiteration, here only data, published until 2008 have been referred to.

Like 16p11.2 locus, copy number variants at **1q21.1** locus have been observed to be associated with autism, MR and schizophrenia, thus suggesting that it could potentially be important in all of these disorders. The manifestation of this CNV cannot be assigned to one phenotype, but could be considered in a broader context of development, as it has been found associated with a wide range of paediatric and developmental abnormalities. The deletion has been detected in some unaffected family members and based on this evidence, it has been proposed that CNV at this locus could produce subtle phenotype features, which could become evident after further clinical assessment (Mefford et al. 2008).

It has been shown that rare copy number variants are a significant factor for developing autism. Similar to MR, in autism spectrum disorder several variants have been replicated in a number of studies which provide evidence that these findings are robust. Furthermore, some of these variants have been implicated as predisposing factors in MR and schizophrenia suggesting that at least partially, similar underlying mechanisms may be relevant to these disorders. It could be speculated that the observed copy number variants act together with genes, other contributing CNVs and environmental factors, which could influence development of MR in some individuals, and autism or schizophrenia in others.

1.2.6.3. Schizophrenia

There is compelling evidence that CNVs play a role in the susceptibility to schizophrenia. Early evidence for structural abnormalities in schizophrenia comprises the following two classical examples: a translocation of the gene disrupted in schizophrenia (*DISC1*) and a deletion at the 22q11.2 locus. In addition, new evidence has started accumulating since the beginning of 2008 with a flurry of reports. Unlike previous experience with lack of replication of statistically significant findings with respect to association with SNPs, the findings from CNVs have shown strong independent replication across different populations.

In a large Scottish pedigree it was observed that schizophrenia and affective disorder cosegragate with a balanced reciprocal translocation between chromosomes 1 and 11 (t(1;11)(q42.1;q14.3)) (Blackwood et al. 2001; St Clair et al. 1990). In the family, the performed linkage analysis showed highly statistically significant association with psychotic illness- LOD score of > 7. Two genes have been identified at this locus- *DISC1* and *DISC2*, and interestingly, both are brain-expressed (Millar et al. 2000). *DISC1* gene has been prioritised as the more probable candidate gene of the two as it is directly disrupted by the translocation. Since discovery, the DISC locus has been studied extensively. The findings have been independently replicated in a number of genetic linkage and association studies (reviewed in (Chubb et al. 2008)). The locus has been implicated to be associated with a number of phenotypes- schizophrenia, bipolar disorder, unipolar and schizoaffective disorder.

The other locus, a deletion at **22q11.2**, has also been studied extensively for association with neuropsychiatric conditions as it confers a substantial risk for developing schizophrenia and is the underlying cause for velo-cardio-facial syndrome (VCFS) (Bassett and Chow 2008). A deletion at 22q11.2 causes DiGeorge (MIM #188400)/velocardiofacial syndrome (VCFS) (MIM#192430). It is commonly referred to as 22q11.2 deletion syndrome (22q11.2 DS). The 22q11.2 deletion syndrome is the most common microdeletion syndrome occurring in ~1 of every 4000 live births (Oskarsdottir et al. 2004). 22q11.2 DS is characterised with developmental delay and wide range of cognitive and neurological deficits, including speech, language, memory and attention (Bearden et al. 2001). Interestingly, it has been found that deletions at this locus are one of the highest known genetic factors for

development of schizophrenia (Murphy and Owen 2001). This suggests that there could be a possible link between genes in this region and susceptibility to schizophrenia. Ivanov et al. estimated the frequency of the deletion to be ~0.6% in schizophrenia cases from early studies (Ivanov et al. 2003). More recent data based on Stefansson et al. and the International Schizophrenia Consortium studies provided evidence for a lower rate of 0.2% to 0.4% (International Schizophrenia Consortium 2008; Stefansson et al. 2008). Nevertheless, it has been shown that the deletion carries the highest odds ratios (OR) for development of psychosis. Murphy et al. found a rate of ~ 30% of adult carriers developing a psychotic illness (Murphy et al. 1999). It is of note, that the deletion has not been observed in healthy controls from the general population even though the number of the screened control individuals has been extremely large especially in the Stefansson et al. study (n=39,299 healthy controls) (Stefansson et al. 2008).

Since, with the help of the development in the methods for detection, evidence for the involvement in schizophrenia of microdeletions and microduplications has become apparent.

Kirov et al. presented one of the first studies of copy number variation in schizophrenia. In the screened 93 cases, 13 CNVs not present in controls were detected. The main findings have been a deletion disrupting *NRXN1* gene (neurexin 1) at locus **2p16.3** (present in two affected siblings) and a *de novo* duplication affecting *APBA2* gene (amyloid beta (A4) precursor protein-binding, family A, member 2) at **15q13.1** locus (Kirov et al. 2008). Deletions in *NRXN1* have been also detected in autism cases (Szatmari et al. 2007). In addition, a duplication affecting *APBA2* has been reported in a patient with autism (Christian et al. 2008).

Schizophrenia is characterised by reduced fecundity and yet the lifetime risk has been stable at around 1% . Although persistence in the population at this rate is explicable by a polygenic model comprising many weak effects (International Schizophrenia Consortium et al. 2009), it is also reasonable to postulate that variants of large effect may be replenished by new mutations implicating *de novo* CNVs in the susceptibility. It has been suggested that a new mutation in an affected individual could be potentially relevant to the pathogenesis of schizophrenia. As a reduction in fecundity has been observed (that is also the case with other neurodevelopmental disorders), due to negative selection, variants of large effect will be found with low

frequencies in the population. Such rare events could arise in multiple unrelated individuals by *de novo* occurrence (Stefansson et al. 2008). Therefore, CNVs with large effect could be enriched in such *de novo* events. To evaluate the potential involvement of rare *de novo* CNVs in schizophrenia, parent-offspring transmissions have been studied (Xu et al. 2008). This team observed that *de novo* CNVs contribute to susceptibility to schizophrenia ($p = 0.00078$). CNVs were observed eight times more frequently in sporadic cases (frequency of 10%) when compared with familial cases and control population (1.3%) (Xu et al. 2008). This provided evidence of the involvement of rare events of large effect size in the liability to schizophrenia.

Furthermore, Walsh et al. observed that rare CNVs were more frequent in cases (15% in cases with adult onset), when compared with healthy controls (5%) ($p = 0.0008$). This effect was more pronounced for cases with childhood onset- 20% of them were found to harbour a CNV ($p = 0.0001$) (Walsh et al. 2008). The overrepresentation was replicated in a sample of 83 trios, where a 2.6-fold enrichment of rare structural variants was observed in cases compared with the untransmitted parental chromosomes ($p = 0.03$). Walsh et al. also found that genes affected by the CNVs were statistically significantly overrepresented in neurodevelopment pathways including neuregulin and glutamate pathways (Walsh et al. 2008).

In 2008, two seminal papers investigating the possible involvement of copy number variation in schizophrenia were published in Nature- the International Schizophrenia Consortium and the Stefansson et al. studies (International Schizophrenia Consortium 2008; Stefansson et al. 2008).

Stefansson et al. identified 66 *de novo* CNVs in a population-based sample by analysing 9878 transmissions from parents to offspring. Three CNVs exhibited an association with schizophrenia in a subsequent study of patients with schizophrenia and controls- 1q21.1, 15q11.2 and 15q13.3. These variants were rare with high OR of 14.83 ($95\%CI: 3.55 - 60.40$; $p = 2.9 \times 10^{-5}$), 2.73 ($95\%CI: 1.50 - 4.89$; $p = 6 \times 10^{-4}$) and 11.54 ($95\%CI: 2.53 - 49.58$; $p = 5.3 \times 10^{-4}$) respectively for the three deletions. The 22q11.2 deletion was observed with a frequency of 0.2% in the cases and 0% in the controls (3838 cases and 39,299 controls) ($p = 4.2 \times 10^{-5}$, $OR = \infty$) (Stefansson et al. 2008).

Noteworthy correspondence was observed between the Stefansson et al. study and the International Schizophrenia Consortium (ISC) study (International Schizophrenia Consortium 2008; Stefansson et al. 2008). ISC analysed large sample set of 3391 patients with schizophrenia and 3181 healthy controls, using high-density microarrays. Similar to the Stefansson et al. study, rare deletions at 22q11.2, 1q21.1 and 15q13.3 were detected in cases with high OR (21.6, 6.6 and 17.9 respectively). In addition, when CNVs with frequency of $< 1\%$ and with size > 100 kb were analysed, it was found that schizophrenia cases have 1.15 higher rate of CNVs when compared with controls. This effect was more pronounced for rarer, single-occurrence CNVs and CNVs affecting genes in patients (International Schizophrenia Consortium 2008). Our department was involved in this study, and the findings in schizophrenia prompted my research interest in the area, which subsequently led to the design of this PhD study.

In conclusion, an increased frequency of rare CNVs in schizophrenia has been observed, suggesting that such structural variants could be important for the liability to the disorder. In addition, some specific CNVs have been found to be involved in susceptibility. It is of note that some of the observed CNV have been independently supported in several studies. Moreover, it has been shown that all of the associated structural variants confer a substantial risk for developing schizophrenia. Some of the implicated variants, viz. 1q21.1, *NRXN1* deletion and 15q13.3 have also been observed to increase the risk for developing other neurodevelopmental disorders, such as mental retardation, autism spectrum disorder and/or variable paediatric phenotypes. Similarly, deletions at 22q11.2 have been implicated in other phenotypes- attention-deficit hyperactivity disorder (ADHD) and bipolar disorder.

The findings of these specific loci account only for a fraction of the schizophrenia cases. There is an increased burden of CNVs in schizophrenia, although these specific loci do not provide explanation for most of this specific burden. Therefore, it could be speculated that other loci remain to be found that could explain the rest of the increased burden. Discovering these variants could be of great benefit, as these CNVs could point to identification of additional risk variants in genes and pathways involved in schizophrenia.

Since these early studies on the whole-genome CNV structure in schizophrenia, a lot more research has been performed. The more recent findings will be considered in the Discussion section (page 153) in order to avoid repetition.

1.2.6.4. Bipolar disorder

Unlike schizophrenia, the copy number variation has been largely unexplored in the susceptibility to bipolar disorder. There had been only one systematic study of CNVs in bipolar disorder, published when the work towards this PhD was underway. The involvement of rare and large CNVs (> 100 kb) in developing BD was assessed in 1001 cases and 1034 controls (Zhang et al. 2008). The copy number variation was inferred using data from Affymetrix 6.0 arrays. No increased burden was observed in cases when compared with controls. However, single-occurrence deletions were found more frequently in cases (16.2%) than in controls (12.3%), with a corresponding significance level of $p = 0.007$. The number of genes affected by singleton deletions did not show a statistically significant difference when the cases were compared with controls ($p = 0.21$). No statistically significant difference was noted between cases and controls with respect to singleton duplications ($p = 0.054$). Nevertheless, when only cases with early disease onset (< 18 years) were investigated, the effect of singleton deletions was found to be more pronounced. A comparison of early onset versus late onset cases revealed a marginally higher frequency of singleton deletions in early onset cases ($p = 0.05$). In addition, Zhang et al. performed pathway analysis which revealed that genes affected by CNVs were overrepresented in pathways involved in psychological disorders and learning behaviours ($p = 6.30 \times 10^{-6}$ and $p = 8.29 \times 10^{-3}$ respectively).

Based on these findings, it was concluded that developing bipolar disorder could be a result of joint function of multiple rare structural variants (Zhang et al. 2008).

Prior to the publication of the above-mentioned report, studies investigating the possible involvement of copy number structural variation in bipolar disorder were anecdotal. The majority of studies analysed limited number of samples.

Namely the studies providing data, based on analysing copy number variation in bipolar disorder were the following:

A balanced chromosomal translocation t(1:11) (q43,q21) affecting the ***DISC1*** gene observed in a Scottish pedigree, segregated with disease. In this family, both schizophrenia and mood disorders have been detected (St Clair et al. 1990).

A duplication at 3q13.3 affecting ***GSK3β*** gene (glycogen synthase kinase 3 beta) and two other genes has been found associated with bipolar disorder (Lachman et al. 2007).

The **22q11.2** locus has also been associated with developing bipolar disorder. This locus has been extensively studied in the neuropsychiatric field as it has been observed that schizophrenia and bipolar disorder are present in some of the affected individuals with VCFS, amongst other symptoms (facial, immunological and cardiovascular) (Carlson et al. 1997; Murphy et al. 1999; Papolos et al. 1996). A more detailed description of the data with respect to 22q11.2 was already presented in section 1.2.6.3 (page 49).

Wilson et al. studied brains from bipolar disorder cases, schizophrenia cases and controls for copy number variation. 1p34.3, 14q23.3 and 22q12.3 affecting the *GLUR7*, *AKAP5* and *CACNG2* genes were found to be associated with disease (Wilson et al. 2006).

In conclusion, it has been shown that copy number variation accounts for some of the susceptibility to neuropsychiatric conditions, such as mental retardation, autism and schizophrenia. The amount of accumulated evidence with respect to some specific CNVs has been large and even more, some of the evidence has been replicated in independent studies.

In addition, previously it has been shown that bipolar disorder and schizophrenia may share some of the underlying genetic factors for the predisposition to disease. Thus, to ask the question of the possible involvement of copy number variation in liability to bipolar disorder is in order. The knowledge with respect to CNVs in the predisposition to developing bipolar disorder has been very limited, with the predominant number of studies focusing on specific candidate genes or has been based on a limited number of samples. There was only one study based on whole-genome CNV investigation in a large sample of bipolar cases and controls at the time when the work towards this PhD was already underway. Therefore, further studies were warranted to elucidate if structural variation is a predisposing factor for the susceptibility to bipolar disorder.

1.3. Summary

Bipolar disorder is a severe neuropsychiatric condition affecting ~1% of the human population worldwide. A strong genetic component has been implicated in its causality. Thus far, a number of genetic variants have been shown to increase the susceptibility (i.e. variants in the *ANK3*, *CACNA1, DGKH, MYO5B* and other genes), although conclusive evidence has proved elusive. Recent developments in the neuropsychiatric field have found support for the possible involvement of rare structural variants in human disorders including autism spectrum disorders, mental retardation and schizophrenia. At the start of this PhD work, the possible involvement of copy number variation in the causality of bipolar disorder was understudied and largely unknown. This stimulated my interest in this area and led to the work described in this thesis.

2. Aims and objectives

The primary aim of this PhD research project was to investigate the possible involvement of copy number variation in the susceptibility to bipolar disorder.

In order to achieve this, I aimed to:

1. Analyse the CNV structure in a large cohort of bipolar cases (n=1868) and healthy controls (n=2938) using SNP genotyping data from Affymetrix 500K arrays. In addition, I wanted to compare bipolar cases with a large number of non-psychiatric controls (~10,000), with respect to their copy number variation structure;
2. Determine whether large (> 100 kb) and rare (frequency- < 1%) copy number variants were associated with susceptibility to bipolar disorder;
3. Determine whether some specific deletions or duplications were involved in developing bipolar disorder;
4. Compare bipolar disorder and schizophrenia with respect to their CNV burden and the possible involvement of some specific deletions/duplications.

3. Material and Methods

In this chapter I will provide a description of the methodology used in the PhD work. Here, I will therefore cover: case and control sample description, laboratory methods, CNV calling procedure, bioinformatical and statistical methodology.

Unless otherwise stated, all work and analyses have been performed by me.

3.1. DNA samples

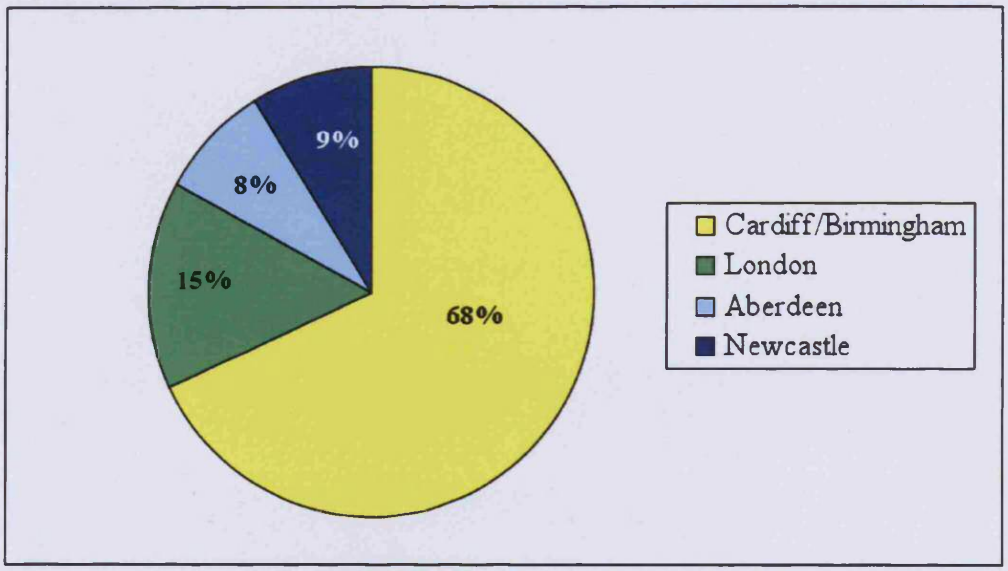
3.1.1. Cases

3.1.1.1. Bipolar disorder

The collection of the bipolar samples involved in this work, including the diagnosis of patients, was performed by Professor Nick Craddock, Dr George Kirov, Dr Ian Jones and other members of the Department of Psychological Medicine, Cardiff University and the Department of Psychiatry, Birmingham University. The total number of the analysed bipolar cases was 1868. The majority of the sample was recruited by the Department of Psychological Medicine, Cardiff University and the Department of Psychiatry, Birmingham University as shown in Figure 10.

The recruitment of the rest of the sample was undertaken throughout the UK by teams based in Aberdeen (Institute of Medical Sciences, University of Aberdeen), London (the Institute of Psychiatry, King's College London) and Newcastle (Neurobiology and Psychiatry, Royal Victoria Infirmary).

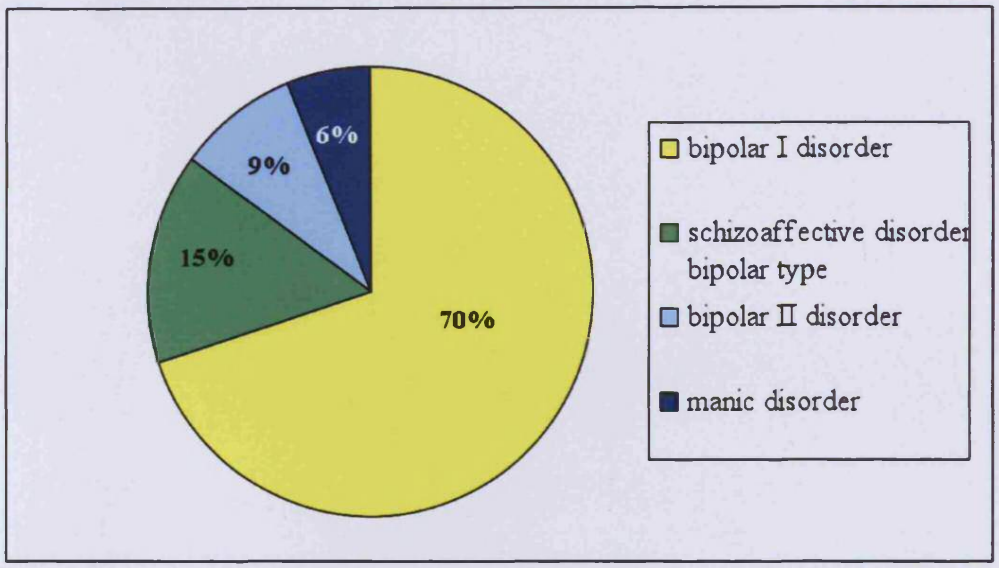
Figure 10 Proportion of the sample set recruited in the four different sites throughout the UK



All participants were Caucasian and of UK origin. 37% of the sample were male cases; mean age 47 (SD 13) years and mean age of onset 26 (SD 11) years. All participants have provided written informed consent to take part in genetic studies. Protocols and procedures were approved by relevant ethical review panels.

Individuals who had been in contact with mental health services were recruited if they suffered with major mood disorder characterised with significant episodes of elevated mood. This was defined as a lifetime diagnosis of bipolar mood disorder according to Research Diagnostic Criteria (Spitzer et al. 1978) and included the bipolar subtypes that have been previously shown in family studies to co-segregate (Rice et al. 1987): bipolar I disorder, schizoaffective disorder bipolar type, bipolar II disorder and manic disorder (Figure 11).

Figure 11 Proportion of the sample diagnosed with disorders from the bipolar spectrum



Lifetime best-estimate method was used to make the diagnoses (Leckman et al. 1982) by a trained psychologist or psychiatrist. All diagnoses were confirmed by a consensus of two people. The diagnoses were based on data from semi-structured interview- Schedules for Clinical Assessment in Neuropsychiatry (SCAN) (Wing et al. 1990), information obtained from psychiatric case records and an OPCRIT checklist (McGuffin et al. 1991). The Bipolar Affective Disorder Dimension Scale (BADDS) was used to rate key clinical variables relating to psychosis (Craddock et al. 2004).

The 1868 bipolar cases were exactly the same cases that took part in a large collaborative project to perform genome-wide single nucleotide polymorphism association study of 14,000 cases of seven common diseases and 3000 shared controls (that is the Wellcome Trust case control consortium study, WTCCC). The studied common diseases comprised: bipolar disorder, Crohn's disease, rheumatoid arthritis, type 1 diabetes, type 2 diabetes, hypertension and coronary artery disease (Wellcome Trust Case Control Consortium 2007).

3.1.1.2. Schizophrenia

Along with the bipolar cases, 520 unrelated schizophrenia/ schizoaffective disorder cases were also studied in this PhD work. All were white and born in the UK. The mean age at first psychiatric contact was 23.8 (SD 7.9) years and the mean at ascertainment was 44.8 (SD 13.1) years. Of those 520, 471 survived copy number detection filtering and 440 of them were further analysed for this PhD. The rest were excluded as they had diagnoses of schizoaffective disorder.

All patients were interviewed by trained psychiatrists or psychologists using Present State Examination (Wing 1974) or SCAN interview (Wing et al. 1990).

Based on the interview information, along with review of case notes, consensus diagnoses of schizophrenia were made according to DSM-IV criteria (American Psychiatric Association 1994), by two independent raters. Cases were screened to exclude substance-induced psychotic disorder or psychosis due to a general medical condition. Ethics committee approval was obtained from all regions where patients were recruited.

These schizophrenia cases were genotyped in the same pipeline as the WTCCC study and were processed at the same time and with the same methods along with the bipolar cases and the controls (O'Donovan et al. 2008).

3.1.2. Controls

Control subjects were derived from two sources: half from 1958 Birth Cohort and the remainder from a new UK Blood Service sample (National blood service, NBS). The DNA preparation, DNA extraction, quality control and normalization of the control DNAs were performed at the Wellcome Trust Laboratories at Sanger, Cambridge. These controls were part of the WTCCC study (Wellcome Trust Case Control Consortium 2007).

3.1.2.1. 1958 Birth Cohort

The 1958 Birth Cohort (also known as the National Child Development Study) comprised all births in Scotland, England and Wales during one week in March, 1958. From the original sample of ~ 17,500 births, the survivors were followed up at 7, 11, 16, 23, 33 and 42 years of age (http://www.cls.ioe.ac.uk/studies.asp?section=000100020003) (Power and Elliott 2006). At the age of 44-45 years, biomedical examination along with blood sample collection were carried out (http://www.b58cgene.sgul.ac.uk/collection.php) (Strachan et al. 2007). From the blood samples of the consenting 7692 individuals, Epstein-Barr virus-transformed cell lines were prepared.

For the purposes of the WTCCC study, DNA samples were extracted from 1504 cell-lines of self-reported white ethnicity and representative of gender and each geographical region in agreement with the gender and geographical region distribution of the 14,000 cases that took part in the WTCCC study (Wellcome Trust Case Control Consortium 2007).

3.1.2.2. National blood service (NBS)

WTCCC in collaboration with the UK Blood Services set up a national repository of anonymized samples of DNA from 3622 consenting blood donors, aged 18-69 years. A set of 1500 samples was selected from the total number. The selection was based on the sex and geographical region in order to reproduce the distribution of the samples in the 1958 Birth Cohort control group. The subjects were of self-reported European Caucasian ancestry. The DNA was extracted from blood (Wellcome Trust Case Control Consortium 2007).

3.2. DNA preparation of the bipolar DNA samples

The DNA preparation involved DNA extraction, assessment of the quality of the extracted DNA samples and DNA preparation for the WTCCC study.

3.2.1. DNA extraction

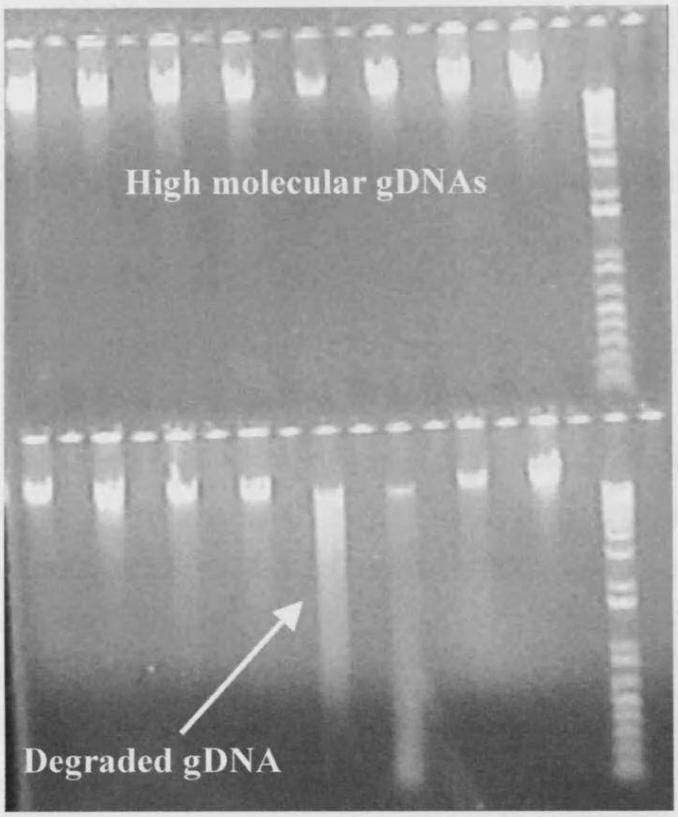
From the 1868 analysed bipolar cases, ~ 23% were extracted by me. The rest were extracted by members of staff from the Department of Psychological Medicine, the Department of Psychiatry, Birmingham University or other centres in UK who took part in the bipolar WTCCC study (Institute of Medical Sciences, University of Aberdeen and the Institute of Psychiatry, London).

High molecular weight genomic DNA (gDNA) was obtained from lymphocytes from venous whole blood. Each sample was prepared via standard phenol-chloroform DNA extraction followed by ethanol precipitation. The DNA was subsequently diluted in TE buffer. The short term storage of the DNA samples was at $4^0 C$ and at $-20^0 C$ for a long period of time.

3.2.2. DNA quality control using degradation test

A proportion of the extracted DNA samples was tested to determine if DNA degradation, smearing or low genomic product were present. DNA samples were run on a gel, able to detect samples with high molecular mass (AquaPor™ with DNA resolution between 1 and 50 kb). The gel was stained with ethidium bromide and DNA fragments with bound ethidium bromide were visualised using ultraviolet light. Since the DNA samples predominantly consist of fragments > 20kb in length, the DNA forms one slow moving band on the gel. Any smearing is indicative of DNA with low molecular weight suggesting that DNA has degraded. An example of high molecular DNA samples and degraded DNA samples is presented in Figure 12.

Figure 12 Example of a degradation test



3.2.3. DNA PCR check

A proportion of the extracted DNAs was tested using a known, previously optimised PCR and the products were then visualised (using ethidium bromide to stain the PCR product and subsequent exposure to ultraviolet light) on a gel to examine if the DNA samples are in a viable condition.

3.2.4. DNA quantification

The PicoGreen® dsDNA Quantitation Reagent was used for DNA quantification. The PicoGreen dye is highly sensitive fluorescent nucleic acid stain for quantifying double-stranded DNA (dsDNA). It possesses strong affinity for binding to dsDNA and it undergoes fluorescent excitement upon binding to it. The fluorescence is measured using a microplate fluorometer.

The assay was performed as outlined in a standard operating procedure used in the department. The DNA samples were diluted 200 times and pipetted in 96 well plates. Subsequently, the fluorescent intensity of the PicoGreen dye incorporated in the dsDNA was measured at excitation wavelength of 485 *nm* and recording the emission wavelength at 535 *nm*. A standard curve prepared by me was then used to calculate the concentration of DNA for each sample.

3.2.5. Preparation of case DNA samples for WTCCC

Unique sample identifiers were issued by WTCCC for all the samples that were sent for genotyping. In an Excel spreadsheet, provided by WTCCC, the following information was logged: unique WTCCC identifier, our laboratory sample ID, position in the shallow 96-well box, barcode of the box, case/control status, DNA concentration, DNA extraction method, DNA source, gender and geographical region. The WTCCC project included sending DNA samples to the Wellcome Trust Laboratories at Sanger Institute, Cambridge in two waves in 2005 and 2006.

In 2005, 1229 DNA samples were sent, and in 2006- another 1055 samples. From these samples, 1868 were selected for genotyping by WTCCC and were further analysed. Half of the samples sent in the first wave were prepared by Mrs. Rachel Raybould and the other half by me. All the DNA preparation for the second wave in 2006, was done by me.

The DNA preparation involved measuring all DNA samples with PicoGreen® quantitation assay, then selecting the DNA samples which had sufficient DNA concentration (DNA concentration $> 70ng/\mu l$) and sufficient DNA quantity ($>100\mu l$). The selected DNA samples were diluted or concentrated to $\sim 100ng/\mu l$. The preparation of the DNA samples also involved performing a degradation test for a proportion of the samples.

After adjusting the concentration of the samples to $\sim 100ng/\mu l$, they were aliquoted into 96-wells Abgene® plates and barcoded accordingly to the file that was prepared with the required information by the Wellcome Trust Laboratories.

Many people from the Department of Psychological Medicine were involved in the recruitment, the selection process of the cases and file preparation of the required information for the project: Prof. Nick Craddock, Dr. George Kirov, Dr.

Elaine Green, Dr. Liz Forty, Dr. Elen Russell, Mr. John Tredget, Mrs. Christine Fraser, Mrs. Rachel Raybould, and myself. At the Department of Psychiatry, Birmingham University, Mrs. Sian Caesar and Dr. Katherine Gordon-Smith, were also involved in the project. As the WTCCC project involved collaborators from the Institute of Medical Sciences (University of Aberdeen), the Institute of Psychiatry (King's College London) and the department of Neurobiology and Psychiatry (Royal Victoria Infirmary, Newcastle), I coordinated the DNA samples and case information receipt at our department. Subsequently, the DNA samples received from our collaborators throughout the UK were prepared by Mrs. Rachel Raybould or me, in exactly the same manner as our samples: normalised to the required DNA concentration, aliquoted in 96-wells Abgene® plates, barcoded and the file with the required information was prepared.

In addition, in exactly the same manner the schizophrenia cases were prepared by Dr. Hywel Williams, Mrs. Sarah Dwyer, Dr. Liam Carroll, Dr. Nadine Norton, Dr. Lyudmila Georgieva and Dr. Tim Peirce.

3.3. Wellcome Trust Case Control Consortium quality control measures and selection of samples for genotyping

The quality of all case and control DNA samples was assessed at the Wellcome Trust Laboratories, Sanger Institute, Cambridge (Wellcome Trust Case Control Consortium 2007).

On receipt at the Sanger Institute, the samples had DNA concentration measured by PicoGreen®, then were checked for DNA degradation on a 0.75% agarose gel, and subsequently genotyped with up to 38 SNPs arranged in two multiplex reactions using the MassExtend® (hME) and/or iPLEX® assay Sequenom® assays. The above SNPs were used to obtain concordance rate between Sequenom® and Affymetrix® assays (25 of the 38 SNPs were present on the Affymetrix GeneChip® 500K arrays) and experimentally confirming the gender of each sample. Samples with concentrations $\geq 50ng/\mu l$, showing limited or no degradation, having minimum of seven SNPs successfully genotyped out of the 10-plex (hME® reaction) and/or 14 SNPs successfully genotyped out of the 23-plex (iPLEX® reaction) SNPs, and having the sex markers in agreement with the supplied information from us, were

deemed fit for whole-genome genotyping. The SNPs in the hME multiplex assay were replaced with a second iPLEX® reaction in the course of the project to increase marker density. 1868 samples were selected from the bipolar cases collection and ~1500 samples from the two control collections- 1958 cohort and NBS. The selected samples were normalized to $50ng/\mu l$ and re-arrayed robotically into 96-well plates, so that each plate comprised 94 samples representing at least two different collections (that is, samples from the seven separate disorders that took part in WTCCC) at a ratio of 1:1.

3.4. Genotyping with Affymetrix platform

DNA from all cases (bipolar and schizophrenia) and controls was genotyped with the Affymetrix GeneChip® 500K Mapping Array Set at the Affymetrix Services Lab in California. The GeneChip® 500K Array Set comprises two arrays: NspI and StyI. NspI array comprises 262,000 SNPs and StyI -238,000 SNPs respectively. The SNPs chosen by Affymetrix to be present on the arrays, have been selected, based on genotyping accuracy, call rate and linkage disequilibrium analysis across the genome. The SNPs present on the 500K Mapping array, have been selected in a way that 85% of the human genome is within 10 kb of a SNP (http://www.affymetrix.com).

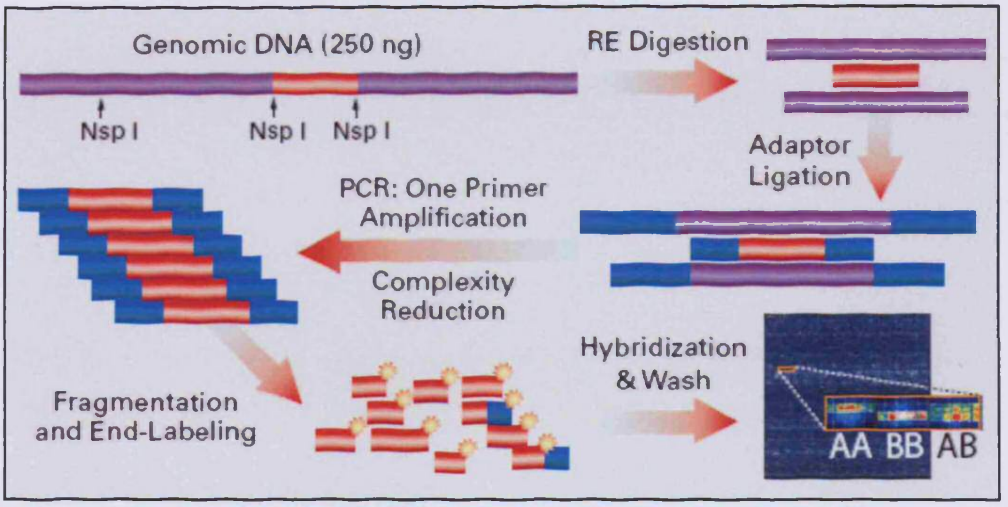
The genotyping procedure is as follows (Figure 13, page 67): two aliquots of $250ng$ of genomic DNA are digested with restriction NspI and StyI enzymes. Subsequently, an adaptor is ligated. The adaptor recognises the "sticky" four base-pair ends that are produced after the digestion with the enzymes. A generic primer that recognises the adaptor sequence is used to amplify the adaptor-ligated fragments. Preferentially only fragments with size range between 200 and 1100 bp are amplified due to the selected PCR conditions. The amplified DNA is fragmented, labelled and hybridised to the corresponding GeneChip® 250K array (NspI or StyI). Each array in the GeneChip® Mapping set, comprises more than 6.5 million features, each consisting of more than one million copies of a 25-bp oligonucleotide probe of a defined sequence. Each SNP is interrogated by 6- or 10-probe quartets, where each probe quartet comprises perfect match and a mismatch probe for each allele. Each enzyme preparation is hybridised to the corresponding SNP array- NspI or StyI.

Up to the hybridisation step, samples were processed in batches of 96-well plate format. The processing of the samples in batches proved to have a profound

effect on the copy number variation identification and is further discussed in detail in section 4.1.3 (page 92).

Each 96-well plate carried a positive and negative control. CEL files contained the row hybridisation signal intensity data of the various probes on each array.

Figure 13 Affymetrix GeneChip® 500K Mapping overview



(http://www.affymetrix.com)

3.5. Quality control of Affymetrix genotyping

The quality control of the genotyping with the Affymetrix platform was performed by me. The bipolar case DNA samples were genotyped for 46 markers in two separate panels, using the iPLEX® Sequenom MassARRAY platform (http://www.sequenom.com). The first panel comprised markers that were genotyped by Sanger institute for quality control of the DNA samples. In addition, these markers were also present on the Affymetrix GeneChip 500K array. The second iPLEX panel comprised markers that were present on the Affymetrix 500K platform.

The Sequenom MassARRAY genotyping system allows highly accurate genotyping of SNPs by combining iPLEX® GOLD primer extension chemistry with MALDI-ToF® (Matrix Assisted Laser Desorption Ionisation-Time of Flight) Mass Spectrometry (http://www.sequenom.com). IPLEX® GOLD involves single nucleotide primer extension over the polymorphism of interest and the examination of

the mass of the extended product to discern the genotype of the sample. Results are stored and analysed using the Typer Analyzer software. The main advantage of this genotyping system is the high accuracy combined with high multiplexing level (up to a 40-plex) in a 384 microlitre plate.

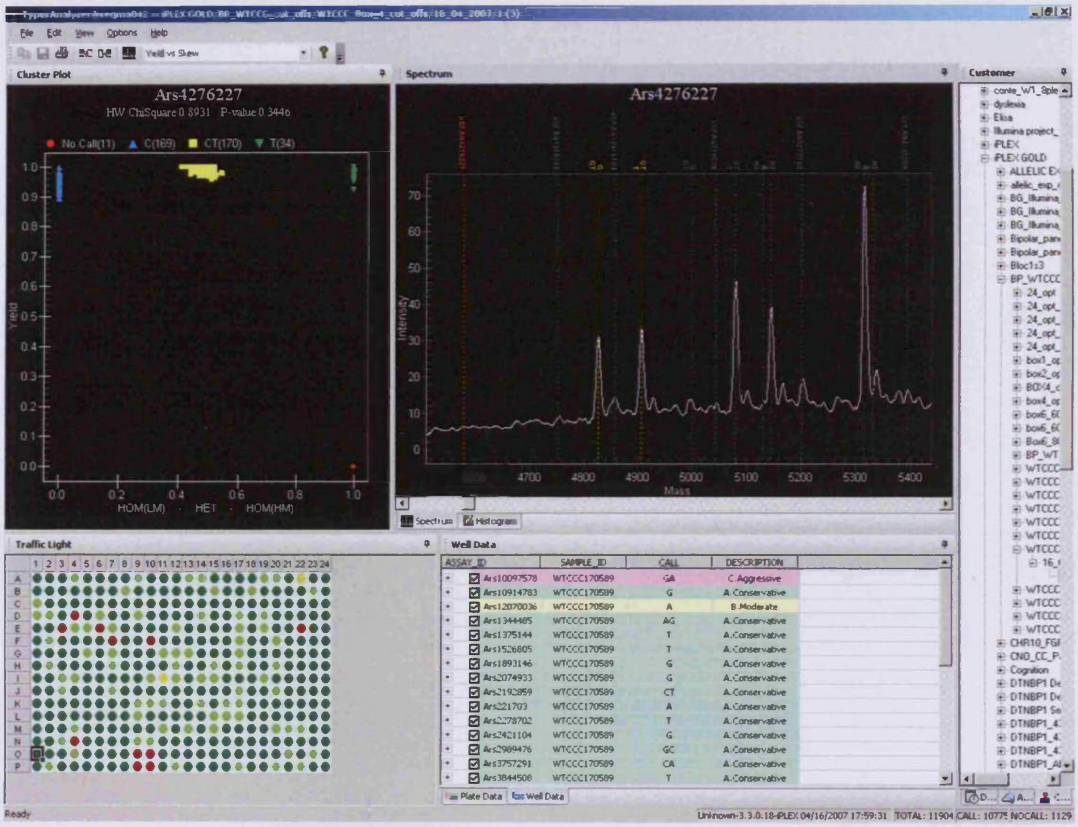
It involves an initial PCR of all the SNPs in a 384 sample format. Should the 1st step pass an initial quality control for contamination and PCR efficiency, the PCR reaction is cleaned up in order for the unincorporated dNTPs, primers, DNA polymerase and salts to be removed. The clean product is then subject to a primer extension reaction. It involves the addition of optimised concentrations of unextended primers, along with dideoxy nucleoside triphosphates.

After the extension reaction, desalting the solution with special resin, provided by Sequenom, is performed, in order to remove all ions that may alter the spectra of the sample and affect the subsequent analysis.

Samples are automatically spotted onto the Sequenom MassARRAY SpectroCHIP using a nanodispenser liquid handler. Each chip comprises 384 spots composed by a combustible matrix (3-hydroxypicolinic acid), which allows ioinisation of the product when excited by a laser (Buetow et al. 2001).

Each ioinised, extended and unextended MassEXTEND primer product differs in mass and is therefore amenable to MALDI-ToF mass spectrometry analysis using MassARRAY RT software (SpectroAcquire, Sequenom). The software estimates genotypes for each sample, based upon certain parameters such as peak height (intensity of mass signal) of each allele and also extension primer yield (successful extension of the unextended primer compared to the residual unextended primer). These genotypes can be then viewed and manually revised by the user using the Typer Analyzer software (Figure 14).

Figure 14 Screenshot from Typer Analyzer software for one of the tested SNPs (rs4276227)



3.6. Copy number variation detection

3.6.1. Affymetrix Genotyping console

Copy number variation was inferred from the raw intensity genotyping data (CEL files). The computer used for the analysis of the CNV data had the following specifications:

Operating System: Microsoft Windows XP Professional

Processor: Intel ® Core ™ 2 Quad, 2.4 Giga Hertz

Random Access Memory: 4 Giga Bytes

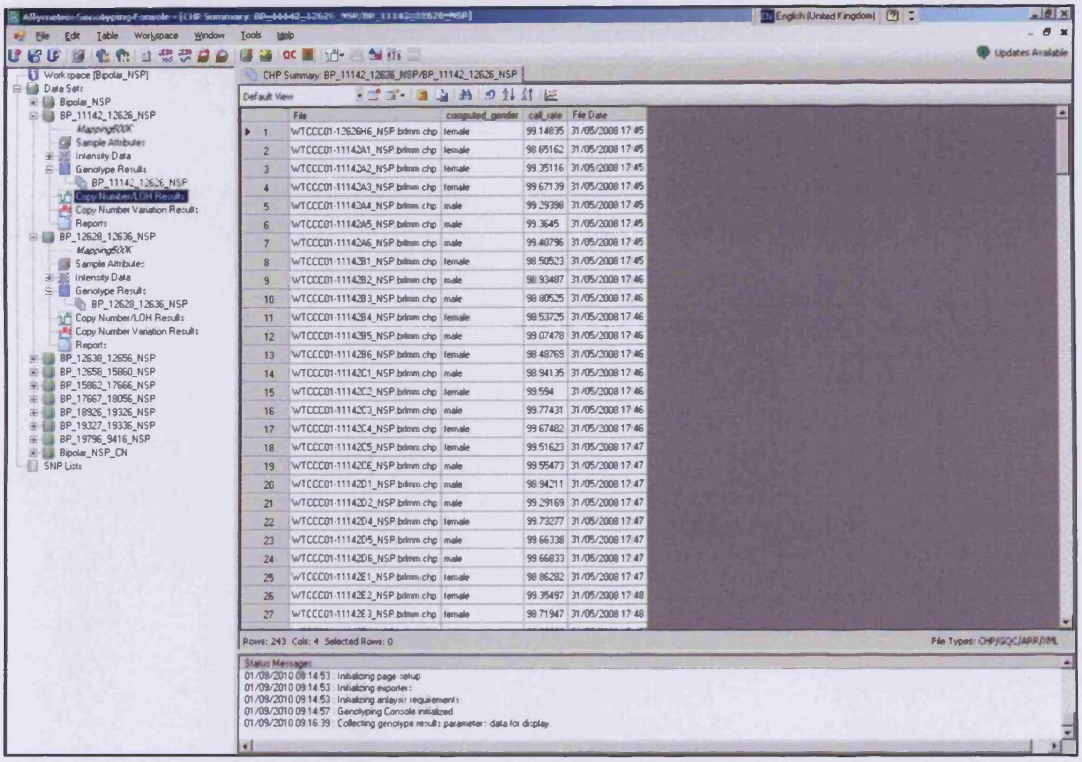
Hard-Disk Drive: 500 Giga Bytes

Due to the large size of the raw intensity files (CEL), the amount of studied individuals and the size of some of the files generated by the Affymetrix Genotyping

console, an external hard-drive was also required (Western Digital, dual drive storage system, 2 TB drive).

The CNV calling was carried out using stand-alone Windows-based application provided by Affymetrix, namely- Affymetrix® Genotyping console 2.1 (Figure 15). The software was downloaded from the Affymetrix website (http://affymetrix.com).

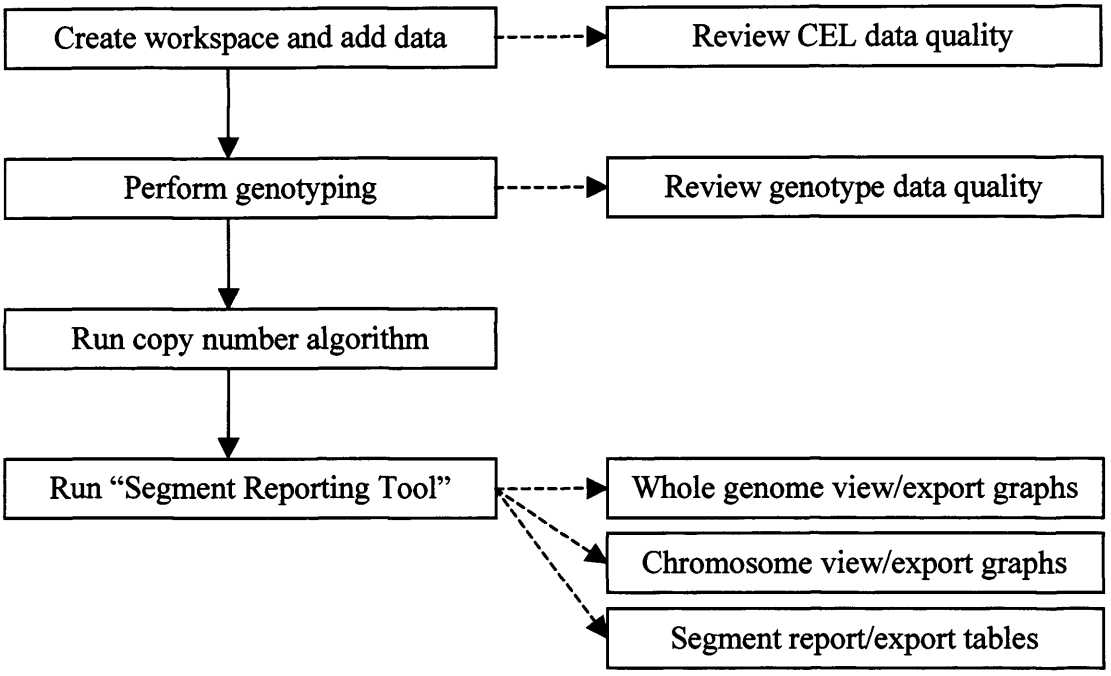
Figure 15 Affymetrix® Genotyping console 2.1



The software required intensity data files (CEL) for each individual. These data files have been generated by other Affymetrix software, when a microarray has been scanned. The CEL files comprise information about the intensity values of the individual probes present on the array. These files were provided by WTCCC. As the Affymetrix GeneChip® array set comprises two arrays- NspI and StyI, the CEL intensity files from one individual for the two arrays were analysed separately.

The workflow of the separate steps involved in the CNV detection analysis using the Affymetrix genotyping console is presented in Figure 16, page 71.

Figure 16 Workflow of the Affymetrix genotyping console for CNV detection



The steps involved in inferring the CNV structure of the case and control samples using the Affymetrix genotyping console are presented in more detail below:

1. Quality control (QC) of the CEL intensity data- the default QC metrics of the program were used with a QC SNP call rate threshold $\geq 93\%$. This QC measure provides an estimate of the overall quality of the samples. Samples with an average SNP call rate $< 93\%$ were excluded from further analysis. Gender analysis was also performed at this stage. It provided a gender call, which was used at a later step, when individuals were separated into two groups according to gender for the CNV calling step (as males and females were analysed separately). The QC step was performed for all individuals (i.e. all the cases for NspI or StyI arrays at once). It has been shown that removing poor quality CEL files at this stage could improve the quality of the genotypes of the remaining CEL files (Affymetrix Inc 2007);

2. Genotyping- at this stage, genotyping analysis of the intensity data that passed the initial quality control step was performed. The genotype calls (i.e. AA, AB, BB) were generated using a method that has been developed by Affymetrix (that is, Bayesian Robust Linear Model with Mahalanobis distance classifier algorithm) (http://www.affymetrix.com). Due to computational power limitation, the genotyping step was performed for groups of < 250 individuals at a given time;

3. CNV analysis- the input for performing this step was the CEL and the genotyping files. Males and females were analysed separately. In addition, a set of reference samples was created at this stage. At this point of the analysis, a major batch effect in the data was noticed. A batch is a set of up to 96 samples processed at the same time in the laboratory. That is, the samples have been run in the Affymetrix service lab using the same reagents and as all of the samples have not been run at once, there is a possibility that systematic differences between the runs occur due to intrinsic variability of the starting material, hybridisation cocktail, the arrays or the scanning, etc. Therefore, to minimise artefacts due to batch effect, I suggested, to process the samples in the same batches for the CNV calling step as they have been processed up to the hybridisation step of the laboratory protocol. Around 100 individuals were analysed against a reference set of ~ 100 individuals. From this comparison, the sample's copy number data were inferred.

The copy number algorithms depend on comparing signal for each marker in each sample against a reference set formed from a group of samples (http://www.affymetrix.com). The assumption is that for each marker, the reference state in the group will be copy number = 2 (except for the Y chromosome, where the reference state is 1), and hence deviations from the reference can be detected by forming the log ratio of each marker's signal compared to its reference value. For the autosomes, the reference value of a particular probe is estimated using the median signal of all samples in the reference for the specific probe. For the X chromosome, the reference value is estimated using the samples, determined to have two X chromosomes. Respectively, for the Y chromosome, the reference set is estimated by taking into account only the samples, having one Y chromosome. In the process of calculating the signal, various normalisation steps are made in order the signal from each sample to be compared with each other.

To assess the reliability of the copy number calls, the interquartile range (IQR) of the non-normalised $\log 2 ratio$ was used. The IQR is a measure of dispersion/spread of the data. It is the difference between the 75^{th} *percentile* (named 3^{rd} *quantile*) and the 25^{th} *percentile* (named 1^{st} *quantile*). The formula for the IQR is as follows: 3^{rd} *quantile* $- 1^{st}$ *quantile* or alternatively 75^{th} *percentile* $- 25^{th}$ *percentile*. The IQR is a robust measure for data spread as it represents the central 50% of the data and is relatively less affected by outliers.

The default QC cut-off of the program for the IQR was < 0.4 and samples with $IQR > 0.4$, were not analysed further;

4. Segment report step- the segment reporting tool generated copy number segment report for each sample. The report file comprised all the segments that exhibited a copy number change. The filtering criteria used for this step were to take into account fragments comprising ≥ 10 consecutive SNPs and segments that were > 100 kb in length.

The file with the report comprised the genomic location, copy number state, size of the segment, number of markers and an overlap with known CNVs.

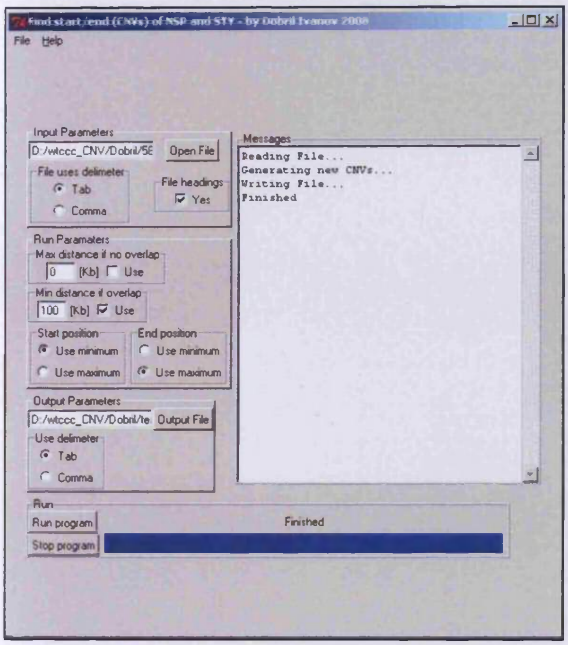
The CNV detection for the bipolar cases was carried out by me. The CNV calling of the controls and the schizophrenia cases was completed by Dr. George Kirov and me. The CNV identification step took approximately four months computer running time. Subsequently the data were manipulated in a number of different ways, which are described in detail in the next section.

3.6.2. Manipulation of data post-CNV detection

After the CNV detection, a quality control step was performed. This step comprised the following:

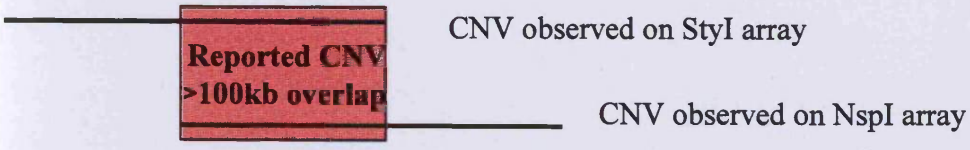
- Only deletions or duplications found independently on both arrays with overlap of the segments identified on the StyI and the NspI arrays of ≥ 100 kb in length were taken into account. Matching of the segments of the two arrays was performed by using a standalone Windows-based program developed specifically for this analysis by Dr. Dobril Ivanov (Figure 17, page 74; available at http://x004.psycm.uwcm.ac.uk/~dobril/combine_cnvs_NSP_STY/);

Figure 17 Program for matching fragments identified from the two Affymetrix arrays (that is NspI and StyI)



The program worked with a file containing the identified segments obtained from the Affymetrix® Genotyping console. The minimum distance of overlap between the called segments per person between the NspI and StyI arrays was set to be 100 kb. The most conservative approach in reporting the CNVs that were observed on both arrays at a given genomic location was chosen- only the segments produced from the start and end position of the exact overlap between the two arrays were used for further analysis (Figure 18);

Figure 18 Schematic representation of the rationale for matching fragments identified from the two Affymetrix arrays (NspI and styI)



- Samples with ≥ 20 deletions and/or duplications were excluded from further analysis;
- CNVs with low SNP density (< 3 SNPs per 100 kb) were also excluded.

The post-CNV identification manipulation of the data was performed using SPSS® 16.

Common copy number variants (found in ≥ 1% of the samples) were excluded from the subsequent analysis. CNVs that overlapped with > 50% of their length to common CNV regions were also excluded.

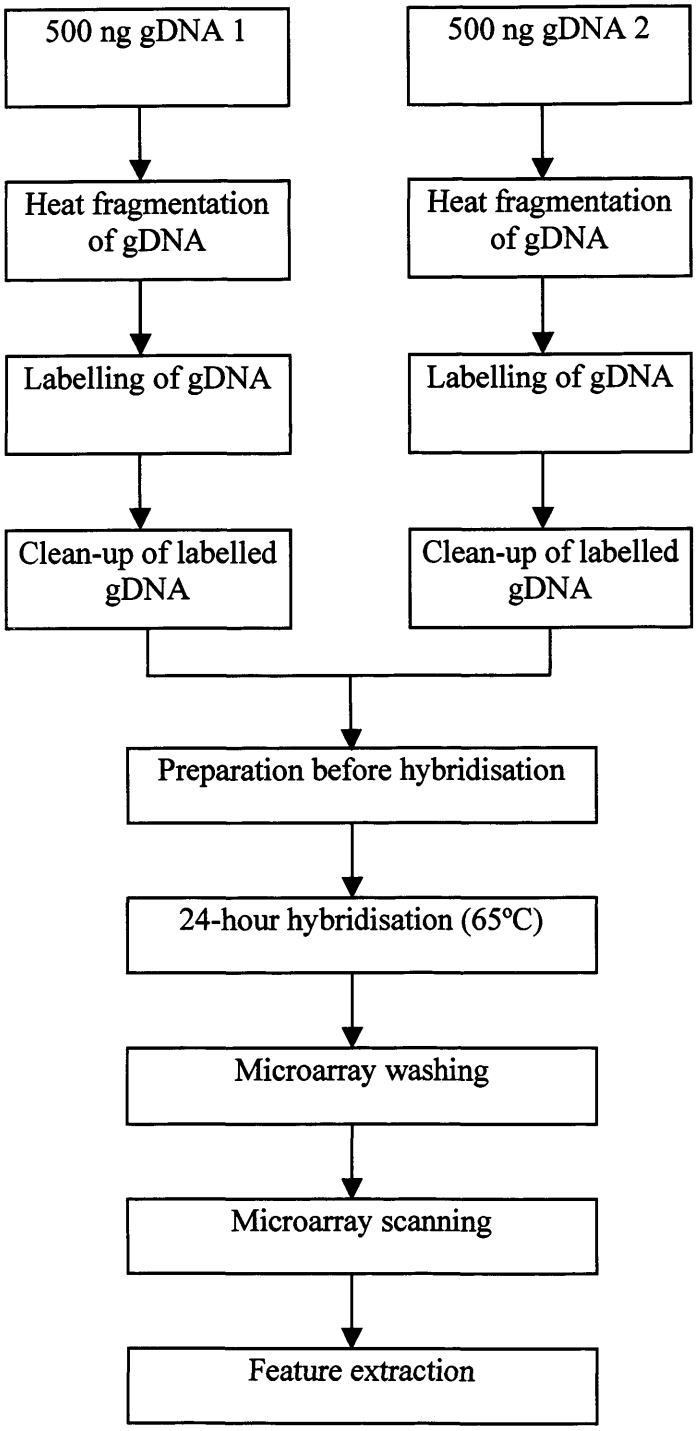
Using PLINK (Purcell et al. 2007) version 1.05, all rare deletions and duplications which remained after applying the above-mentioned criteria, were listed in a custom track for visualisation in USCS Genome Browser (http://genome.ucsc.edu/) and is also publicly available for download at http://x004.psycm.uwcm.ac.uk/~detelina/.

The genomic coordinates used in this PhD are based on March 2006 human genome sequence assembly (UCSC Hg18, NCBI build 36).

3.6.3. Validation of large copy number variants (≥1 Mb)

The validation of the large copy number variants was performed by me. An oligo array Comparative Genomic hybridisation (aCGH) platforms- Agilent Human genome CGH 4x44K microarray, was the method of choice for validation of some of the CNVs inferred by the Affymetrix platform. Agilent oligo aCGH platform enables studying genome-wide copy number variations at a high resolution. The probes on the Agilent aCGH microarrays are 60-mer oligonucleotides synthesized *in situ* using Agilent's inkjet SurePrint™ technology. On the 44K array that was used for validation, ~ 43,000 60-mer oligonucleotide probes are printed on a glass slide. It provides coverage with an average resolution of 30 – 35 kb. The probe selection was biased towards genes, with 84% of the probes being in intragenic regions. Well-characterised genes are represented by at least one probe (http://www.chem.agilent.com). The workflow of the experiment is presented in Figure 19 (page 76).

Figure 19 Workflow of the Agilent microarray experiment



adapted from http://www.chem.agilent.com

As presented in Figure 19, the preparation of the Agilent microarrays included the following: fragmentation of the DNA, chemical labelling, subsequent hybridisation onto the array and scanning of the glass slide. A detailed explanation of the steps is provided below:

1. DNA preparation

The required quantity of the DNA for the Agilent experiment was $500ng$, with minimum concentration of $62.5ng/\mu l$. The DNA concentration was estimated using PicoGreen® dsDNA reagent in a Fluoroscan Ascent Fluoremeter as described in section 3.2.4 (page 63). The samples were diluted to a final concentration of $70ng/\mu l$ and subsequently aliquoted into 96-well Abgene skirted plates. Instead of using a healthy control with which to compare the DNA case sample, for cost-effective reasons, it was decided to use only DNA from cases. This was possible, as we were interested in testing only certain regions of the genomes of the studied individuals, and as long as we were comparing cases with different CNVs, we were confident that we were going to detect the CNVs of interest if they were genuine.

2. Heat Fragmentation

The DNA samples were incubated at $95^0 C$ in a PCR machine with heated lid for 10 minutes. The heat fragmented DNA was stored on ice until ready for chemical labelling;

3. Chemical Labelling

The Agilent labelling kit differentially and directly labels gDNA samples with fluorescent dyes in one step, using non-enzymatic procedure. Sample DNA from the test and the reference human genomes were labelled with different fluorescent dyes-Cyanine 3 (Cy3) and Cyanine 5 (Cy5) respectively. The fluorescent signal intensity at each spot of the array subsequently serves as a measure, indicative of the amount of the sample that has been bound to the DNA sequence of the particular probe. The ratio between the intensities of Cyanine 3 and 5 at each probe shows the relative quantities of the test and reference DNA. This ratio was used to detect copy number.

The Cy3 and Cy5 labelling master mixes were prepared by combining the components presented in Table 10:

Table 10 Master mix for the labelling reaction

Reagents	Per reaction (μl)	Per slide (μl)
Nuclease-free water	0.5	2.5
Cy3 or Cy5	0.5	2.5
10 x labelling solution	1	5
Final volume of Labelling Master Mix	**2**	**10**

- The appropriate amount of Labelling Master Mix was added to each well containing the gDNA to make a total volume of $10\mu l$;
- The PCR tubes were incubated at 85°C for 30 minutes;
- $10\mu l$ of nuclease-free water was added to each PCR tube to make a total volume of $20\mu l$;
- The PCR tubes were stored on ice until dye removal with KREApure columns.

4. Removal of non-reacted fluorescent dyes:

Non-reacted Cy3 or Cy5 dyes could interfere with the subsequent microarray experiment and increase background noise, if they are not efficiently removed prior to hybridization. Agilent KREApure columns were used to remove non-reacted fluorescent dye. $20\mu l$ of the labelled gDNA was put onto the Agilent KREApure column and subsequently centrifuged at speed of minimum 16,000g to collect the purified labelled gDNA in a collection tube.

Afterwards, the Cy3 and Cy5 labelled DNA samples were combined into eppendorf tubes to a total volume of $37\mu l$.

5. Preparation of labelled Genomic DNA for Hybridisation

The combined samples were concentrated to $22\mu l$ using speed vacuum centrifuge. Subsequently, the samples were transferred to a skirted Abgene plate and the hybridisation master mix was added. The hybridisation master mix was prepared according to volumes given in Table 11.

Table 11 Hybridisation Master mix

Reagents	Volume per Hybridization (μl)	Volume per slide (including excess) (μl)
Cot-1 DNA ($1mg/ml$)	5	25
Agilent 100X CGH Blocking Agent	1	5
Agilent 2X CGH Hybridization Buffer	55	275
Final volume of hybridization master mix	**61**	**305**

- After adding the hybridisation master mix, the samples were put on PCR machine to incubate at $95^0 C$ for 3 min;
- The samples were transferred immediately to a heating block at 37°C and incubated for 30 minutes;
- The Agilent-CGH Block was added to each tube containing the labelled gDNA and Hybridisation Master mix to make the final volume of hybridisation sample mixture of $110 \mu l$.

6. Microarray Hybridisation

Each microarray was printed on the side of the glass slide containing the "Agilent"-labelled barcode. This side is called the "active side". The numeric barcode is on the "inactive side" of the glass slide. The hybridization sample mixture is applied directly to the gasket slide and not to the active side of the microarray slide. Then the active side of the microarray slide is placed on top of the gasket slide to form a "sandwich slide pair". The hybridisation sample mixtures were dispensed onto the gasket. Eight samples (four labelled with red and four labelled with green dye) were placed onto one glass array. Hybridisation chamber was used to cover the sandwiched slides. The assembled slide chambers were put onto a rotator rack in a hybridisation oven set to 65°C for a 24-hour hybridisation.

The 24-hour hybridisation was carried out at the Department of Pathology using their Agilent Microarray Hybridization Chambers and Agilent Microarray Hybridization Oven. Dr. Yumin Teng from the Department of Pathology provided advice with regards to working with the hybridization chamber, oven and the subsequent washing steps.

7. Microarray Washing

The washing of the arrays was done following the conditions in Table 12:

Table 12 Microarray washing procedure

Step	Dish	Wash Buffer	Temperature	Time
Diassembly of the array-gasket sandwich	1	Oligo aCGH Wash Buffer 1	Room temperature	
1st wash of the array	2	Oligo aCGH Wash Buffer 1	Room temperature	5 minutes
2nd wash of the array	3	Oligo aCGH Wash Buffer 2	37°C	1 minute
Acetonitrile wash of the array	4	Acetonitrile	Room temperature	1 minute
3rd wash of the array	5	Stabilization and Drying Solution	Room temperature	30 seconds

8. Microarray scanning using Agilent Scanner

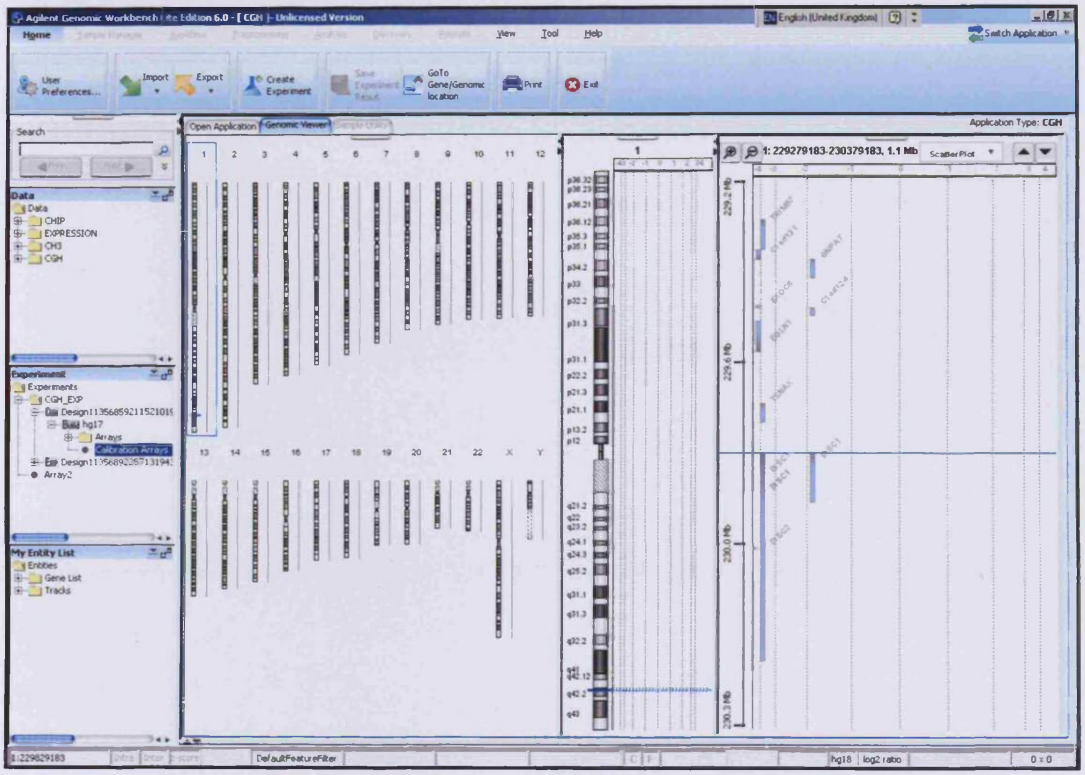
The scanning of the arrays was performed at the Department of Medical Genetics by trained personnel. The data in TIFF format were transferred onto an external hard disk.

9. Data extraction using Feature Extraction Software

The Feature extraction software (downloaded from http://www.chem.agilent.com) was used for data extraction of microarray TIFF images of the scanned Agilent CGH microarrays. This software used the intensity data captured in the TIFF image at each probe spot to produce a file with the $\log 2 ratio$ of the red and green intensities of each probe.

Subsequently, the result files were analysed using SPSS® 16 and software developed by Agilent specifically for the analysis of microarray data- Agilent Genomic Workbench edition 6.0 (Figure 20, page 81).

Figure 20 Agilent Genomic Workbench edition 6.0



3.6.4. Follow-up of 2 loci implicated in schizophrenia in two families with a proband affected with bipolar disorder

Two families with a proband harbouring a specific CNV were genotyped using Human610-quad arrays (Illumina Inc, San Diego, California) in order for the results to be put through second platform for validation.

The DNA samples from the two families were normalised to the required concentration for performing the Illumina assay by me.

The Illumina assays were carried out by Dr. Masashi Ikeda and Dr. Irina Zaharieva following manufacturer's protocols (http://www.illumina.com).

The used Human610-quad Illumina array provides comprehensive genomic coverage with access to more than 550,000 loci.

The protocol features single-tube sample preparation and whole-genome amplification without PCR or ligation steps. After hybridizing an unlabelled DNA sample to the array, two-step allele detection provides high call rates and accuracy. Selectivity and specificity are accomplished in two steps. Target hybridization to bead-bound 50-mer oligonucleotides provides selectivity and enzymatic single-base extension provides specificity. The single base extension also incorporates a labelled nucleotide for assay readout. Then the signal intensities are detected using Illumina iScan System (http://www.illumina.com). The assay data were analysed using Illumina GenomeStudio Genotyping Module.

3.7. Bioinformatical and statistical methods

3.7.1. PLINK

CNV association analyses were performed using whole genome data analysis toolset- PLINK version 1.05, obtained from the following website- http://pngu.mgh.harvard.edu/~purcell/plink/ (Purcell et al. 2007). PLINK was run on a Linux-based multiprocessor server with the following specifications:

Operating System (OS): Fedora Core 12 Linux x86_64

Processor: 4 Dual-Core AMD Opteron 8224 SE @ 3.2GHz

Random Access Memory (RAM): 132 Giga Byte

Hard-Disk Drive: 1.8 TB

As the CNV analyses performed with PLINK were not requiring large resources of computer power, the same analyses could be executed on a desktop computer as well.

PLINK is a genetic-analysis tool designed to handle large data sets of whole-genome data (Purcell et al. 2007). It provides an environment for a rapid manipulation and analyses of such data sets. PLINK includes data management, summary statistics, population stratification, association analysis and identity-by-descent estimation. In addition to whole-genome SNP analysis, PLINK offers functions for downstream analyses of CNV data. It is **not** a tool for CNV identification, but for CNV data management and statistical analyses.

There is a comprehensive web-based documentation http://pngu.mgh.harvard.edu/~purcell/plink/cnv.shtml with respect to CNV analyses.

With the help of PLINK the following tasks were performed:

1. Basic support for CNV data and file types: The basic files that were required by PLINK were: CNV file, MAP file and FAM file.

The **CNV file** comprised the following information:

Family ID

Individual ID

Chromosome

Start position (base-pair)

End position (base-pair)

Type of variant (1=deletion; 3=duplication)

Confidence score associated with the variant

Number of probes in the variant

The confidence score and the number of sites values were not used in a direct way and did not impact the analysis performed with PLINK.

Prior to any analysis, PLINK required a **MAP file**. It was created using PLINK on the basis of the CNV file and the result MAP has entries corresponding to the start and end positions of all segments. This file was needed by PLINK for subsequent parsing and analysis of the CNV data.

The MAP file is created using the following command:

plink --cnv-list BP_CNV.cnv --cnv-make-map

The **FAM file** contained ID and phenotype information:

Family ID

Individual ID

Paternal ID (used only when trio data were analysed, 0 was used for the purposes of the current study)

Maternal ID (used only when trio data were analysed, 0 was used for the purposes of the current study)

Gender (1=male; 2=female)

Phenotype (1=unaffected; 2=affected)

The basic commands for reading a list with CNV data were:

plink --cnv-list BD_CNV.cnv --fam BD_CNV.fam –map BD_CNV.map

plink --cfile BD_CNV

2. Filtering steps/PLINK commands:

A. CNV filtering based on frequency:

plink --cfile BP_CNV --cnv-freq-exclude-above 45 --cnv-overlap 0.50 --cnv-write --BP_CNV_rare

B. Filtering based on size:

--cnv-kb 100 (minimum size in kb that PLINK will take into account when analysing CNV data, the number is put for illustration purposes)

--cnv-max-kb 500 (maximum size in kb that PLINK will take into account when analysing the CNV data, the number is put for illustration purposes)

C. Filtering based on CNV type:

--cnv-del (PLINK will take into account only deletions in the analysis)

--cnv-dup (PLINK will take into account only duplications in the analysis)

If neither deletions nor duplications are specified, then the analysis includes all types of CNVs.

3. Association analysis:

A global test of CNV burden in cases versus controls was performed, based on permutations to generate empirical p-values (10,000 permutations). All p-values reported throughout the PhD are 2-sided, based on comparing the rate of CNVs in two independent samples (usually cases versus controls).

The following commands were used:

--cnv-indiv-perm

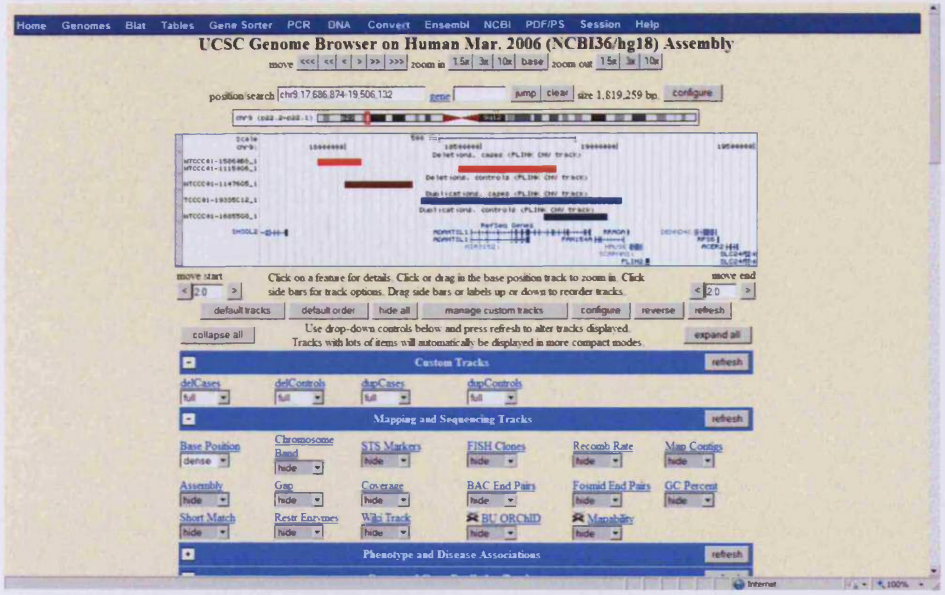
--mperm 10000

--cnv-2sided

4. Visualisation of data:

PLINK was used to generate custom annotation track files (in BED format). In order to visualise the data, these custom-built annotation files were uploaded to UCSC Genome Browser (http://genome.ucsc.edu/cgi-bin/hgGateway) (Figure 21).

Figure 21 Screenshot from the UCSC Genome Browser



5. Miscellaneous PLINK commands:

--cnv-intersect regions.list/--cnv-disrupt- were used to extract a specific set of segments that overlap with one or more specified regions. This command was applied to test if segments intersect with genes. The file with the gene information and gene location was provided by PLINK (downloaded from http://pngu.mgh.harvard.edu/~purcell/plink/dist/glist-hg18);

--segment-group- used in order to group and report sets of segments that span a particular position. More specifically, the command takes all segments in a given region (whole genome unless otherwise specified) and forms "pools" of overlapping segments. Several pools of overlapping segments will be created; these will be listed in order of decreasing size (number of segments). Using this command it was tested whether segments are observed more often in cases than controls. After "pooling" segments found in cases and controls for a specific region, a Fisher's exact test (3.7.2, page 86) was applied, by using the number of segments in cases and controls;

--cnv-write- given a set of filters is applied, a new CNV file can be generated comprising only samples/segments that satisfy the provided criteria.

3.7.2. Association analysis

To provide an estimation whether CNVs at a particular locus (i.e. 16p11.2, 15q13.3, 17p11.2, etc.), is associated with disorder, a ***Pearson's*** χ^2 test with one degree of freedom (df) can be calculated (Altman 1991). The test compares the observed frequencies with the expected frequencies under the assumption of independence (i.e. CNVs are not associated with an increased frequency in cases and the rate is similar in cases and controls). The χ^2 test statistic is an appropriate statistical measure, when 80% of the cells in the contingency table (i.e. $2x2$ table) have expected counts ≥ 5, and all cells show expected counts > 1 (Altman 1991). The following website was used to calculate the significance level (p-value): http://www.graphpad.com/quickcalcs/contingency1.cfm, 2-tailed χ^2 with Yates' correction.

In the case of small expected frequencies, two-tailed ***Fisher's exact test*** was used. The significance level (p-value) was then compared with a pre-specified $\alpha = 0.05$ value or Type I error. Results with a p-value $\leq \alpha$ were deemed statically significant. Fisher's exact test is a statistical significance test used in the analysis of contingency tables where sample sizes are small. The significance of the deviation from the null hypothesis is estimated exactly rather than relying on approximation (this is the case with χ^2) (Field 2009). An online Fisher's exact test calculator was used (http://www.langsrud.com/fisher.htm).

3.7.3. Calculation of statistical power

The power of a statistical test represents the probability to detect an effect size of a particular magnitude with a specified Type I error rate (α, or false positive rate) and a particular sample size, $power = 1 - \beta$ (β is Type II error rate, or false-negative rate). The analysis for power requires the assumption that genuine effect exists in the

studied populations. The power calculations were performed using pwr.chisq.test package which is part of the R statistical language (http://cran.r-project.org/).

3.7.4. Pathway analysis

Genes, participating in the nervous system pathways (long-term potentiation, long-term depression and neurotrophin signalling pathway), development (axon guidance), and circadian rhythm (Homo sapiens) were downloaded from KEGG (Kyoto Encyclopedia of Genes and Genomes). More specifically, KEGG PATHWAY database was used (http://www.genome.jp/kegg/pathway.html).

KEGG is a database comprising 16 main databases, containing systems information, genomic information, and chemical information (Kanehisa and Goto 2000; Kanehisa et al. 2006). One of these 16 databases, KEGG PATHWAY, is a collection of manually curated maps (based on published materials) for metabolism, other cellular processes, human diseases, organismal systems, etc.

In addition, the IDs of brain-expressed genes were obtained using Ensembl via BioMart Genome browser (http://www.ensembl.org, (Flicek et al. 2010)).

PLINK, version 1.07 was used to perform the pathway analysis (Raychaudhuri et al. 2010). The rates of CNVs affecting specific gene sets in cases versus controls were compared, i.e. case-control strategy is employed to test the above-mentioned pathways for association with bipolar disorder. 2-sided asymptotic p -values are used to estimate the significance level of association.

The following PLINK commands were applied:

A. To test if genes from a certain pathway are enriched in the CNV data, relative to all observed CNVs

--cfile BD

--cnv-count glist-hg18.dat

--cnv-subset pathway.txt

--cnv-enrichment-test

B. To test if genes from a certain pathway are enriched in the CNVs, relative to all CNVs affecting genes:

A list with CNVs affecting genes was generated using PLINK:

--cfile BD

--cnv-intersect glist-hg18.dat

--cnv-write

--out my_genic_CNV

Subsequently, the genes from the studied pathway were tested for enrichment in the CNV data, relative to all genic CNVs:

--cfile my_genic_CNV

--cnv-count genes.dat

--cnv-subset pathway.txt

--cnv-enrichment-test

3.8. Comparison between bipolar cases and an extended control group

The CNV data from the whole WTCCC study that bipolar disorder took part in, became publicly available in 2010 (Wellcome Trust Case Control Consortium 2007).

The studied sample comprised 11,909 UK Caucasian individuals. This sample set included 1650 bipolar cases and 10,259 individuals affected with several non-psychiatric diseases (Table 13).

Table 13 Cohorts that participated in the WTCCC study

Sample set (abbreviation)	Number of studied individuals
bipolar disorder (BD)	1650
controls (1958 cohort +NBS)	2777
coronary artery disease (CAD)	1855
Crohn's disease (CD)	1450
hypertension (HT)	1864
rheumatoid arthritis (RA)	1374
type 1 diabetes (T1D)	1903
type 2 diabetes (T2D)	1813

These individuals are of white European ancestry and live in the UK. All these subjects have been studied in the WTCCC study for associations of single nucleotide polymorphisms with the corresponding disorders (Wellcome Trust Case Control Consortium 2007). In addition, the WTCCC study also included 1434 controls born in one week in 1958, and 1343 controls from the National Blood Transfusion Service (NBS). These controls have already been analysed in this PhD work and therefore were not used in this analysis.

The genotyping was performed using Affymetrix 500K arrays comprising the NspI and StyI arrays. Subsequently, using the intensities ($\log 2ratio$) at each SNP, copy number state was inferred. The CNV identification was performed by WTCCC and the results with the already called CNVs per individual were sent to us through a secure ftp server.

The procedure used to detect CNVs based on the raw intensity data involved normalisation of the data and subsequent CNV discovery (Barnes et al. 2008; Olshen et al. 2004; Price et al. 2005). Several quality control (QC) filter steps have been applied, i.e. removal of samples which failed SNP genotyping, removal of samples that had too many CNVs and samples producing outlier intensities (personal communication with WTCCC).

The CNV calling and the above-mentioned QC filtering measures were performed at the Sanger Centre and subsequently the data have been made publicly available. It is of note that the analysis of the CNVs in the bipolar cases involved exactly the same raw SNP genotyping intensity data that were analysed in this PhD work.

Subsequently, I applied additional QC filtering steps on the data received from WTCCC. Only CNVs surpassing the following thresholds were chosen for analysis:

- rare CNVs observed in $\leq 1\%$ of the sample;
- CNVs ≥ 100 kb;
- CNVs with density of > 3 SNPs per 100 kb
- CNVs comprising ≥ 10 consecutive SNPs per variant.

All chromosomal positions with respect to results obtained from WTCCC, were according to NCBI Build 35 (hg17), May 2004 of the UCSC Genome Browser. To allow comparisons with the data generated in this PhD work (which are according

to NCBI Build 36 coordinates), the genome coordinates were converted to the appropriate versions of the genome browser, before examining the regions of interest.

PLINK and SPSS® 16 were applied to manipulate and analyse the CNV data obtained from WTCCC (Purcell et al. 2007).

For comparison purposes, schizophrenia samples, coming from two sources were also used:

- Part of publicly available International Schizophrenia Consortium (ISC) study (International Schizophrenia Consortium 2008). The sample set included 3391 cases and 3181 controls, genotyped with Affymetrix 6.0 or 5.0 arrays. The ISC data were also according to NCBI Build 35 (hg17);
- Schizophrenia cases (n=471) genotyped with the same array as WTCCC study (Affymetrix 500 K array) and analysed using the same pipeline as the samples in this PhD work (Kirov et al. 2009a). More information about this sample set has been already presented in section 3.1.1.2 (page 60).

4. Results

4.1. Results from quality control

4.1.1. Wellcome Trust case-control consortium quality control

2284 bipolar DNA samples were prepared to 100 μg / ml dilution and ~70 μl of each DNA sample was sent to the Wellcome Trust Sanger institute, Cambridge. Based on the quality control (QC) measures taken at the Sanger institute, 1997 DNA samples were approved for genotyping. The QC measures performed at the Sanger institute were based on: DNA quantification, test for degradation and PCR amplification (Wellcome Trust Case Control Consortium 2007).

1997 bipolar DNA samples were sent to Affymetrix Services laboratories in USA for genotyping. 3004 controls were selected using the same criteria and were sent along with the case's DNA to be genotyped in the same pipeline.

After the genotyping with the Affymetrix GeneChip 500K array, more quality control steps were performed at the Sanger Institute. These included: sample call rate, duplication of samples, relatedness and evidence of non-European ancestry. Samples which did not pass the above-mentioned criteria were excluded from further analyses. 129 bipolar individuals were excluded at this stage. 66 samples were excluded from the control set (24 from the 1958 cohort and 42 from the NBS). In Table 14 are presented the QC measures and the corresponding number of individuals not passing the QC criteria separately for the cases and controls (Wellcome Trust Case Control Consortium 2007).

Table 14 Quality control measures, performed at Sanger Institute, after the genotyping with the Affymetrix 500K array

Collection	Number of individuals							
	Initial	Missingness	Discordance	Non-European ancestry	Duplicate	Relative	Total excluded	Accepted
1958 cohort	1504	9	4	6	4	1	24	1480
NBS	1500	8	5	14	0	15	42	1458
BD	1997	30	0	9	77	13	129	1868

These quality control steps resulted in a final dataset of **1868** bipolar individuals and **2938** controls.

4.1.2. Quality control of the Affymetrix genotyping

Markers already present on the 500K Affymetrix array, were chosen to be genotyped with a separate genotyping platform-Iplex® Sequenom MassARRAY platform. The reason for performing the genotyping was to estimate the concordance between the two platforms in an attempt to double check the genotyping accuracy of the Affymetrix array. Furthermore, this will also provide information for possible sample swaps.

All 1868 bipolar case DNAs that were already genotyped by Affymetrix, were genotyped with the Iplex chemistry for 46 SNPs by me. Overall, ~75,000 genotypes were analysed and > 99.5% agreement between the two genotyping platforms was observed.

The correspondence rate between the two platforms was very high, which indicated that the obtained results from the genotyping with the Affymetrix array are robust and trustworthy and that no sample swaps have occurred.

4.1.3. Quality control of the CNV detection algorithm

The Affymetrix ® Genotyping console ™ v2.1 software was used for inferring the CNV structure. The console utilises intensity raw data from the SNP genotyping.

All intensity data (CEL files) from the WTCCC Sanger institute were put through the CNV calling software.

The numbers of individuals, mentioned in the 'Initial' column in Table 14 (page 92) (i.e. 1997 BD individuals, 1504 from the 1958 cohort and 1500 from the National Blood Service sample set) were analysed for copy number variants. The reason for using all of the individuals was that these samples were genotyped by Affymetrix and at that point it was not known which individuals should be excluded from the analysis. Consequently, problems were observed with regard to duplicated samples, relatedness and non-Caucasian ancestry. At a later stage, the samples that showed any inconsistencies with regard to duplications, relatedness or non-Caucasian ancestry were excluded.

For quality control of the CNV calling, a number of filtering steps were applied during the CNV detection procedure. Namely these were:

A. The default parameters of the Affymetrix ® Genotyping console software were used for initial quality control filter (intensity quality control threshold $< 93\%$). Intensity QC call rate of $< 93\%$ is indicative of a low sample quality. At this stage, 32 samples were excluded from the BD sample set and six and 13 from 1958 cohort and NBS sample sets respectively;

B. The interquartile range (IQR) of the $\log 2$ ratio was used to evaluate the quality of the arrays for the copy number analyses. Samples with $IQR \geq 0.40$ were excluded from the analyses. In total, 76 individuals were excluded due to $IQR \geq 0.40$ (47 from the bipolar sample, 11 and 18 from the 1958 cohort and NBS sample sets respectively);

When many samples produced high IQR measures (≥ 0.40), this was indicative of presence of a batch effect. At first, at the CNV calling step, samples from one batch were analysed for copy number change using reference sets from a separate processing batch. Using this method, many samples were being filtered out due to $IQR \geq 0.40$. This effect was more pronounced for the StyI array.

Figure 22 and Figure 23 represent histograms with IQRfrequency data for NspI and StyI arrays for the 1958 cohort, showing the difference in the quality of the arrays.

Figure 22 Histogram showing the frequency of distribution of *IQR* of 1958 cohort for the NspI array

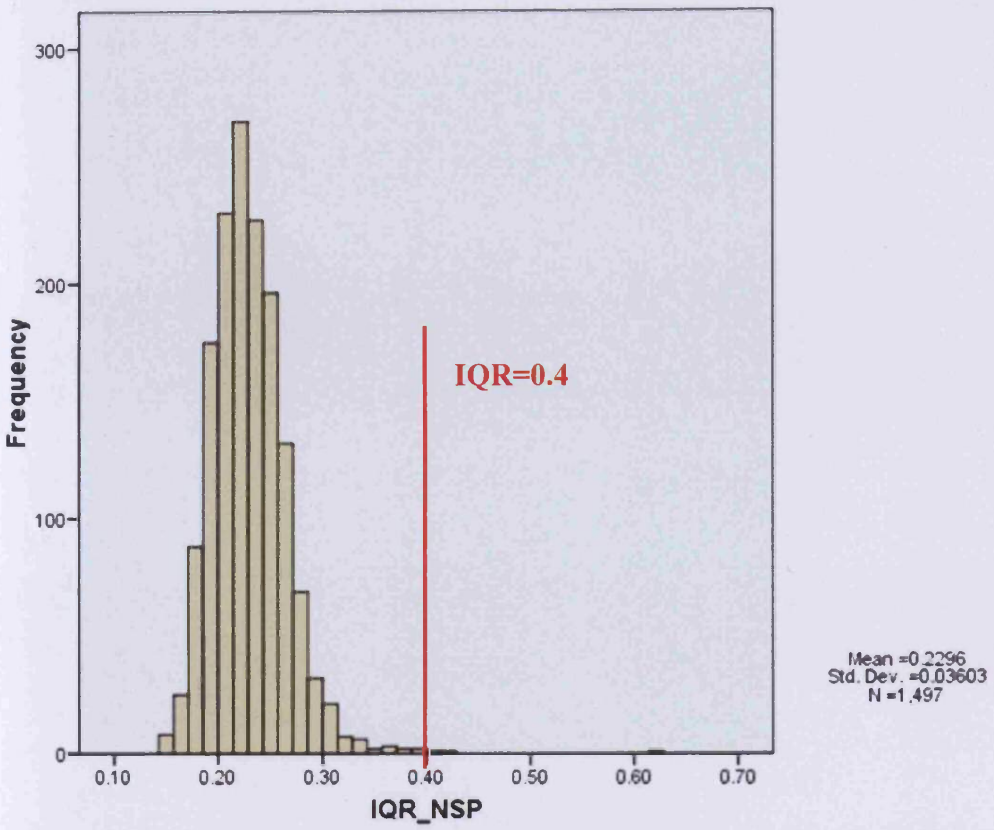
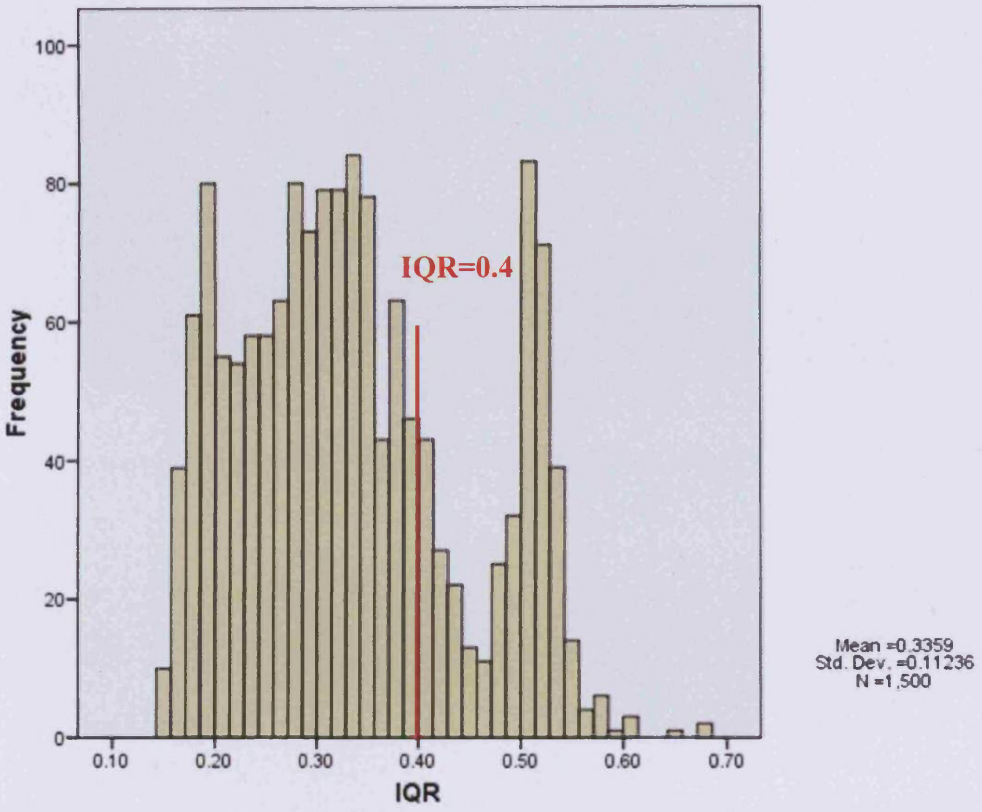
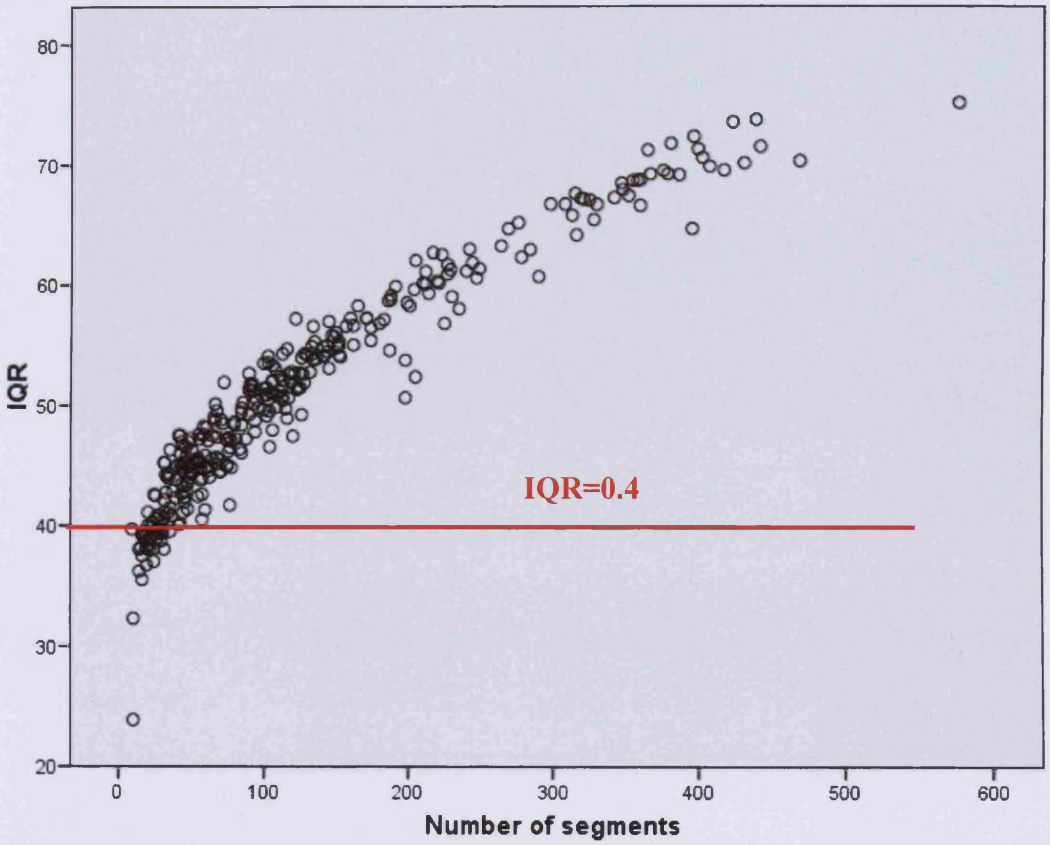


Figure 23 Histogram showing the frequency of distribution of IQR of 1958 cohort for the StyI array



When the IQR was ≥ 0.40, many copy number variants were observed, in the range of hundreds per person. This detected CNV rate per person, including the controls, was much higher than the average reported in the literature (Itsara et al. 2009; McCarroll and Altshuler 2007; Redon et al. 2006). In addition, it is highly unlikely that on one plate no individuals would have more than 20 CNVs per person, while on other plate most of the individuals would have copy number variants in the range of hundreds per person. Clearly, this was due to some technical issue. This pointed that these were spurious data due to systematic confounding effects. Figure 24 presents an example of a number of segments and the corresponding IQR for samples with high IQR (the NBS control set genotyped with the StyI array). A strong positive correlation was observed between IQR and the corresponding number of segments, with many segments detected when the IQR was elevated (Pearson's correlation coefficient, $\rho = 0.948$).

Figure 24 An example of the observed number of segments and the corresponding *IQR* for samples with high *IQR*



NBS cohort genotyped with StyI array

Therefore, the samples were analysed against samples coming from the same processing batches.

When samples were analysed using reference sets from the same processing batches, the number of individuals being filtered out due to $IQR \geq 0.40$, dropped dramatically. For comparison, Figure 25 and Figure 26 present NBS cohort CNV analyses before and after having the batch effect corrected.

Figure 25 Frequency of *IQR* in the NBS sample, not taking the batch effect into account

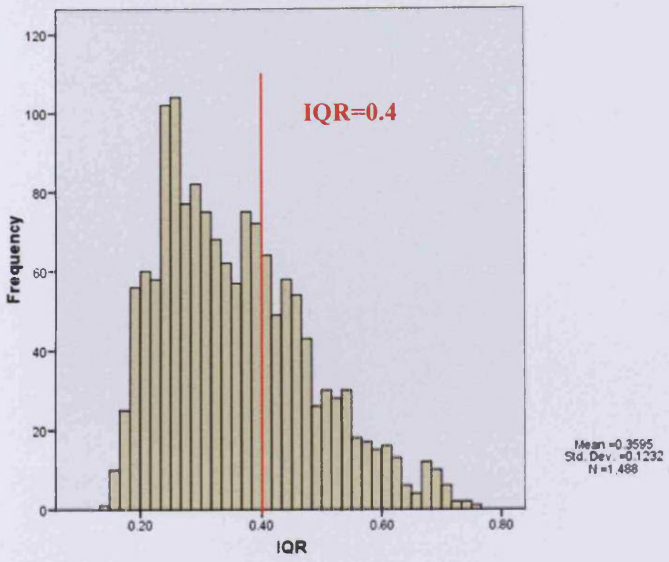
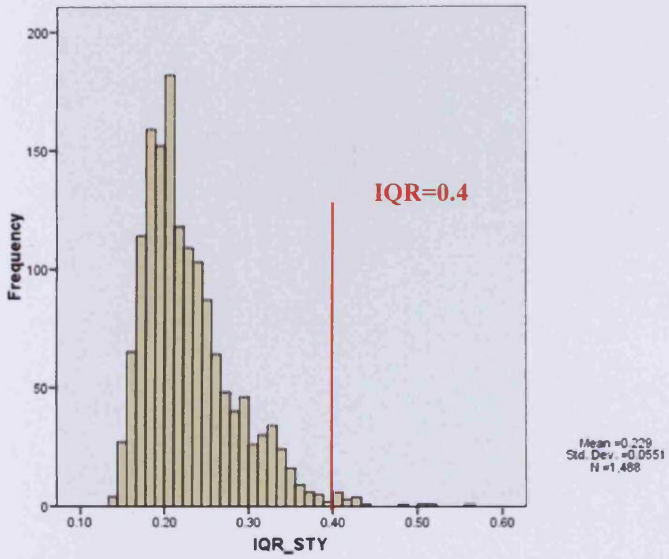


Figure 26 Frequency of *IQR* in the NBS sample after the batch effect was corrected



C. The following parameters of the Affymetrix software were used for CNV calling: only CNVs ≥ 100 kb and only if they spanned ≥ 10 consecutive SNPs on each of the arrays were included;

D. Samples with ≥ 20 segments were excluded from the analysis.

In total, 298 individuals were excluded (178 from the bipolar sample, 73 and 47 from the 1958 cohort and NBS sample sets respectively);

Samples with ≥ 20 segments were excluded even if the observed IQR was ≤ 0.4. Samples with $IQR \geq 0.4$ produced with almost no exception > 20 deletions and/or duplications (and > 100 CNVs when IQR was > 0.5), suggesting that observations corresponding to large number of CNVs per person were likely to be false-positive (Figure 24, page 96);

E. Deletions or duplications observed only on one of the arrays (i.e. only on NspI or StyI array) were excluded from the analysis. Effectively this meant that one array was validated with the other one;

F. Segments with very low SNP density (< 3 SNPs per 100 kb) were also excluded. Such segments tend to intersect with low copy repeats or 'difficult' regions of the genome and have an increased probability of being false-positive.

After applying the above-mentioned criteria, 1697 cases and 2806 controls (1411 individuals from 1958 cohort and 1395 from NBS) were further analysed.

4.1.4. Validation with Agilent array

Schizophrenia cases were also analysed in this PhD work. These cases were also part of the WTCCC study and were processed at the same time in the same pipeline as the bipolar cases and the controls. In addition, the CNV calling, QC and analysis were performed in exactly the same way as in bipolar cases and controls.

Validation of some relevant CNVs in schizophrenia samples was performed with Agilent Human Genome CGH 4x44K microarrays. The chosen CNVs (Table 15, page 99) for validation were deletions ≥ 1 Mb (as they provided the most statistically significant results in the schizophrenia study) and the CNVs that were regarded as being pathogenic in schizophrenia (that is, CNVs found to be associated with schizophrenia in a number of independent studies, e.g. 1q21.1, 15q11.2, etc.) (Kirov et al. 2009a). Each CNV was confirmed, reflecting the very stringent quality control criteria employed, including the fact that each call was made independently on both StyI and NspI arrays. Additional family members, where available, were also tested for these loci to examine the segregation with illness and possible *de novo* status.

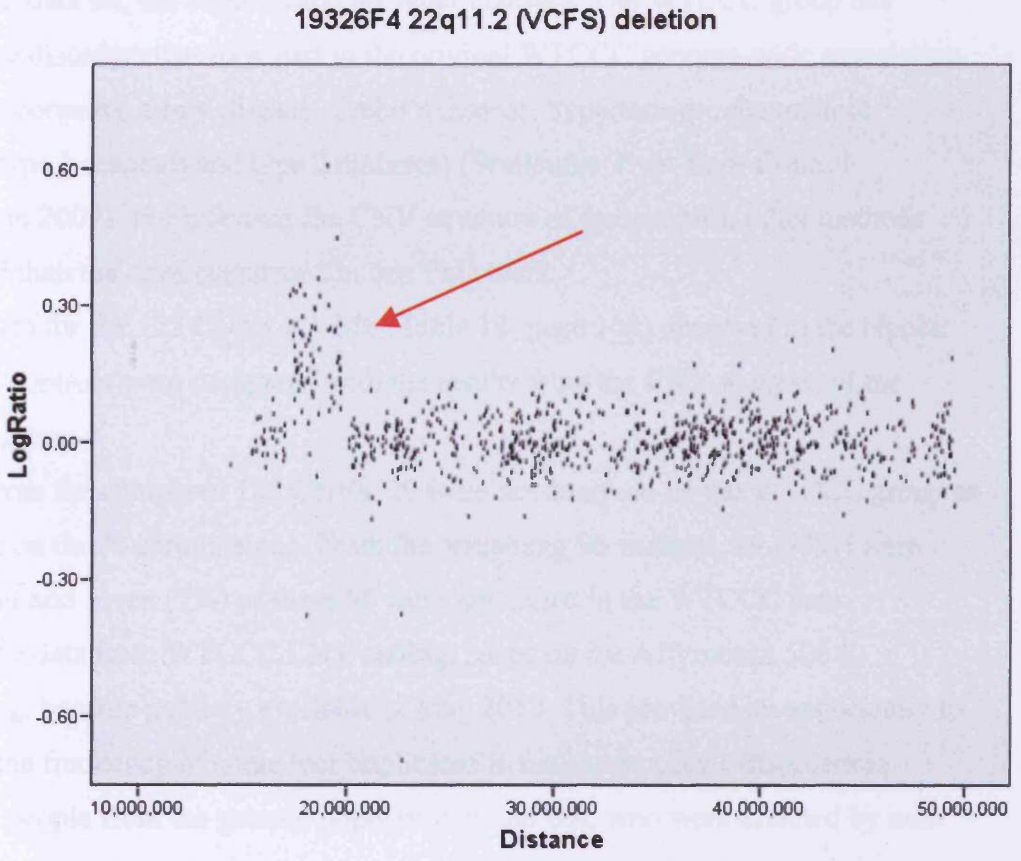
Table 15 Selected CNVs in schizophrenia cases for validation with an Agilent 4x44K microarray

Sample ID	Chromosome position	Start position [Mb]	End position [Mb]
19327C5	2p16.3 (*NRXN1*)	50	51.2
19336A5	2q22.1	139.2	140.2
20577A5	3p24.3	17.2	20.6
19326C3	5p15.33	0.2	1.2
19338E3	6p12.3	48.1	49.3
19339E4	7q34	137.8	139.3
19328D4	7q36.3	157.7	158.8
19338A2	10p13	12.7	14.6
19335B1	15q11.2	20.2	20.8
19335D1	15q11.2	20.2	20.8
19337A5	15q11.2	20.2	20.8
19337C5	15q11.2	20.2	20.8
19339G4	15q11.2-q13.1	21.2	26.2
19327C3	16p13.11	14.9	16.4
19327F2	16p13.11	14.9	16.4
20577B1	16p13.11	14.9	16.4
19326A3	16p11.2	29.5	30.1
19328A3	16p11.2	29.5	30.1
19328G5	16p11.2	29.5	30.1
19325A4	17p12	14	15.4
19338F6	17p12	14	15.4
19338B6	22q11.2 (VCFS)	17.2	19.8
19326F4	22q11.2(VCFS)	17.2	19.8
20579B3	Xp22.33	0.3	2.7
19339D3	Xp22.31	6.5	8

Figure 27 (page 100) presents one of the confirmed copy number variations, namely the 22q11.2 deletion. The deletion exhibited an elevated $\log 2$ ratio of the intensity of the region in question (marked with a red arrow). The deletion appears as duplication on the figure due to intricacy of the specific design of the experiment. In order to make the experiment as cost-efficient as possible, no control DNA was used as reference sample. Samples with different CNVs were analysed against each other which was possible as only specific regions were attempted to be validated and the experiment was not intended to serve as CNV discovery method. In this specific case,

the sample with the 22q11.2 deletion was used as a "control" which will make the other sample appear to harbour duplication at this locus.

Figure 27 Example of 22q11.2 deletion successfully validated with an Agilent 44K microarray in a schizophrenia sample



As the schizophrenia samples, controls and bipolar cases were prepared, genotyped and analysed for copy number variants in the same pipeline, and all relevant (and large) CNVs in the schizophrenia sample validated successfully, it was concluded that there was no need to validate any of the CNVs in the bipolar cases. Furthermore, each CNV had already been called independently by two arrays (StyI and NspI). Nevertheless, two duplications observed in bipolar cases were confirmed because parental DNA was available and in this way the segregation with illness was also tested. This time Illumina arrays were used. These results are presented separately in section 4.2.7.1.1 (page 115).

4.1.5. Comparison between the CNV analysis performed in this PhD work with an independently performed CNV analysis at WTCCC, Sanger Institute

Independently, the WTCCC group have analysed the copy number variation in the bipolar data set, the controls and six other diseases. The WTCCC group has studied the disorders that took part in the original WTCCC genome-wide association study (i.e. coronary artery disease, Crohn's disease, hypertension, rheumatoid arthritis, type 1 diabetes and type 2 diabetes) (Wellcome Trust Case Control Consortium 2007). For inferring the CNV structure of the samples, other methods were used than the ones performed in this PhD work.

Data for the 125 CNVs ≥ 1 Mb (Table 18, page 104) observed in the bipolar cases and controls were compared with the results from the CNV analysis of the WTCCC group.

From the compared 125 CNVs, 29 were not analysed by the WTCCC group as they were on the X-chromosome. From the remaining 96 variants, 89 (93%) were concordant and seven (7%) of these 96 were not called in the WTCCC data.

The data from WTCCC CNV calling, based on the Affymetrix 500 K genotyping, became publicly available in May 2010. This provided an opportunity to examine the frequency of some loci implicated in neuropsychiatric disorders in > 10,000 people from the general population in the UK, who were affected by non-psychiatric disorders. Thus, it provided a chance to compare the observed CNVs in bipolar disorder cases with the ones observed in this large sample set. The results with respect to these analyses are presented in section 4.5 (page 126).

4.2. CNVs in bipolar disorder

4.2.1. Global burden of CNVs

Initially all of the observed CNV were analysed. The results are presented in Table 16.

Table 16 Global burden analysis of all observed CNVs

CNV type	Cases (n=1697)	Controls (n=2806)	CNV per case	CNV per control	case/control ratio	*p*
Del	912	1635	0.54	0.58	0.93	0.05
Dup	1838	3142	1.08	1.12	0.96	0.25
Total	2750	4777	1.62	1.70	0.95	0.03

The *p* is empirical 2-sided based on 10,000 permutations; Del=deletions; dup=duplications

No statistically significant difference was observed between the mean number of duplications when the cases were compared with controls ($p = 0.25$). When deletions were analysed, a trend for overrepresentation in controls was observed ($p = 0.05$). When the total number of CNVs was compared between cases and controls, the controls exhibited an elevated number of CNVs- 1.70 per control when compared with cases-1.62 per case ($p = 0.03$). No correction for multiple testing was applied to these tests.

These observed CNVs were not analysed further as they were not going to lead to reliable results due to difficulties, inherent to the methodology of genotyping common CNVs (which are included in the above-mentioned results). Common CNVs tend to intersect with 'problematic' regions in the human genome and that was one of the main reasons why they were excluded from early Affymetrix arrays, similar to the one that is used in the present study. It has been shown that common CNVs correspond to 'bald spots' in the physical coverage of the early SNP arrays (McCarroll et al. 2008). Moreover, due to their frequency (i.e. being common, > 1%), the reference set will undoubtedly contain a combination of 1, 2, 3, etc. copies of these CNVs which makes the detection unreliable. Furthermore, studies in schizophrenia have shown that the burden of rare (≤ 1%) CNVs is increased in cases when compared with controls (International Schizophrenia Consortium 2008; Walsh et al. 2008; Xu et al. 2008).

Therefore, common CNV (> 1% of the samples) were excluded. In addition, CNVs that overlapped with > 50% of their length to common CNVs were also excluded. Based on these criteria, a data set of rare CNVs was created and was used for all of the subsequent analyses.

4.2.2. Global burden of rare CNVs

The total number of rare CNVs (with frequency ≤ 1% in the population) and the corresponding p -value from the comparison between cases and controls are presented in Table 17. There was no statistically significant increase in the rate of CNVs between cases and controls, and there was a statistically significant association in the opposite direction (i.e. fewer CNVs in cases) when only deletions were analysed ($p = 0.01$). There was not a statistically significant difference between the cases and controls with regard to the number of duplications and the total number of CNVs.

Table 17 Global burden of rare CNVs

CNV type	Cases (n=1697)	Controls (n=2806)	CNV per case	CNV per control	case/control ratio	p
Del	324	632	0.191	0.225	0.85	**0.01***
Dup	538	901	0.317	0.321	0.99	0.84
Total	862	1533	0.508	0.546	0.93	0.09

The p is empirical 2-sided based on 10,000 permutations. Del=deletions; dup=duplications;
*This significant result is for fewer CNVs in cases than controls

4.2.3. Global burden with respect to CNV size

As previous studies in schizophrenia have suggested that the CNV effect comes from CNVs larger than a certain size (International Schizophrenia Consortium 2008; Kirov et al. 2009a), similar analysis was performed in the bipolar cases.

The observed CNVs were divided into size categories and were compared between cases and controls (presented in Table 18). Apart from a number of small deletions (≤ 200 kb) in cases and controls, statistically significant differences between cases and controls were not observed for any other size ranges. The statistically significant result with respect to deletions ≤ 200 kb is for fewer CNVs in cases than controls ($p = 0.03$). It is of note, that nearly all types of CNVs were observed at a lower frequency in cases as compared to controls.

Table 18 Global burden with respect to CNV size

CNV type, size [kb]	Cases (n=1697)	Controls (n=2806)	CNV per case	CNV per control	case/control ratio	*p*
Del<200	142	293	0.084	0.104	0.80	**0.03**
Dup<200	155	261	0.091	0.093	0.98	0.88
Total<200	297	554	0.175	0.197	0.89	0.09
Del 200-500	149	285	0.088	0.102	0.86	0.15
Dup 200-500	277	436	0.163	0.155	1.05	0.55
Total 200-500	426	721	0.251	0.257	0.98	0.72
Del 500-1000	26	34	0.015	0.012	1.26	0.42
Dup 500-1000	73	139	0.043	0.050	0.87	0.35
Total 500-1000	99	173	0.058	0.062	0.95	0.66
Del>1000	7	20	0.004	0.007	0.58	0.24
Dup>1000	33	65	0.019	0.023	0.84	0.47
Total>1000	40	85	0.024	0.030	0.78	0.20

The *p* is empirical 2-sided based on 10,000 permutations. Del=deletions; dup=duplications;

4.2.4. Singleton CNVs

Zhang et al. studied the CNV structure in bipolar disorder cases and made comparisons with healthy controls (Zhang et al. 2008). An increased overall burden of rare structural genomic variants was not observed, although an increase was found with respect to deletions that appear only once in the dataset (16.2% of BD cases in contrast to 12.3% in controls, $p = 0.007$).

In order to test if a similar effect was present in WTCCC BD data, the global burden of single-occurrence deletions and single-occurrence duplications was estimated. Deletions-only and duplications-only datasets were generated. In these datasets, a deletion was defined as a single-occurrence event, even if there was a duplication present in the same region. The same rule was applied to duplications. Thus, the number of single-occurrence deletions and single-occurrence duplications did not sum up to the total number of single-occurrence events.

No statistically significant difference was observed between the proportion of cases and controls with singleton CNVs when deletions and duplications were combined or when deletions and duplications were considered separately (Table 19, page 105).

Table 19 Global burden with respect to singleton CNV events

Single CNV type	Cases (n=1697)	Controls (n=2806)	CNV per case	CNV per control	Case/control ratio	*p*
Del	104	179	0.06	0.06	0.96	0.77
Dup	130	251	0.08	0.09	0.86	0.17
Total	203	355	0.12	0.13	0.95	0.55

The *p* is empirical 2-sided based on 10,000 permutations. Del=deletions; dup=duplications;

Zhang et al. reported that singleton deletions were particularly more common in cases with mania with an onset of illness ≤ 18 years of age (Zhang et al. 2008). In the 65 patients in the WTCCC BD sample with such an early age of onset, there was no statistically significant difference in the rate of singleton CNVs, compared with the controls (Table 20).

Table 20 Global burden with respect to singleton CNV events in bipolar cases with early onset

singleton/age onset mania	Cases (n=65)	Controls (n=2806)	CNV per case	CNV per control	case/control ratio	*p*
Del	4	214	0.06	0.08	0.81	0.83
Dup	4	308	0.06	0.11	0.56	0.29
Total	7	446	0.11	0.16	0.68	0.36

The *p* is empirical 2-sided based on 10,000 permutations. Del=deletions; dup=duplications;

4.2.5. CNVs disrupting genes and analysis of pathways

Following previous studies in schizophrenia (International Schizophrenia Consortium 2008; Walsh et al. 2008), the burden of CNVs that delete, duplicate or disrupt genes was examined in the bipolar samples. This analysis was performed for all CNVs and for CNVs that occurred only once in the data (i.e. singleton events). Only CNVs that overlapped with at least one gene, based on the hg18 genomic coordinates, were taken into account. Results for all CNVs and for singleton CNVs are presented in Table 21 and Table 22 respectively.

Table 21 Copy number variants disrupting genes

	Cases (n=1697)	Controls (n=2806)	CNV per case	CNV per control	*p*
Del	102	210	0.06	0.07	0.97
Dup	269	469	0.16	0.17	0.77
Total	371	679	0.22	0.24	0.94

Table 22 Single copy number variants disrupting genes

	Cases (n=1697)	Controls (n=2806)	CNV per case	CNV per control	*p*
Del	44	87	0.03	0.03	0.86
Dup	85	187	0.05	0.07	0.99
Total	118	248	0.07	0.09	0.98

No statistically significant differences were observed in CNVs (single-occurrence or not) that disrupt genes.

Genes affected by CNVs that were found only in cases but not in controls, were also studied. A list of all genes disrupted by CNVs in cases that were not disrupted in any of the controls is presented in Table 38 (page 177 in the Appendix). The list comprises not only singleton CNVs (those found once in the dataset), but also CNVs found in more than one case, but not in controls. Where several CNVs mapped to the same region, they did not always fully overlap. The start and end base pair positions in Table 38 refer to the total region covered by such a cluster of CNVs.

The results with respect to the pathway analysis are presented in Table 23.

Table 23 Pathway analysis

	Brain-expressed genes *p*	**Circadian rhythm** *p*	**Neurotrophin signalling** *p*	**Axon guidance** *p*	**Long-term potentiation** *p*	**Long-term depression** *p*
Enrichment of pathway genes, relative to the whole genome	0.03	1	0.98	0.6	0.03	0.99
Enrichment of pathway genes, relative to all genic CNVs	0.01	1	0.96	0.61	0.03	0.99

The genes from the brain-expressed and long-term potentiation pathways exhibited an overrepresentation in the CNV bipolar data. As these statistically significant results were not corrected for multiple hypotheses testing, they should be treated with caution. After correcting for the 12 tests performed, no pathway showed enrichment in the CNV data.

4.2.6. Are there CNV loci occurring more often in bipolar cases than in controls?

CNVs occurring in cases were compared with those in controls. Although an overall increase of CNV burden in bipolar cases compared with controls was not observed, some individual CNVs were found more often in cases than in controls. Nevertheless, none were significantly associated after correcting for multiple hypothesis testing. Results of CNVs that exhibited an uncorrected for multiple testing nominal significance level ($\alpha \leq 0.05$) are presented in Table 24.

Table 24 Regions showing more CNVs in cases than controls at nominal significance $p \leq 0.05$

Locus	Start bp	End bp	Type	BD cases	Controls	Fisher's exact test	Gene
1q25.1	173,769,777	173,978,862	dup	5	1	0.03	*TNR*
9q31.1	104,826,097	104,885,068	del/dup	3	0	0.05	No genes
12p11.21	31,202,250	31,301,551	dup	19	16	0.03	*OVOS2*
18p11.21-11.1	14,694,694	15,092,421	dup	3	0	0.05	*ANKRD30B*
19p12	20,001,614	20,177,979	del/dup	5	1	0.03	*ZNF682, ZNF90, ZNF486*
19p12	24,013,968	24,295,825	dup	3	0	0.05	*ZNF254*

Locus refers to the chromosome band; Start bp and end bp provide the start base pair position and the end base pair position of the CNVs at the specified locus

4.2.7. Follow up of previously implicated CNVs in psychiatric disorders

4.2.7.1. CNVs implicated in autism spectrum disorders, mental retardation, epilepsy or schizophrenia

In Table 25 are summarised the main chromosomal regions reported to be associated with autism spectrum disorder, schizophrenia, mental retardation or epilepsy. The respective number of the observed CNVs in the bipolar disorder cases and in the controls is also included. Only loci reported in multiple studies or with strong statistical support from at least one large study are considered.

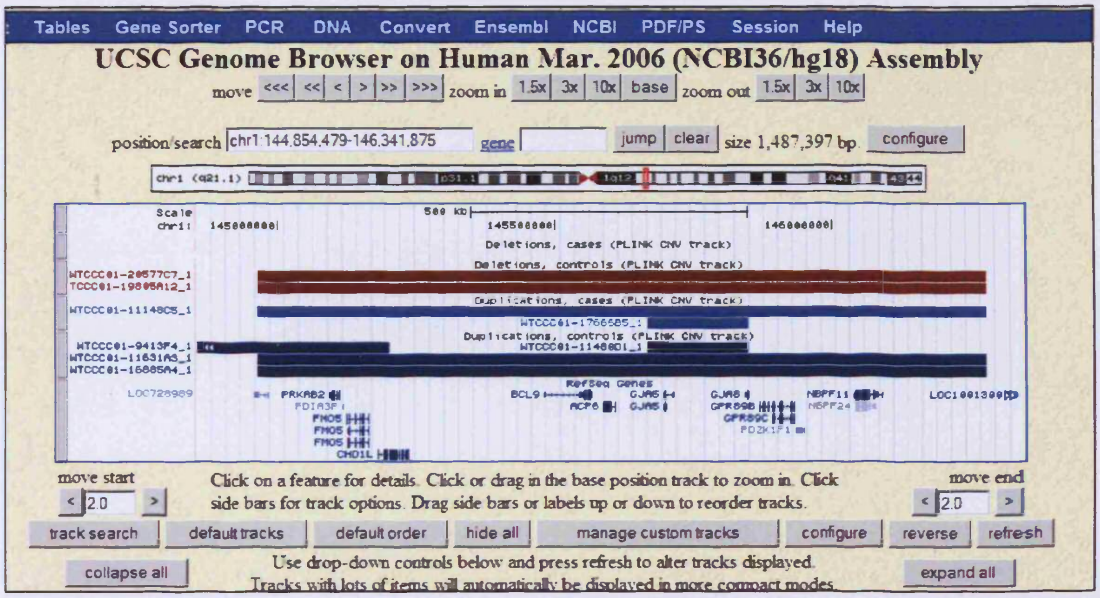
Table 25 CNVs implicated in schizophrenia, autism spectrum disorders/mental retardation (MR)/epilepsy and the corresponding number of CNVs in the current study

Locus	Position	Type	Article/ SZ	Article autism/MR/ epilepsy	BD Cases (n=1697)	Controls (n=2806)
1q21.1	144.9- 146.3	del dup	(International Schizophrenia Consortium 2008; Stefansson et al. 2008; Walsh et al. 2008)	(Brunetti-Pierri et al. 2008; Christiansen et al. 2004; Mefford et al. 2008)	1 dup	2 del; 2 dup
2p16.3 (*NRXN1*)	50-51.3	del	(Guilmatre et al. 2009; Ikeda et al. 2010; Kirov et al. 2008; Kirov et al. 2009b; Need et al. 2009; Rujescu et al. 2009; Walsh et al. 2008)	(Glessner et al. 2009; Guilmatre et al. 2009; Kim et al. 2008; Szatmari et al. 2007; Weiss et al. 2008)	0	4 del (various length)
7q34-36.1	145.6-148.7	del	(Friedman et al. 2008)		0	0
15q11-q13	21.2-26.2	dup	(Ingason et al. 2010; Kirov et al. 2009a)	(Christian et al. 2008; Glessner et al. 2009; Jacquemont et al. 2006; Marshall et al. 2008; Sebat et al. 2007)	0	0
15q11.2	20.2-20.8	del	(Kirov et al. 2009a; Stefansson et al. 2008)	(de Kovel et al.; Doornbos et al. 2009; Mefford et al. 2009)	3 del; 7 dup	14 del; 10 dup
15q13.3	29-30.30	del dup	(International Schizophrenia Consortium 2008; Kirov et al. 2009a; Stefansson et al. 2008; van Bon et al. 2009)	(Dibbens et al. 2009; Helbig et al. 2009; Miller et al. 2009; Pagnamenta et al. 2008; Sharp et al. 2008; van Bon et al. 2009)	2 dup	0
16p11.2	29.5-30.3	dup del	(Guilmatre et al. 2009; McCarthy et al. 2009; Walsh et al. 2008)	(Bijlsma et al. 2009; Kumar et al. 2008; Marshall et al. 2008; McCarthy et al. 2009; Mefford et al. 2008; Sebat et al. 2007; Weiss et al. 2008)	3 dup	3 del; 1 dup
16p13.1	15.0-16.2	dup del	(Ikeda et al. 2010; Ingason et al. 2009; International Schizophrenia Consortium 2008)	(de Kovel et al.; Hannes et al. 2009; Mefford et al. 2009; Ullmann et al. 2007)	2 dup	5 dup; 1 del
17p12	14.0-15.4	del	(Kirov et al. 2009a)		1 dup	0
22q11.2	17.4-19.8	del dup	Multiple publications	(Christian et al. 2008; Guilmatre et al. 2009; Marshall et al. 2008; Szatmari et al. 2007)	0	8 dup

The phenotypic data of the bipolar cases found to harbour schizophrenia implicated CNVs are presented below:

At locus **1q21.1** (Figure 28), a duplication has been detected in one bipolar case (6014-1). 6014-1 suffers from bipolar I disorder with psychotic symptoms. The father of the patient also has bipolar I disorder. More information about the segregation of this CNV in the 6014 family is presented in section 4.2.7.1.1 (page 115).

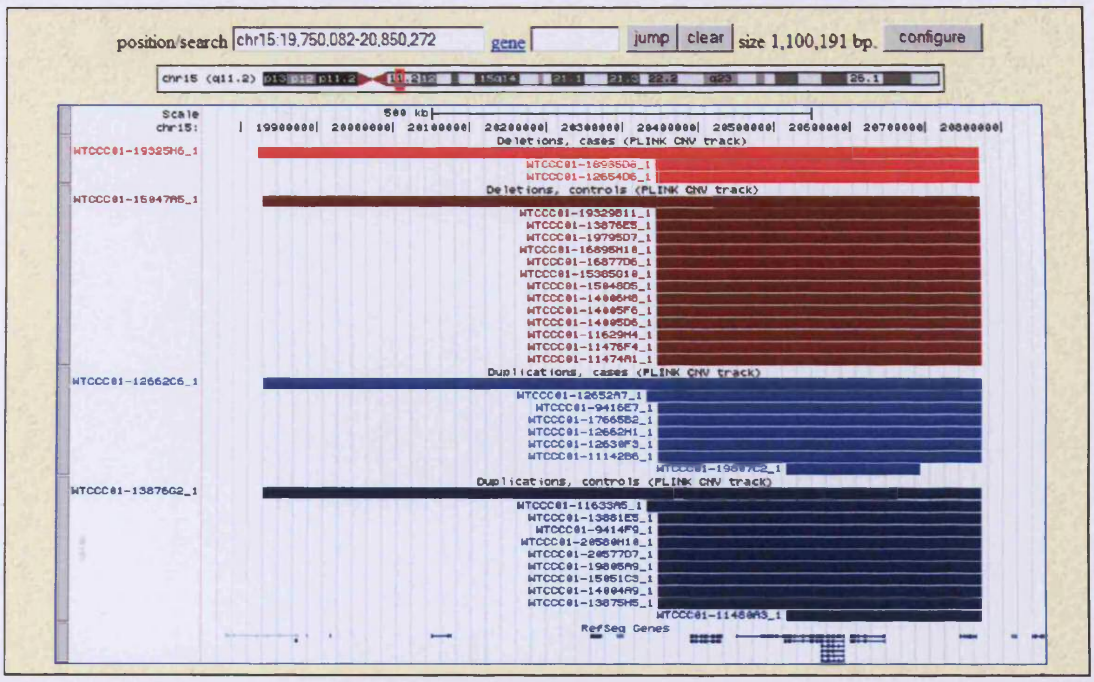
Figure 28 Locus 1q21.1



UCSC Genome Browser on Human Mar. 2006 (NCBI36/hg18) Assembly

Three bipolar cases and 14 controls had deletions at **15q11.2** locus (Figure 29). For two of the cases (10823-1 and 10720-1), phenotypic data were available. 10823-1 had schizoaffective disorder, bipolar type with an age of onset of 14 years of age. In addition, the case manifested panic disorder, obsessive-compulsive disorder and had postnatal depression episode two week after childbirth. Furthermore, the patient also manifested psychotic symptoms during episodes of affective disturbance (i.e. auditory hallucinations, delusions of influence, persecutory delusions and perplexity). The second case (10720-1) with a deletion at 15q11.2 was diagnosed with BPNOS. Borderline personality disorder and eating disorder were diagnosed as well. Auditory hallucinations were also observed. Additionally, more than 100 suicide attempts have been noted. The age of onset was at 22 years of age.

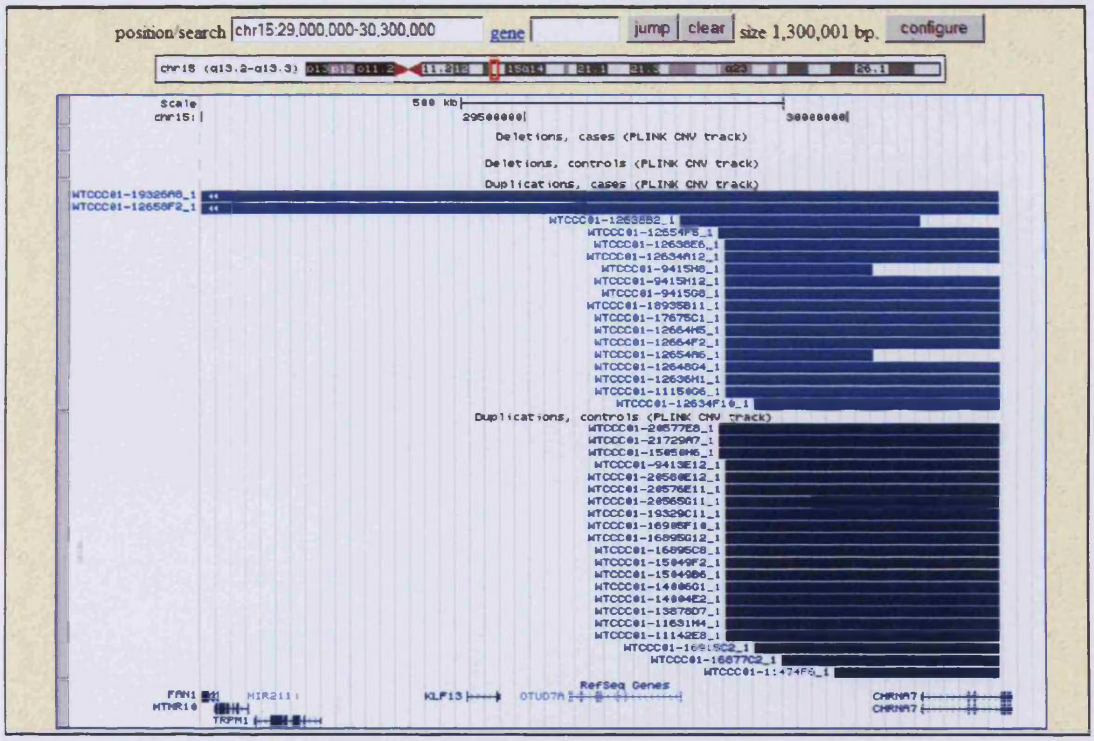
Figure 29 Locus 15q11.2



UCSC Genome Browser on Human Mar. 2006 (NCBI36/hg18) Assembly

No deletions at locus **15q13.3** (Figure 30) were observed in the sample, although two cases and no controls were found to harbour duplications in this region (Fisher's Exact Test, $p = 0.14$). Both of the bipolar cases with the duplication at this locus (10088-1 and 5344-1) were suffering from bipolar I disorder. No psychosis was observed. Interestingly, both had positive family history; 10088-1 had a daughter with bipolar I disorder, while 5344-1 had elder sister suffering from depression and a son suffering from manic depression.

Figure 30 Locus 15q13.3



UCSC Genome Browser on Human Mar. 2006 (NCBI36/hg18) Assembly

A region on **16p11.2** has been implicated in a range of neuropsychiatric phenotypes. Duplications at 16p11.2 were found in three cases and one control, resulting in a 5-fold increase in the frequency, although the result did not reach statistical significance (0.2% vs. 0.04%, Fisher's Exact test, $p = 0.15$). The deletion associated with this region was found in three controls and interestingly in none of the cases. It has to be noted that the 16p11.2 locus is covered by seven markers on the NspI array. Therefore it would have been filtered out by the applied quality control criterion of ≥ 10 consecutive markers on each array. To target this region, every individual who showed a CNV in this region, detected by the SNPs on the StyI array, was individually inspected with the Affymetrix® Genotyping console™ v2.1 software. Such a CNV was accepted if in addition to the SNPs on the StyI array, all markers on NspI array within this region also showed a deviation in the log2 ratios in the same direction. This region was then inspected in the CNV data that were received from the Wellcome Trust. It was reassuring that these duplications, detected with different methodology, were also observed by the Wellcome Trust.

The phenotypic information with respect to the three cases harbouring duplication at this locus is listed below:

8026-1- suffered from bipolar I disorder, with an age of onset at 27 years. The onset was four days after she gave birth to a child; the illness was characterised by severe episodes with affective features and perplexity. Episodes of puerperal psychosis followed the births of her two children;

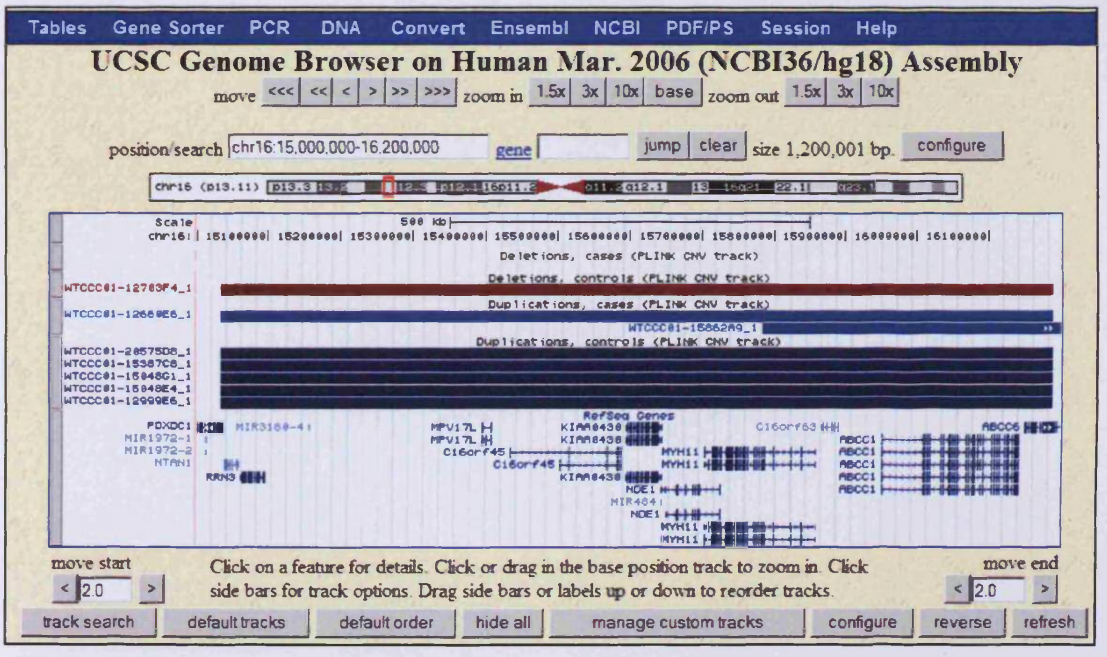
F400-4- bipolar I case with an age of onset at 16 years of age; the patient left school at 15 without taking any exams; no psychosis was observed. In total, 42 hospital admissions were noted;

In the 3rd patient (6023-1), the duplications had arisen *de novo* and was not present in the proband's father who also suffered from bipolar disorder. More information with regards to this family is provided in section 4.2.7.1.1 (page 115).

In addition, 16p11.2 locus will be discussed in more detail in the Discussion section, as the analysis from this PhD work has contributed to the evidence in favour of this locus in the susceptibility to bipolar disorder (McCarthy et al. 2009).

With respect to the **16p13.1** duplication (Figure 31), two cases and five controls harboured the CNV. Phenotypic information was available for one of the cases. 5475-1 was diagnosed with bipolar I disorder. The age of onset was at 35 years of age. Psychotic symptoms were also observed (visual hallucinations and persecutory delusions).

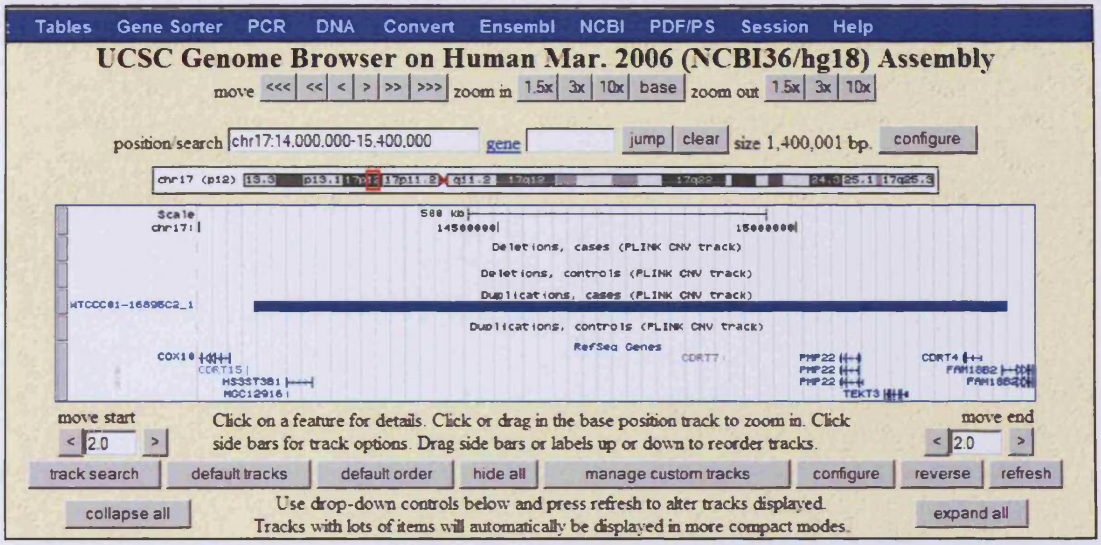
Figure 31 Locus 16p13.3



UCSC Genome Browser on Human Mar. 2006 (NCBI36/hg18) Assembly

With regard to **17p12** locus, in the bipolar sample, one case had duplication. No duplications or deletions were observed in the controls (Figure 32).

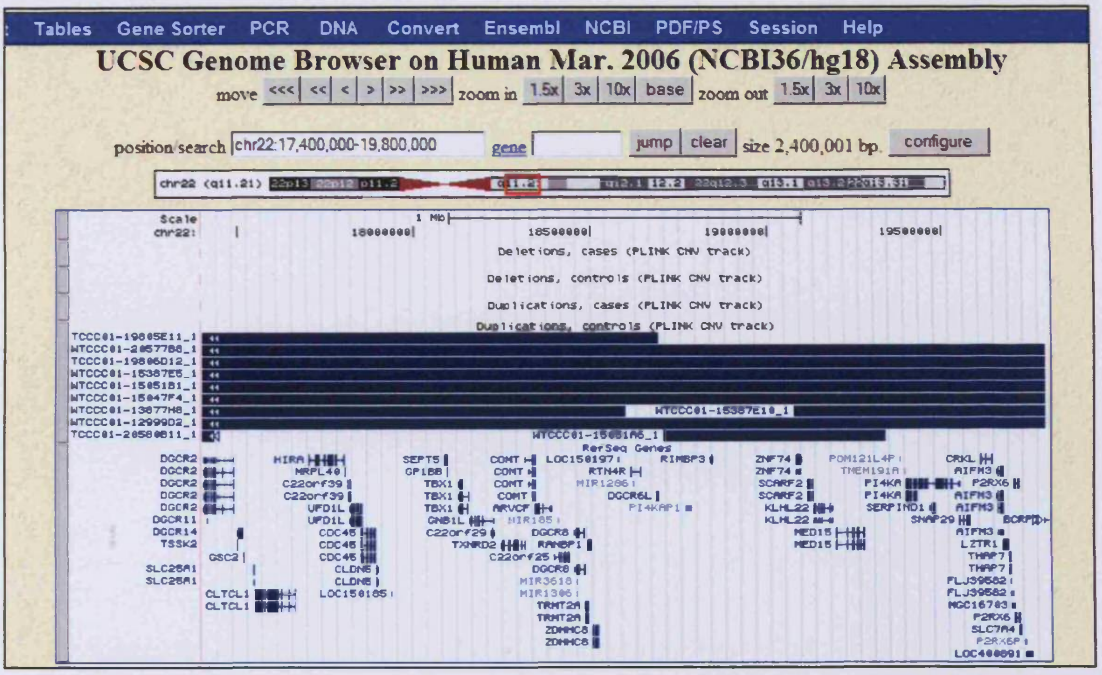
Figure 32 Locus 17p12



UCSC Genome Browser on Human Mar. 2006 (NCBI36/hg18) Assembly

No deletions were observed in the bipolar dataset at **22q11.2** locus. However, eight people harbouring a reciprocal duplication in this region (six of them spanning the full region) were found in the control dataset. All were part of the National Blood Service sample set (Figure 33).

Figure 33 Locus 22q11.2



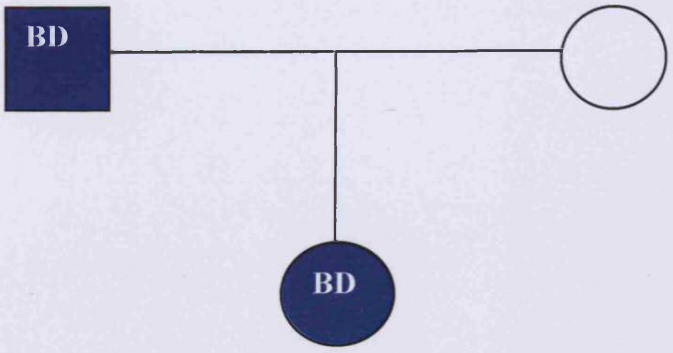
UCSC Genome Browser on Human Mar. 2006 (NCBI36/hg18) Assembly

4.2.7.1.1. Follow up with Illumina arrays of 1q21.1 and 16p11.2 in bipolar probands

For two of the individuals with CNVs that are reported in Table 25 (page 109), DNA from both parents was available. The probands and the parents were genotyped with an independent platform (Illumina HumanHap610 quad arrays) to validate the CNVs and to examine the transmission in the family.

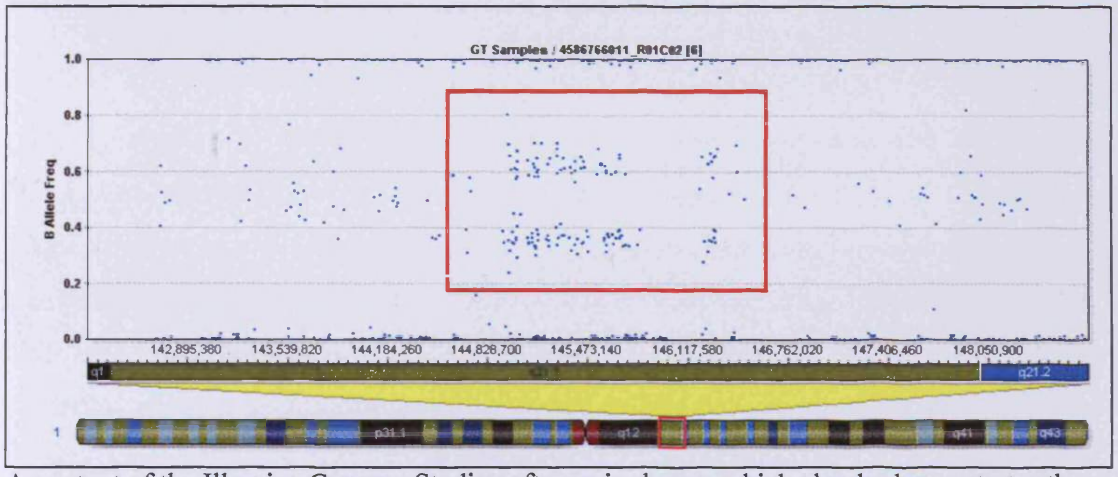
1q21.1 duplication

Figure 34 Transmission of 1q21.1 duplication in family 6014



The proband was affected with bipolar I disorder (Figure 34). The 1q21.1 duplication was confirmed in the proband (Figure 35) and was also found in the father (Figure 36), but not in the mother (Figure 37). Figure 35, Figure 36 and Figure 37 demonstrate the findings from the Illumina arrays produced using Illumina GenomeStudio Genotyping Module (http://www.illumina.com/). It is of interest, that the father who transmitted the duplication also suffers with bipolar I disorder.

Figure 35 A proband with 1q21.1 duplication



An output of the Illumina Genome Studio software is shown, which clearly demonstrates the presence of heterozygous SNPs of the type AAB, or ABB, i.e. two alleles from one parent and one from the other parent, for the duplicated interval

Figure 36 Father with identical 1q21.1 duplication

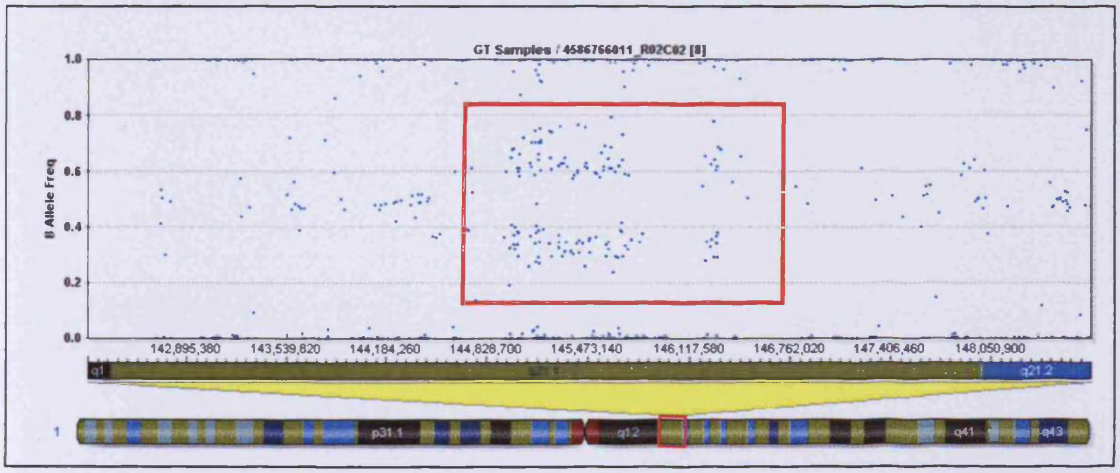
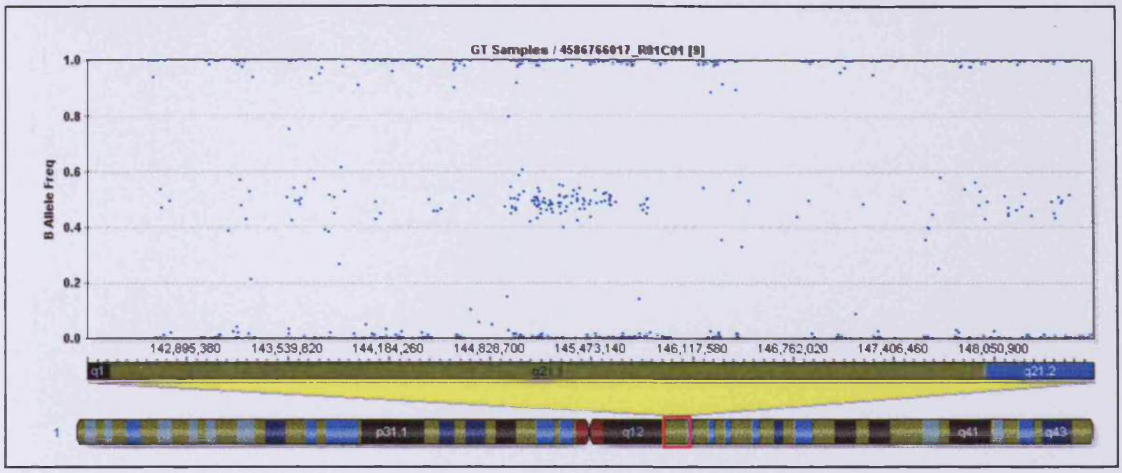
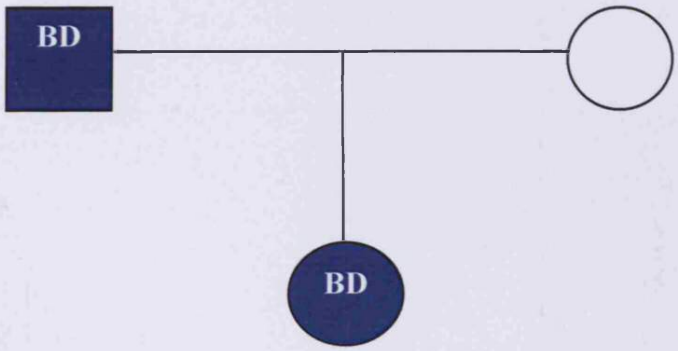


Figure 37 Mother of the proband



16p11.2 duplication

Figure 38 Transmission of 16p11.2.1 duplication in family 6023



In total, three bipolar cases were observed to harbour a duplication at the 16p11.2 locus. For one of these probands, DNA for both parents was available. The proband was affected with bipolar I disorder, as well as the father (Figure 38). The duplication was confirmed in the proband, but was not found in either parent. Paternity was confirmed from the SNP data on the array. SNP analysis showed that the duplication arose from the maternal genome.

The daughter and the father had a very similar presentation of illness, with severe manic and depressive episodes, but no psychotic features. The traces from the

Illumina arrays for the proband, father and mother are presented in Figure 39, Figure 40 and Figure 41.

Figure 39 A proband with *de novo* 16p11.2 duplication

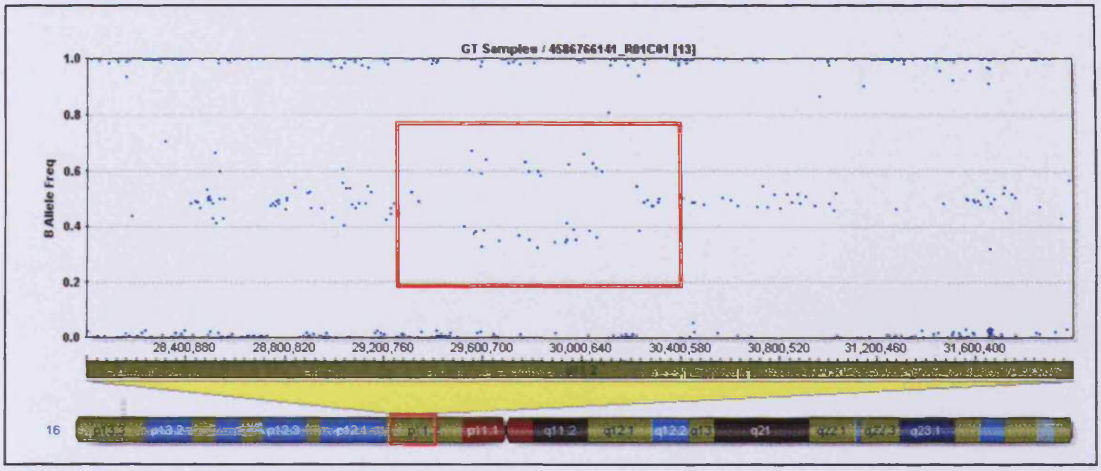


Figure 40 The father of a proband with 16p11.2 duplication

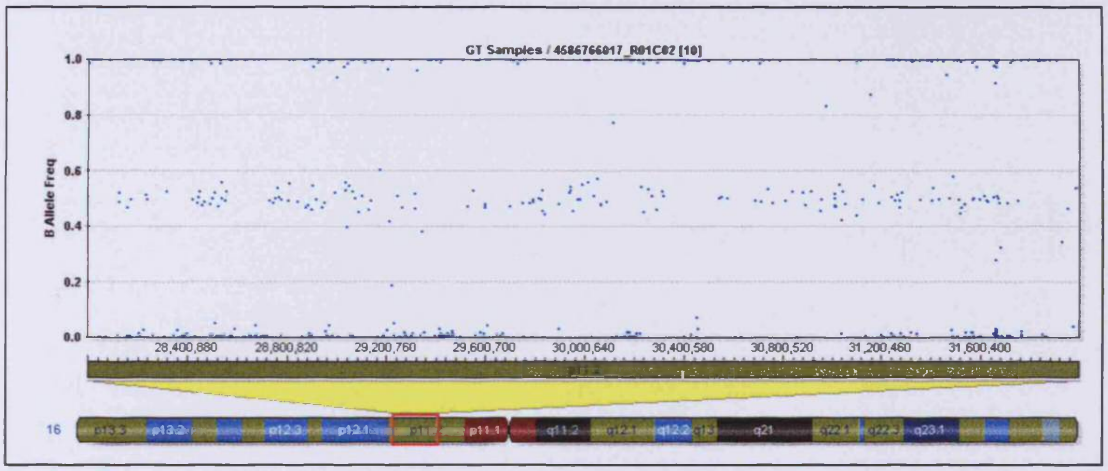
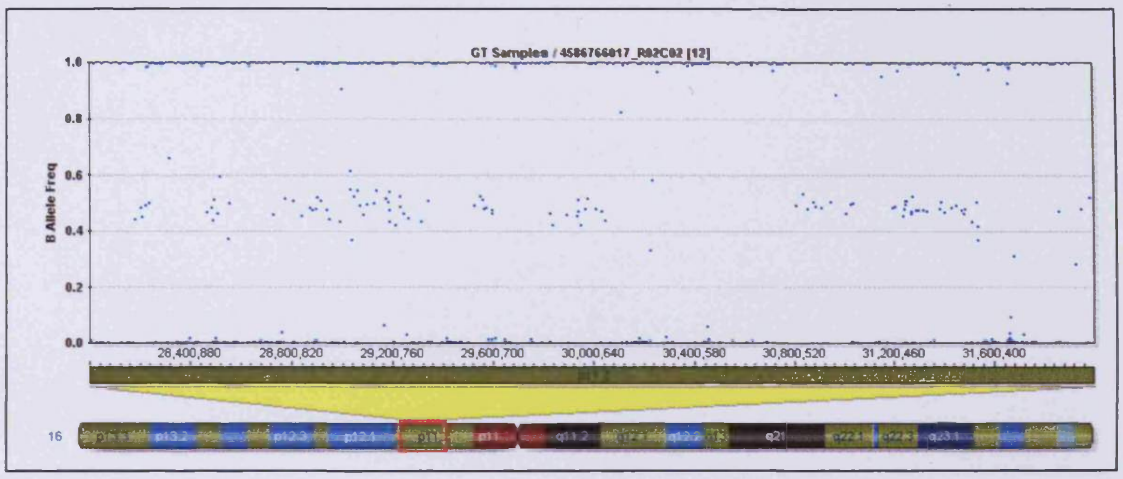


Figure 41 The healthy mother of a proband with 16p11.2 duplication



4.2.7.2. CNV previously implicated in bipolar disorder

CNV regions previously implicated in susceptibility to bipolar disorder were investigated for their occurrence in the WTCCC bipolar case data.

The following loci: **1p34.3**, **14q23.3** and **22q12.3**, affecting the *GRIK3*, *AKAP5* and *CACNG2* genes respectively, were examined (Wilson et al. 2006). No deletions or duplications were found in the bipolar dataset.

Locus **3q13.3**, reported to harbour a duplication in the *GSK3β* gene was also examined (Lachman et al. 2007). No CNVs were detected at this locus in the bipolar cases.

A study by Yang et al. examined the CNV structure in a large family with members affected with bipolar disorder (Yang et al. 2009). Certain regions have been found to harbour CNVs more often in the affected members of the family (6q27, 9q21.11, 12p13.31 and 15q11). No overrepresentation was observed in the WTCCC data in these loci.

4.2.8. Exploratory analysis of CNVs in bipolar subphenotypes

The bipolar disorder sample set consists of cases with bipolar disorder I (n=1451), bipolar II disorder (n=121), bipolar disorder not otherwise specified (BPNOS) (n=38) and schizoaffective disorder (n=87). These diagnoses are based on DSM-IV criteria. To reduce the heterogeneity in the sample, these sample sets were

compared separately to the controls. This analysis was not performed for BPNOS due to the limited number of cases.

In Table 26 are presented the results with regard to bipolar I disorder samples. The CNVs are split according to size.

Table 26 Global CNV burden in bipolar I cases

CNV type, size (kb)	Cases (n=1451)	Controls (n=2806)	CNV per case	CNV per control	case/control ratio	*p*
Del	270	632	0.186	0.225	0.83	**0.01**
Dup	444	901	0.306	0.321	0.95	0.45
Total	714	1533	0.492	0.546	0.90	**0.03**
Del<200	119	293	0.082	0.104	0.79	**0.03**
Dup<200	121	261	0.083	0.093	0.90	0.33
Total<200	240	554	0.165	0.197	0.84	**0.02**
Del 200-500	125	285	0.086	0.102	0.85	0.13
Dup 200-500	233	436	0.161	0.155	1.03	0.72
Total 200-500	358	721	0.247	0.257	0.96	0.55
Del 500-1000	20	34	0.014	0.012	1.14	0.67
Dup 500-1000	62	139	0.043	0.050	0.86	0.34
Total 500-1000	82	173	0.057	0.062	0.92	0.55
Del>1000	6	20	0.004	0.007	0.58	0.29
Dup>1000	28	65	0.019	0.023	0.83	0.44
Total>1000	34	85	0.023	0.030	0.77	0.21

Statistically significant results were observed with regard to the overall number of deletions and the total number of CNVs, when cases were compared with controls ($p = 0.01$ and $p = 0.03$ respectively). When the CNVs were split according to their size, it became clear that these results were predominantly driven by deletions of size ≤ 200 ($p = 0.03$). The rest of the comparisons between the cases and controls were not statistically significant. It has to be noted that relatively more CNVs were observed in controls than in cases, and for all of the above-mentioned significant results, the controls had more CNVs than cases. None of the p -values were corrected for multiple hypotheses testing.

Similar analyses were performed for the bipolar II disorder sample (Table 27). None of the comparisons between cases and controls were statistically significant.

Table 27 Global CNV burden in bipolar II cases

CNV type, size (kb)	Cases (n=121)	Controls (n=2806)	CNV per case	CNV per control	case/control ratio	*p*
Del	23	632	0.190	0.225	0.84	0.44
Dup	41	901	0.339	0.321	1.06	0.75
Total	64	1533	0.529	0.546	0.97	0.81
Del<200	11	293	0.091	0.104	0.87	0.67
Dup<200	13	261	0.107	0.093	1.16	0.64
Total<200	24	554	0.198	0.197	1.00	1.00
Del 200-500	9	285	0.074	0.102	0.73	0.39
Dupl200-500	21	436	0.174	0.155	1.12	0.64
Total 200-500	30	721	0.248	0.257	0.96	0.86
Del 500-1000	2	34	0.017	0.012	1.36	0.66
Dupl500-1000	4	139	0.033	0.050	0.67	0.54
Total 500-1000	6	173	0.050	0.062	0.80	0.71
Del>1000	1	20	0.008	0.007	1.16	1.00
Dupl1000	3	65	0.025	0.023	1.07	1.00
Total>1000	4	85	0.033	0.030	1.09	1.00

The results with regard to the schizoaffective disorder, bipolar type sample set are presented in Table 28.

Table 28 Global CNV burden in schizoaffective disorder, bipolar type

CNV type, size (kb)	Cases (n=87)	Controls (n=2806)	CNV per case	CNV per control	case/control ratio	*p*
Del	23	632	0.264	0.225	1.17	0.49
Dup	33	901	0.379	0.321	1.18	0.41
Total	56	1533	0.644	0.546	1.18	0.25
Del<200	10	293	0.115	0.104	1.10	0.88
Dup<200	17	261	0.195	0.093	2.10	**0.01**
Total<200	27	554	0.310	0.197	1.57	**0.02**
Del 200-500	9	285	0.103	0.102	1.02	1.00
Dup 200-500	11	436	0.126	0.155	0.81	0.59
Total 200-500	20	721	0.230	0.257	0.89	0.67
Del 500-1000	4	34	0.046	0.012	3.79	**0.03**
Dup 500-1000	4	139	0.046	0.050	0.93	1.00
Total 500-1000	8	173	0.092	0.062	1.49	0.27
Del>1000	0	20	0.000	0.007	0.00	0.66
Dup>1000	1	65	0.011	0.023	0.50	0.72
Total>1000	1	85	0.011	0.030	0.38	0.38

Statistically significant differences are shown in red

In the schizoaffective disorder, bipolar type three of the performed comparisons were found to be statistically significant (duplications ≤ 200 kb, $p = 0.01$; total CNVs ≤ 200 kb, $p = 0.02$ and deletions $200 - 500$ kb, $p = 0.03$). The p-values are reported without a correction for multiple hypothesis testing.

4.3. CNVs in schizophrenia sample

The schizophrenia sample set (n=440, DSM-IV criteria) was analysed for copy number variation. The number of schizophrenia cases analysed here and the number of schizophrenia cases, CNV analyses on which, we have previously published (n=471), differ (Kirov et al. 2009a). This is due to the fact, that for the analyses in the PhD, I excluded ~30 individuals due to diagnosis of schizoaffective disorder and then performed the analyses.

The schizophrenia sample set was compared with the control set with regard to the total number of CNVs and CNVs split according to size. The results from these comparisons are presented in Table 29. No statistically significant difference was observed with regard to the overall number of CNVs in cases as compared to controls. When the CNVs were split into different sizes, a statistically significant difference was observed with regard to CNVs of size ≥ 1 Mb. Such large deletions and duplications were two times more frequent in cases than in controls ($p = 0.005$). This effect was most prominent in large deletions, observed three times more frequently in cases when compared with controls ($p = 0.004$). After correction for multiple hypotheses testing (eight independent test, i.e. four size intervals for each deletion and duplication), the association was still statistically significant- $p = 0.03$. Although, large duplications did not exhibit a statistically significant difference between cases and controls, they were more frequently found in cases (~1.5 times more frequent than in controls).

Table 29 Global CNV burden in schizophrenia cases

CNV type, size (kb)	Cases (n=440)	Controls (n=2806)	CNV per case	CNV per control	case/control ratio	*p*
Del	99	612	0.225	0.218	1.03	0.780
Dup	141	893	0.320	0.318	1.01	0.960
Total	240	1505	0.545	0.536	1.02	0.840
Del<200	33	273	0.075	0.097	0.77	0.170
Dup<200	35	253	0.080	0.090	0.88	0.500
Total<200	68	526	0.155	0.187	0.82	0.130
Del 200-500	51	285	0.116	0.102	1.14	0.420
Dup 200-500	65	436	0.148	0.155	0.95	0.750
Total 200-500	116	721	0.264	0.257	1.03	0.800
Del 500-1000	5	34	0.011	0.012	0.94	1.000
Dup 500-1000	25	139	0.057	0.050	1.15	0.570
Total 500-1000	30	173	0.068	0.062	1.11	0.680
Del>1000	10	20	0.023	0.007	3.19	**0.004**
Dup>1000	16	65	0.036	0.023	1.57	0.110
Total>1000	26	85	0.059	0.030	1.95	**0.005**

Statistically significant differences are shown in red

As the only group of CNVs that was significantly found more often in schizophrenia cases when compared with controls was the CNVs of size ≥ 1 Mb, it is likely that a number of these CNVs could be pathogenic. These CNVs are presented in Table 30.

Table 30 CNVs of size ≥ 1 Mb observed in schizophrenia cases

Sample ID	Chr	Start [bp]	End [bp]	Size [kb]	Number observed in controls
Deletions					
19336A5	2	139,150,475	140,198,728	1048	0
20577A5	3	17,179,123	20,586,352	3407	0
19338E3	6	48,110,720	49,298,238	1188	0
19328D4	7	157,744,094	158,798,338	1054	0
19338A2	10	12,673,372	14,643,305	1970	0
19325A4	17	14,048,304	15,357,533	1309	0
19338F6	17	14,048,304	15,357,533	1309	0
19338B6	22	17,275,227	19,791,017	2516	0
20579B3	X	323,881	2,726,346	2402	0
19339D3	X	6,505,282	8,051,350	1546	3
Duplications					
19338A5	1	17,118,758	18,187,208	1068	0
20577F4	2	106,275,670	107,792,241	1517	0
20579B4	3	58,527,369	60,146,377	1619	2
19336D3	7	47,255,074	49,037,822	1783	0
19338B1	7	157,623,546	158,798,338	1175	0
19337A5	8	136,873,012	138,138,546	1266	0
19325E5	12	33,576,567	34,650,850	1074	4
19337F4	13	22,466,197	23,791,694	1325	2
19339B1	14	80,656,427	82,414,847	1758	0
19335E2	15	28,748,073	30,231,488	1483	0
19327C3	16	15,032,942	16,189,808	1157	6
19327F2	16	15,032,942	16,189,808	1157	6
20577B1	16	15,032,942	16,189,808	1157	6
19338A5	18	26,180,575	27,565,032	1384	0
19326C1	X	281,199	2,726,346	2445	4
19325H5	X	6,505,282	8,051,350	1546	6

Of the large deletions, presented in Table 30, the majority were observed only once. This makes drawing conclusions regarding their pathogenicity difficult.

One of these deletions was observed two times in the cases and was not observed in any of the controls, making it significantly associated with schizophrenia (Fisher's Exact Test $p = 0.018$). This deletion is affecting the **17p12** locus. One of the probands with the 17p12 deletion was part of a family with several members affected with neuropsychiatric conditions. One parent and three siblings (including the proband) had schizophrenia and another sibling had major depressive disorder. CNV analysis of all available family members, performed with Agilent arrays, showed that the affected parent, and three of the four siblings with a major psychiatric disorder, also carried the deletion. The absence of the deletion in one schizophrenic

member in this family is not compatible with the hypothesis that most of the risk in this family is attributable to a relatively highly penetrant CNV.

Of the large duplications, one was observed in more than one patient. Namely, a duplication at **16p13.1** locus was found in three cases and six controls. The combined burden of duplications in this locus did not exhibit a statistically significant association (0.7% in cases and 0.25% in controls, Fisher's exact test, $p = 0.11$).

Even though the CNV involvement in schizophrenia is not a study aim in this PhD, the aforementioned analyses were presented as one of my main aims was to compare bipolar disorder and schizophrenia with respect to copy number variation (which follows in the next section).

4.4. Comparison between CNVs in bipolar disorder and schizophrenia

The bipolar disorder cases were compared with a set of schizophrenia cases that was examined using the same methodology and was presented above. CNVs were classified into size categories and results are presented in Table 31. Compared with bipolar disorder, the schizophrenia sample exhibited a significant excess of large deletions (≥ 1Mb) ($p = 0.0009$), total number of large CNVs (≥ 1Mb) ($p = 0.0006$), and a trend for an excess of large duplications ($p = 0.053$). The observed p-values were not corrected for multiple testing. Those that would withstand a correction for multiple hypothesis testing were the excess of large (≥ 1Mb) deletions in schizophrenia compared with controls or bipolar cases.

Table 31 Comparisons of the rates of CNVs between controls, bipolar and schizophrenia cases

CNV type	Size range (kb)	CNV per control	BD/control ratio (p)	SZ/control ratio (p)	BD/SZ ratio (p)
Deletions	100-200	0.104	**0.801 (0.03)**	0.771	1.045
	200-500	0.102	0.864	1.141	0.758
	500-1000	0.012	1.264	0.938	1.348
	>1000	0.007	0.579	**3.189 (0.004)**	**0.181 (0.0009)**
Total deletions		0.225	**0.85 (0.01)**	1.032	0.825
Duplications	100-200	0.093	0.982	0.882	1.126
	200-500	0.155	1.051	0.951	1.105
	500-1000	0.050	0.868	1.147	0.757
	>1000	0.023	0.839	1.570	0.53 (0.053)
Total duplications		0.321	0.990	1.007	0.984
Total (Del+Dup)		0.546	0.930	1.017	0.918

Statistically significant differences are shown in red; the non-statistically significant p - values are not shown

4.5. Comparison between CNVs in bipolar disorder, schizophrenia and WTCCC non-psychiatric diseases

A cohort of 10,259 UK Caucasian individuals, affected with several non-psychiatric diseases was studied. These are coronary artery disease (CAD), Crohn's disease (CD), hypertension (HT), rheumatoid arthritis (RA), type 1 and type 2 diabetes (T1D and T2D). All of these sample sets have been examined in the WTCCC study (Wellcome Trust Case Control Consortium 2007) for associations of single nucleotide polymorphisms with these diseases. The bipolar cases and controls were part of the WTCCC project and all three sample sets were genotyped in the same pipeline along with CAD, CD, RA, HT, T1D and T2D including the schizophrenia samples (O'Donovan et al. 2008; Wellcome Trust Case Control Consortium 2007). This allowed comparing the CNV calling performed in this PhD work with independent calling performed by the WTCCC. Initially, I checked if the CNVs of size ≥ 1 Mb observed in the WTCCC study corresponded to those detected in this PhD work. These results were already presented in section 4.1.5 (page 101; correspondence rate between the two studies-93%). In addition, I checked the correspondence for loci implicated in other psychiatric disorders, in the bipolar

disorder sample and controls. The correspondence rate between the calls for these large CNVs was 100% (36 relevant CNVs called by both teams). This permits a conclusion that these CNVs, even though called using different analytical methods, can be compared between the separate phenotypic groups.

4.5.1. Comparison with respect to loci, previously implicated in schizophrenia

Loci implicated to increase the susceptibility to schizophrenia were chosen for further examination, as these were already analysed in this PhD work. The reason for choosing these loci was that they had received replicated evidence from multiple large studies, or from reviews on these CNVs.

In Table 32 are presented the frequencies of the studied loci implicated previously in schizophrenia, in bipolar disorder, schizophrenia and WTCCC non-psychiatric phenotypes. The corresponding significance levels derived from comparison between bipolar cases, schizophrenia, WTCCC controls (1958 cohort and NBS, analysed in this PhD work) with WTCCC non-psychiatric phenotypes are also presented. The WTCCC controls (2806 individuals, analysed in this PhD work) and the new dataset of 10,259 of non-psychiatric controls are not statistically different for any of the loci, while schizophrenia cases show a strong trend for an increased rate of these loci when compared with the new independent set of controls. There is no increased rate of these CNVs in the bipolar cases when compared with the new controls. There was a trend for overrepresentation only for one locus-16p11.2.

Table 32 Loci implicated in schizophrenia and the corresponding rate in BD, SZ and WTCCC non-psychiatric phenotypes

CNV	Rate				*p* (based on comparison with the rate in WTCCC phenotypes)			Review papers used
	BD	SZ	WTCCC controls	WTCCC phenotypes	BD	SZ	WTCCC controls	
1q21.1 del	0/1650 (0%)	17/7918 (0.2%)	2/2806 (0.071%)	1/10259 (0.01%)	1	$7.8x10^{-6}$	0.23	(International Schizophrenia Consortium 2008; Kirov et al. 2009a; Stefansson et al. 2008)
2p16.3 del *NRXN1*	0/1650 (0%)	14/8798 (0.16%)	2/2806 (0.071%)	7/10259 (0.068%)	0.6	0.08	0.96	(Kirov et al. 2009b; Rujescu et al. 2009)
15q11.2 del	4/1650 (0.242%)	47/7918 (0.59%)	14/2806 (0.499%)	40/10259 (0.39%)	0.51	0.05	0.53	(International Schizophrenia Consortium 2008; Kirov et al. 2009a; Stefansson et al. 2008)
15q13.3 del	0/1650 (0%)	15/7413 (0.2%)	0/2806 (0%)	4/10259 (0.039%)	1	0.002	0.66	(International Schizophrenia Consortium 2008; Kirov et al. 2009a; Stefansson et al. 2008)
16p11.2 dup	3/1650 (0.181%)	26/8590 (0.3%)	1/2806 (0.036%)	4/10259 (0.039%)	0.06	$4x10^{-6}$	0.94	(McCarthy et al. 2009)
16p13.1 dup	2/1650 (0.121%)	23/6920 (0.33%)	5/2806 (0.178%)	25/10259 (0.24%)	0.57	0.3	0.67	(Ingason et al. 2009; International Schizophrenia Consortium 2008; Kirov et al. 2009a)
17p12 del	0/1650 (0%)	8/5089 (0.16%)	0/2806 (0%)	7/10259 (0.068%)	0.6	0.11	0.35	(International Schizophrenia Consortium 2008; Kirov et al. 2009a; Stefansson et al. 2008)
22q11.2 del	0/1650 (0%)	18/7038 (0.26%)	0/2806 (0%)	0/10259 (0%)	NA	$9.5x10^{-8}$	N/A	(International Schizophrenia Consortium 2008; Kirov et al. 2009a; Stefansson et al. 2008)

Rate in SZ given for comparison purposes and based on the referenced articles, excluding the samples that overlap between the studies; Rate in BD cases, based on the WTCCC Sanger study; Rate in WTCCC controls (1958 cohort and NBS)-based on the PhD study for comparison purposes

The total number of implicated CNVs in schizophrenia in each of the phenotypes comprising the WTCCC study is presented in Table 33. For comparison,

the corresponding numbers in the cases affected with schizophrenia, that were genotyped and analysed in the same pipeline as the WTCCC study (Kirov et al. 2009a) and the schizophrenia cases and controls from the ISC study are also shown (International Schizophrenia Consortium 2008). When the total rate of previously implicated in schizophrenia loci was compared between bipolar cases and the WTCCC non-neuropsychiatric controls, no statistically significant difference was observed ($p = 0.19$). In contrast, when the combined sample of Cardiff and ISC schizophrenia cases was compared with the WTCCC non-psychiatric controls (3862 SZ cases with 94 CNVs vs. 10,259 non-psychiatric controls with 93 CNVs), a highly statistically significant difference was observed- $p = 5 \times 10^{-11}$.

Table 33 Rate of all combined CNVs implicated in schizophrenia in the separate cohorts that comprise the WTCCC study

Phenotype	Number individuals	Number CNVs	Frequency (%)
BD	1650	9	0.54
Cardiff: SZ cases	471	15	3.18
ISC: SZ cases	3391	79	2.33
ISC: controls	3181	22	0.69
NBS+1958 Cohort	2777	26	0.94
RA	1374	9	0.65
CD	1450	13	0.90
HT	1864	16	0.86
T1D	1903	18	0.95
T2D	1813	22	1.21
CAD	1855	15	0.81

Numbers in the International Schizophrenia Consortium study (ISC) are also given for comparison purposes

4.5.2. Comparison with respect to loci ≥1 Mb

In addition, this data set was also examined for CNVs ≥ 1 Mb and found in < 1% of the sample. The results with respect to this analysis are presented in Table 34. It is of note that bipolar disorder and rheumatoid arthritis have lower rate of this type of CNVs compared to the controls. None of the rates in the other phenotypes are significantly different from those in the 2777 controls, using 2-tailed tests with 10,000

permutations (that is excluding schizophrenia). The rate in schizophrenia was derived from Kirov et al. (Kirov et al. 2009a).

Table 34 Rate of deletions and duplications >1Mb in size in the WTCCC samples

Phenotype	N individuals	Rate of del per person	Rate of dup per person	Rate of CNV per person
controls	2777	0.007	0.020	0.027
BD	1650	0.004	0.014	0.018
CAD	1855	0.009	0.025	0.033
CD	1450	0.01	0.021	0.032
HT	1864	0.011	0.018	0.029
RA	1374	0.005	0.012	0.017
T1D	1903	0.011	0.021	0.032
T2D	1813	0.009	0.028	0.037
SZ	471	0.028	0.040	0.068

When bipolar cases were compared with the 10,259 non-psychiatric controls with respect to the cumulative burden of CNVs ≥ 1 Mb, statistically significant differences were observed with respect to deletions and total number of CNVs ≥ 1 Mb (Table 35). These differences were in the opposite direction than expected, i.e. the rate in the control sample was increased.

Table 35 Burden with respect to CNVs ≥ 1 Mb in bipolar disorder

Type CNV	BD N=1650	Rate BD	WTCCC phenotypes N=10,259	Rate WTCCC phenotypes	*p*
del	6	0.004	95	0.009	0.02*
dup	23	0.014	218	0.02	0.06
total	29	0.018	313	0.03	0.003*

p -values derived using two-tailed Fisher's exact test

*This significant result is for fewer CNVs in cases than controls

Schizophrenia cases were also compared with the 10,259 non-psychiatric controls. All types of CNVs ≥ 1 Mb were overrepresented in the schizophrenia cases (Table 36).

Table 36 Burden with respect to CNVs ≥ 1 Mb in schizophrenia

Type CNV	SZ N=471	Rate SZ	WTCCC phenotypes N=10,259	Rate WTCCC phenotypes	*p*
del	13	0.03	95	0.009	0.001
dup	19	0.04	218	0.02	0.02
total	32	0.07	313	0.03	0.0001

p -values derived using two-tailed Fisher's exact test

5. Discussion

In the final chapter of this PhD, I will discuss the methodological issues that I encountered whilst performing the analyses of the possible involvement of copy number variation in bipolar disorder. Then I will discuss the results and will put them in context of what is known with respect to copy number variation in neuropsychiatric disorders in general, and bipolar disorder in particular. In addition, I will consider what the results from studying the copy number variation in bipolar disorder tell us and finally will consider possible future strategies.

5.1. Main objectives

Recent studies have suggested that copy number variation in the human genome is extensive and may play an important role in susceptibility to disease. It has been shown that CNVs may confer susceptibility to infectious and complex diseases and play a role in adaptation to environment (Armengol et al. 2009; Flint et al. 1986; Gonzalez et al. 2005). Theoretically, relative to SNPs, large CNVs are more likely to induce phenotypic effects by altering genomic functionality (Cooper et al. 2007). In addition, it has been shown that CNVs could alter gene expression within and flanking the CNVs (Stranger et al. 2007; Yang et al. 2009). Furthermore, CNVs in general have been shown to have a possible causal role in severe neurodevelopmental syndromes and schizophrenia (Christian et al. 2008; Cook and Scherer 2008; Kirov 2010; Marshall et al. 2008; Sebat et al. 2009; Sharp et al. 2008). In addition, some specific copy number variants have been implicated as susceptibility factors for schizophrenia (Ingason et al. 2010; Ingason et al. 2009; International Schizophrenia Consortium 2008; Kirov et al. 2009a; Kirov et al. 2009b; McCarthy et al. 2009; Rujescu et al. 2009; Stefansson et al. 2008; Walsh et al. 2008).

Based on studying the genetic factors predisposing to schizophrenia and bipolar disorder (e.g. SNPs), previously it has been shown that they might overlap, causing susceptibility to both of the disorders (Gottesman et al. 2010; Lichtenstein et al. 2009; Moskvina et al. 2009).

On the basis of the possible overlap in the genetic factors between bipolar disorder and schizophrenia and evidence of the possible association between

deletions/duplications and the propensity to schizophrenia, the primary hypothesis in this PhD work was that individuals with bipolar disorder could potentially exhibit a greater burden of CNVs and a greater burden with respect to specific CNVs that have been previously shown to increase risk for schizophrenia.

The potential involvement of CNVs in bipolar disorder has received little attention to date. Thus, the current PhD project is a formal attempt to try to shed some light on the relative involvement of CNVs in bipolar disorder. When this project was initiated (at the beginning of 2008), there were no genome-wide data published with respect to investigating the possible role of copy number variation in the susceptibility to bipolar disorder.

More specifically, the main objective of the PhD thesis was to determine if large (≥100 kb) and rare (found in ≤ 1% of the general population) copy number variants are associated with susceptibility to bipolar disorder. The possible involvement of some particular deletions and duplications as a factor for developing bipolar disorder was also examined.

A direct comparison between schizophrenia and bipolar disorder with respect to CNV structure was also attempted.

My role in the project was crucial as I saw the opportunity to infer CNVs using SNP intensity array data, outlined the project and suggested it to my PhD supervisors. Subsequently, I closely participated in every step of the project (DNA case sample preparation, genotyping quality control, CNV calling, CNV calling quality control, CNV validation, CNV analyses and interpretation).

5.2. Quality control of the data

For the CNV analysis, 1868 Caucasian bipolar disorder cases and 2938 controls were used. Overall, 1697 cases (91% of the initial 1868) and 2806 controls (96%) passed the quality control filtering for the CNV analysis and were used for further CNV examination.

DNA from the case and control samples was genotyped with the Affymetrix 500K Array set comprising two arrays: NspI and StyI. Using the intensity data from the SNP genotyping, deletions and duplications were inferred. A variety of quality control filtering steps was applied in an attempt to reduce the false-positive rate:

- analysing copy number variants found in $\leq 1\%$ of the sample;
- analysing only deletions and duplications ≥ 100 kb;
- a deletion or duplication was called only if it spanned ≥ 10 consecutive SNPs on each of the two arrays;
- samples with an interquartile range (IQR) ≥ 0.4 were excluded;
- subjects with ≥ 20 segments were omitted from any further analysis;
- CNVs with very low SNP density were also excluded (≤ 3 SNPs per 100 kb);
- only CNVs observed on both of the arrays were analysed further.

Array-based experimental approach offers efficient and cost-effective method for whole-genome study of copy number variation in the human genome. McMullan et al. performed a study in which three European diagnostic centers assessed the use of Affymetrix 500K SNP arrays for molecular karyotyping in patients with mental retardation. The study was an attempt to examine if the array could be used for reliable detection of rare and large CNVs (≥ 100 kb). It was concluded that the platform provided robust results and could be used for a routine postnatal diagnostics in a clinical setting (McMullan et al. 2009). Other studies have also successfully interrogated SNP data to detect CNVs using the Affymetrix 500 K array (Friedman et al. 2009; Gardina et al. 2008). This provided evidence that Affymetrix 500K array can be confidently used for the purposes of the current study.

The Affymetrix 500K array set is one of the earliest whole-genome SNP genotyping platforms and has relatively lower resolution for CNVs as compared to later platforms. SNPs on the Affymetrix GeneChip 500K, have a median spacing of 2.5 kb. However, due to the variable resolution across the genome, CNV detection has a lower limit of 10- 40 kb (Carter 2007). Platform comparisons between high-density bacterial artificial chromosomes, single-nucleotide and oligonucleotide microarrays have shown that the Affymetrix 500K array set has difficulties in detecting CNVs < 100 kb (Hehir-Kwa et al. 2007). Therefore, in a conservative approach, only copy number changes ≥ 100 kb were analysed. This relatively low resolution of the Affymetrix 500K array has much less impact on large CNVs (the focus of this study), but because of this limitation, the contribution of smaller CNVs (≤ 100 kb) to disease susceptibility was not tested.

It is of note that CNVs are associated with highly repetitive regions in the human genome, i.e. low copy repeats (Lupski and Stankiewicz 2005). Such regions in the early generation arrays were not interrogated as they created difficulties for the SNP genotyping and such SNP data failed to qualify for inclusion on earlier generation of the commercial arrays (Carter 2007; Cooper et al. 2008; McCarroll et al. 2008). Because the probes on the Affymetrix 500K array were not uniformly distributed and were sparse in repetitive regions in the human genome, such arrays have poor abilities to call common CNVs. Thus, only CNVs observed in $\leq 1\%$ of the data were examined. Nevertheless, it has been demonstrated that using early generation arrays, rare CNVs had little coverage bias and are captured relatively well (McCarroll et al. 2008).

Additionally, common CNVs were not considered as the copy-number detection is relative. That is, the CNV identification requires a comparison to a reference dataset (Scherer et al. 2007). In this PhD study, the case samples were compared to reference sets that comprised individuals in the range of 100 to 250. If a CNV is common and is found in both groups, then it will not be detected due to the smaller relative difference in the intensities between the two groups. This could potentially result in a decreased power to detect variants in highly polymorphic regions in the genome.

An important consideration that has to be taken into account when performing CNV analysis is that CNV studies are prone to batch effects, as they rely on differences in signal intensities between case and reference groups. Differences in the signal intensities could stem from DNA quality or from the actual processing of DNA in the laboratory. Batch effects could lead to spurious data and detecting false deletions or duplications. Initially, at the start of the project, I observed that when some individuals, always within one processing batch, were compared with a reference sample from another processing batch, the produced data comprised many CNVs (in the order of hundreds) per person. When the same batch was compared with a reference sample set from the same processing batch, the number of CNVs observed was much smaller (< 20 segments per person). As CNV detection relies on intensity data, when samples from separate processing batches were compared against each other, a large number of falsely-detected CNVs were observed. In this PhD work, this problem was overcome by comparing samples with a reference set from the same processing batches. The importance of such batch effects and the means how to solve

this problem are now well recognised in the scientific community, but these were largely unknown when this PhD was initiated. In this PhD study, I suggested a way of how to resolve this issue, although it has to be said that it is not always straightforward to get around the batch effect. For instance, Zhang et al. could not examine the specific regions previously found to be associated with schizophrenia in bipolar cases. This was due to a proximity of these loci to regions with plate batch effects (Zhang et al. 2008).

Part of the quality control methodology in the current PhD study included that a deletion or duplication was inferred only if it was observed on both of the arrays. This potentially introduced a tendency to underestimate the size of the CNVs as only the overlapping fragments between the NspI and StyI arrays were taken into account. However at the same time, in a way, this was a validation of the findings based on one of the array with the other. Additionally, a validation using separate platforms was also performed.

Generally, it is regarded that the quality of SNP genotyping exceeds the quality of the CNV calling methods. In a recent review, CNVs were referred to as "a headache to identify and classify" (Baker 2010). Therefore, validation is required to ensure that the CNV calling is reliable. For confirming some of the CNVs, inferred using the data from the Affymetrix arrays (25 CNVs in total), an Agilent CGH array was used. This array provided a genome-wide coverage with an average resolution of ~30-35kb. It has been previously shown that using the Agilent array CGH platform, high quality data with respect to copy number changes could be obtained (De Witte et al. 2006). In addition, two CNVs in two bipolar disorder subjects (specifically 1q21.1 duplication and 16p11.2 duplication) were genotyped using Illumina Human Hap610 quad arrays. All of the CNVs that were put through validation, were confirmed. This reflects the stringent QC criteria applied throughout this PhD work.

Further support that the applied methodology produces reproducible results was the comparison with data obtained from WTCCC. Wellcome Trust examined the same raw genotyping SNP data (1868 bipolar cases and 2938 control individuals), analysed in this thesis for CNVs, using different methods for CNV calling (Barnes et al. 2008; Olshen et al. 2004; Price et al. 2005). When 96 CNVs, ≥ 1Mb, detected with the methods in this PhD study were compared with the data from Wellcome Trust, 93% (90 out of 96), were concordant between the two studies. In addition, all of the specific loci which were examined in detail in this PhD were double-checked using

the data from WTCCC. With respect to these loci, 100% concordance of the results between the two studies was observed.

In summary, the excellent validation rate corresponded to the stringent QC criteria, which were applied throughout this PhD work. This provided evidence that the produced CNV data were robust and trustworthy.

5.3. Global burden of CNVs in cases and controls

When bipolar disorder cases were compared with healthy controls, the main observation was that the rate of CNVs was ***not*** increased in bipolar disorder as compared to controls ($p = 0.1$, Table 17, page 103), contrary to expectations. There was a statistically significant association in the opposite direction for deletions ($p = 0.01$). It has to be noted that this statistical significance would not withstand a correction for multiple hypothesis testing. As such, this result is very likely to be false positive. Nevertheless, the effect was coming from controls having relatively more deletions than cases (0.23 CNVs per control as opposed to 0.19 CNVs per case).

The controls having more CNVs could be explained by different source of DNA, e. g. cell lines, which could potentially introduce differential bias in the number of CNVs between the cases and controls. The DNA for all of the cases was extracted from blood samples. The control sample used in this PhD work comprises two separate control datasets- National Blood Donor (NBS) and 1958 birth cohort. That is, blood for the NBS and EBV-transformed cell lines for the 1958 cohort. The general assumption is that DNA extracted from cell lines could exhibit an increased mutation rate (Mohyuddin et al. 2004). Therefore, to check this assumption, the difference between the two groups was examined with respect to the number of CNVs per person. When the two control groups were compared against each other, no statistically significant difference with respect to the total number of CNVs was observed ($p = 0.7$), although when the two groups were compared with respect to deletions only, the 1958 controls had relatively more deletions than NBS subjects ($p = 0.03$). Furthermore, more duplications in the NBS sample were observed in comparison to the 1958 cohort ($p = 0.03$). Nevertheless, the significant results with respect to deletions and duplications will not withstand a correction for multiple

hypothesis testing. Therefore, it was concluded that the two control groups did not provide spurious results. In addition, WTCCC group has compared the two control datasets with respect to SNP data and the two groups were not differentially biased in terms of collection and sample preparation. Only a few significant differences were observed between the two control groups even though, they differed with respect to the sampled populations, DNA source, DNA processing and age. This finding justified the combining of the two control groups in the WTCCC SNP study (Wellcome Trust Case Control Consortium 2007). In a recent large study on common CNVs in 16,000 cases of eight common diseases and 3000 shared controls (the same control set used in this PhD work), a clear biological artefact was observed, i.e. systematic CNV differences between DNAs derived from blood and cell lines (Wellcome Trust Case Control Consortium et al. 2010). It was shown that these apply only to particular CNV loci in the human genome. These specific loci have been located in close proximity to the immunoglobulin genes, rearranged via the process of V(D)J recombination, which is inherent to the B cells (Rivera-Munoz et al. 2007). The B cells are the cells used for producing cell lines, when DNA extracted from blood originates from a mixture of white blood cells (Wellcome Trust Case Control Consortium et al. 2010). In the current study, it can not be estimated whether this effect was observed or not.

Since the initial finding of controls having relatively more deletions than cases ($p = 0.01$) was not highly significant in the context of multiple hypothesis testing, this could well be a chance finding. Therefore a conservative interpretation of no increase in burden of CNVs in bipolar cases rather than controls, may be more appropriate. In addition, the bipolar cases were compared with the 10,259 non-psychiatric controls (described in 3.8, page 88) with respect to global burden of deletions, duplications and total number of CNVs. No statistically significant differences were observed with regard to global burden (irrespective of size) of deletions, duplications or total number of CNVs ($p = 0.09$, $p = 0.84$ and $p = 0.22$ respectively, not presented in the Results chapter). This indicated that abovementioned conclusion could be correct.

Other CNV studies in bipolar disorder have shown very similar findings. The overall genetic burden of CNVs was not increased when cases were compared with

controls (Zhang et al. 2008). Furthermore, Yang et al. observed that the average number and size of CNVs that were found in affected individuals were not significantly different from CNVs in unaffected individuals in a large Amish pedigree ($p = 0.3$ and $p = 0.24$ respectively) (Yang et al. 2009).

Similar were the findings presented at the XVII World Congress of Psychiatric Genetics held in San Diego 2009. Dr. Shaun Purcell presented CNV analysis in a large bipolar disorder cohort (~1400 cases and 2000 controls). No increased CNV burden was observed in cases when compared with controls (Purcell 2009). A poster at the same congress presented analogous result with respect to an overall CNV burden, based on a CNV study of ~900 cases and ~900 controls (Priebe et al. 2009).

In a very recent analysis, studying deletions and duplications in the University College London bipolar disorder case-control sample (546 cases and 517 controls) no increase in the rate of copy number variants in cases was observed (McQuillin et al. 2010). When cases were compared with controls with respect to total burden of deletions and duplications (> 100 kb), the observed 2-sided statistical significance was $p = 0.036$ (not corrected for multiple comparisons), with cases having fewer CNV than controls. It is of great interest that in the University College London bipolar disorder case sample, similarly to the results presented in this PhD thesis, a lower rate of CNVs has been detected in the cases, in comparison to healthy controls (McQuillin et al. 2010). The finding of such lower rate in two independent studies suggests that this could indeed be a genuine finding. Further studies could provide evidence if this is the case.

Based on the data published/presented to date with respect to the global burden of large deletions and duplications in bipolar disorder, it could be concluded that the overall burden was not increased and it is not likely that it plays a role in the susceptibility to developing bipolar disorder.

Since the main finding in this PhD work, from studying the genomic burden of rare (≤1%) and large (≥100kb) variants in cases and controls, suggested that such CNVs were **not** involved in the susceptibility to bipolar disorder, one could pose the question whether relatively more **common** variants account for some of the susceptibility. Common CNVs have not been thus far associated with common diseases, such as schizophrenia. Generally, it is regarded that common CNVs are

involved in the natural variation between individuals, where rare variants could cause the development of certain disorders (Kooy 2010). Common CNVs were not explored in this PhD study, but it has to be noted that almost the same bipolar cases and controls have taken part in a large genome-wide association study focused on common copy number variation performed by Wellcome Trust. This study has provided evidence that common CNVs were not likely to contribute to developing common diseases, such as bipolar disorder (Wellcome Trust Case Control Consortium et al. 2010). It was observed that CNVs at three loci affecting genes were associated with several diseases. Namely, these were: the *IRGM* gene in Crohn's disease, the HLA for Crohn's disease, rheumatoid arthritis and type 1 diabetes and the *TSPAN8* gene for type 2 diabetes. Interestingly, these loci were not novel. They had already been implicated in these disorders on the basis of evidence coming from genome-wide SNP studies. These data suggest that the majority of common CNVs were well tagged by SNPs and have already been indirectly studied. The main conclusion from this large genome-wide study was that common CNVs are unlikely to make a great contribution to the genetic basis of common human diseases (Wellcome Trust Case Control Consortium et al. 2010).

The finding of **no** increase in the CNV burden in cases when compared with controls observed in this PhD work and a similar observation in a number of other CNVs studies in bipolar disorder is in the **opposite** direction to findings in schizophrenia studies where an increased CNV load has been detected (International Schizophrenia Consortium 2008; Kirov et al. 2009a; Stefansson et al. 2008; Walsh et al. 2008). It has to be noted that CNV burden analysis alone could not identify specific risk factors (as the CNVs affect multiple genes), but only suggest a potential CNV involvement in disease susceptibility. Nevertheless, the approach used here has been proven pivotal in pointing to the involvement of CNVs in schizophrenia and other disorders (Abrahams and Geschwind 2008; International Schizophrenia Consortium 2008; Stefansson et al. 2008; Walsh et al. 2008). A possible explanation of the finding that schizophrenia and bipolar disorder differ with respect to the CNV involvement of the pathogenesis is discussed in section 5.11 (page 164).

5.4. Global burden with respect to CNV size

When the observed CNVs were split into size categories and compared between cases and controls, apart from small deletions (≤200 kb), statistically significant differences were **not** observed for any of the other size ranges. Significantly fewer CNVs in cases as compared to controls were observed with respect to deletions ≤ 200 kb ($p = 0.03$). If this result was to be corrected for the number of the tests performed, it would not withstand a correction for multiple hypotheses testing, thus suggesting that this was a potentially chance finding.

The finding that the controls had more deletions in total than cases (discussed in section 5.3, $p = 0.01$) is likely to be driven by the number of small deletions. Even though the Affymetrix 500K array has been shown to have a lower limit for CNV detection of about 10-40 kb (Carter 2007), the small CNV (≤ 200 kb) category is the most difficult to call, hence, the most unreliable. Taking into account that this finding is not highly statistically significant and it would not withstand a correction for multiple hypotheses testing, it could be a chance finding. Thus, it could be conservatively interpreted as that cases showed no increase in the overall burden with respect to deletions ≤ 200 kb, as compared to controls.

These results are in agreement with a recent study in bipolar disorder, with respect to CNV size (McQuillin et al. 2010). In particular, no statistically significant observations were noted in the size ranges tested apart from for deletions with size range 200 – 500 kb which were observed more often in controls than cases ($p = 0.039$, not corrected for multiple hypothesis testing) (McQuillin et al. 2010).

This observation in bipolar disorder is in the opposite direction to findings in schizophrenia where very large CNVs (≥ 500 kb or ≥ 1 Mb) have been observed to be enriched in people with schizophrenia (International Schizophrenia Consortium 2008; Kirov et al. 2009a). Discussion of the observed differences between bipolar disorder and schizophrenia is presented in section 5.11 (page 164).

5.5. Singleton observations

While the analysis for this PhD work was been performed, the first systematic genome-wide CNV analysis in bipolar disorder was published (Zhang et al. 2008).

Similar to the main finding presented in this PhD, no increase in the total burden of CNVs in cases was observed when compared with controls.

The primary finding in Zhang et al. was an effect with respect to single deletions that were observed more often in cases when compared with controls (16.2% of the cases versus 12.3% in controls; $p = 0.007$). However, the data from this PhD work did not replicate the finding of a significant increase in singleton CNV events in bipolar cases compared with controls (Table 37).

Table 37 Proportion of samples carrying at least one singleton CNV in this study compared with Zhang et al.

Single CNV type	cases with a CNV (this study) [%]	controls with a CNV (this study) [%]	cases with a CNV (Zhang et al.) [%]	controls with a CNV (Zhang et al.) [%]
Del	6.0	6.2	16.2	12.3
Dup	7.1	8.3	19.7	19.1
Total	11.0	11.7	32.4	29.0

(Zhang et al. 2008); Del=deletions; dup=duplications

Zhang et al. used a higher resolution Affymetrix 6.0 array and had smaller sample size (1001 cases and 1034 controls) (Zhang et al. 2008). These factors could have led to the observation of a higher number of singleton CNVs, and could explain the differences with respect to the frequency of cases/controls with a singleton CNV (Table 37). It has to be acknowledged that the smaller the sample, the more likely is to observe an event only once in the data. In contrast, if the sample is very large, singleton events will be hardly ever observed unless there is an event that have a complete penetrance with bipolar disorder. It has to be noted that an overall reduction of singleton events has been observed in the current study when compared with Zhang et al. study.

In addition, the power of the current study at $\alpha = 0.05$ level to detect a statistically significant difference with the effect size observed by Zhang et al. ($\delta = 0.11$), is very high (95%). For comparison, the power to detect a significant difference with the aforementioned effect size with the sample size in Zhang et al. is 71% ($\alpha = 0.05$). Power will be lower if there is a true, but smaller increase in singleton CNV burden (i.e. smaller effect size), or if the stringent QC criteria resulted in detecting a smaller proportion of the relevant loci in this PhD analysis. With respect to very weak effects, no single study will be powerful enough to provide definitive results. Therefore in this context, a type II error in the present bipolar versus

control analysis could not be fully discounted. However, it should be noted that the findings of Zhang et al. were relatively weaker than those reported in a smaller-sized schizophrenia and autism studies (Guilmatre et al. 2009; Kirov et al. 2009a; Walsh et al. 2008; Zhang et al. 2008). It is possible that these results (in Zhang et al.) could represent type I error as they would not withstand a correction for multiple hypotheses testing concerning the CNV size, type and frequency.

Zhang et al. observed more pronounced effect with respect to single deletions in bipolar cases with early onset of mania (≤ 18 years of age, $p = 0.001$) (Zhang et al. 2008). At the XVII World Congress of Psychiatric Genetics held in San Diego (2009), a poster presented a CNV analysis of 882 patients with bipolar I disorder and 872 healthy controls using genome-wide SNP data (Priebe et al. 2009). The analysis showed that singleton deletions in patients were larger (in terms of average size) than in controls ($p = 0.014$). On average, single microduplications were also relatively larger in cases with an early onset when compared with controls ($p = 0.0048$). None of the performed tests showed statistically significant differences in CNV load between cases with onset ≥ 21 years of age when compared with controls. An effect with respect to singleton observations was observed, and similarly to the Zhang et al. study, this effect was more prominent in cases with an early age of onset. A statistically significant association was detected with respect to cases with early onset and duplications ($p = 0.00022$, after a Bonferroni correction for the performed 72 tests, $p = 0.016$) (Priebe et al. 2009). In addition, another CNV study in bipolar disorder, presented by Dr. Shaun Purcell at the same congress also showed evidence of possible involvement of CNVs in early onset bipolar disorder (Purcell 2009).

These three studies observed an effect of CNVs in early onset bipolar disorder. When the same criterion (age of onset of mania ≤ 18 years of age) was used in this PhD work, no statistically significant effect was detected. In the current bipolar study, there were only 65 cases with onset of mania ≤ 18 years. This sample set was relatively small to draw any definitive conclusions.

McQuillin et al. tested the hypothesis if single CNV observations were involved in the predisposition to bipolar disorder. In contrast to Zhang et al. and in agreement with the results presented in this PhD work, no statistically significant enrichment was noted in bipolar cases (McQuillin et al. 2010). In addition, McQuillin et al. compared the rate of singleton events in cases with early onset (≤ 18 years of

age) with the cases with a later first onset of mania. No statistically significant differences were observed between the two groups.

While additional studies are required to further explore the possible involvement of single CNVs in the susceptibility to bipolar disorder, the analysis presented in this PhD work and the findings of McQuillin et al., provide evidence against the contribution of singleton CNVs in developing bipolar disorder in the size range tested (≥100 kb).

5.6. CNVs disrupting genes

Studies of schizophrenia have noted an elevated burden of CNVs that disrupt genes in cases as compared to controls (International Schizophrenia Consortium 2008; Walsh et al. 2008). Walsh et al. detected a raised load of CNVs that specifically disrupt genes in affected individuals as compared to controls ($p = 0.012$, $OR = 2.79$, $95\%CI : 1.26 - 6.18$) (Walsh et al. 2008). Furthermore, the International Schizophrenia Consortium detected an increased number of genes affected by CNVs in cases when compared with controls (1.41-fold increase, $p = 2 \times 10^{-6}$) (International Schizophrenia Consortium 2008). Similar analysis was carried out in this PhD with respect to CNVs disrupting genes. No statistically significant difference was observed between the number of CNVs disrupting genes in cases when compared with controls.

Although, there was no evidence of an increased burden of disrupted genes in cases when compared with controls, even a single occurrence of a CNV in a disorder could potentially make an important contribution to the understanding of the disease pathogenesis and could lead to possible mechanisms/pathways that underlie the disorder (Sudhof 2008). Such an example has been the observation of a rare deletion in the *NRXN1* gene in probands with schizophrenia (Kirov et al. 2008; Kirov et al. 2009b; Rujescu et al. 2009).

With respect to bipolar disorder, no statistically significant differences were observed when singleton CNVs affecting genes were analysed.

If a particular CNV has been observed only in cases and in no controls, this could potentially be an indication that such a variant is likely to be pathogenic. Thus, an analysis was performed for CNVs intersecting genes found exclusively in the bipolar cases. Several potentially intriguing genes were affected. Examples included

neuroligin 1, neuregulin 3 and alpha-2 catenin. These genes have been previously associated with neuropsychiatric disorders (Chen et al. 2009; Chubykin et al. 2005; Dean and Dresbach 2006; Lesch et al. 2008), although no specific evidence was observed to support their contribution in the susceptibility to bipolar disorder. This comes from the fact, that it is not currently known how disruption in any of these genes could mechanistically translate into an illness. In addition, similar results were obtained when the number of genes disrupted in cases was analysed (i.e. no statistically significant differences at the 0.05 level).

The analysis with respect to overall genomic burden and analysis with respect to burden of CNVs disrupting genes did not point to possible genetic underpinnings of bipolar disorder. An approach which could potentially identify the involved genes and mechanisms in the disorder pathophysiology is pathway analysis.

5.7. Pathway analysis

It has been suggested that changes in many genes could lead to the expression of the bipolar disease phenotype. Such an effect can be explored by analysis that focuses on pathways rather than on single variants. This investigation could potentially reveal an insight into the genetic basis of the disease. The goal of such analysis is to determine if the susceptibility variants associated with bipolar disorder are clustered in particular biological pathways that are pertinent to the disease pathophysiology.

CNV data on its own do not provide answers to questions with respect to which biological processes are affected by the observed CNVs. The burden analysis that was performed in the PhD work does not help to identify genes that increase the susceptibility to the illness and overall does not provide resource for further research. How exactly a potential CNV burden translates into developing the disease is currently unknown. The main observation when the bipolar cases were compared with controls was that the CNV burden was not elevated in cases. Although, no increase was observed in the CNV load, it is possible that in cases, specific pathways are affected which are not disrupted in controls. Examining such pathways could indicate which biological processes are affected in cases. For instance, Walsh et al. analysed CNV data with respect to disrupted pathways and observed that CNVs in schizophrenia cases disproportionately disrupted genes belonging to pathways

considered to be important for brain development. These pathways included neuregulin ($p = 0.008$) and glutamate signalling ($p = 0.003$), synaptic long-term potentiation ($p = 0.0005$), synaptic long-term depression ($p = 0.017$) and axonal guidance signalling ($p = 0.015$) (Walsh et al. 2008). For comparison, an overrepresentation of any pathway was not observed when control CNV data were analysed. Based on these data, it was speculated that CNV disrupting genes from neurodevelopmental pathways, can lead to developing neuropsychiatric psychopathology (Walsh et al. 2008). Similar results were detected in bipolar disorder. Pathway analysis showed that genes affected by CNVs were disproportionately disrupted in cases in pathways categories such as "psychological disorders" ($p = 6.3 \times 10^{-6}$) and "learning behaviours" ($p = 8.29 \times 10^{-3}$). When controls were studied, genes involved in such pathways were not found to be disrupted by CNVs (Zhang et al. 2008) .

To address the question if genes in certain pathways are disproportionately more likely to be disrupted in cases when compared with controls, pathway analysis was performed. The studied pathways were: brain-expressed genes, long-term potentiation, long-term depression, neurotrophin signalling, axon guidance and circadian rhythm pathways. Reasoning why these pathways were studied in bipolar disorder is provided below.

CNVs have been shown to account for some of the susceptibility to neuropsychiatric disorders. One hypothesis of how the specific CNVs are exerting their function is by affecting brain-expressed genes. If this is true, then the CNVs could affect brain-functioning genes more often than compared with controls. In addition, it has been shown that genes participating in synaptic plasticity could be involved in susceptibility to bipolar disorder (Schloesser et al. 2008). Long-term potentiation and long-term depression are forms of synaptic plasticity. With long-term potentiation is indicated the long-lasting stimulation of transmission of a signal between two neurons which have been induced at the same time, and long-term depression meaning the opposite process. A correlation has been observed between long-term potentiation/depression and learning and memory (Cooke and Bliss 2006). It is of note that in CNV bipolar data a statistically significant association has been observed with "learning behaviours" pathway (Zhang et al. 2008). With respect to the other studied nervous pathway (neurotrophin signalling), it has been previously

observed that lithium, one of the drugs used to treat bipolar disorder, upregulates genes from the neurotrophin signalling pathway (i.e. brain-derived neurotrophic factor (*BDNF*), nerve growth factor, neurotrophin-3 (*NT3*), etc.). In addition, it has been shown that lithium exerts its function by acting upon receptors to these growth factors in animal brain (Young 2009). The reasoning for studying the axon-guidance pathway was that neuregulin (*NRG1*), a psychosis susceptibility gene, has been known to influence axon guidance and neuronal migration (McIntosh et al. 2008). In the case of circadian rhythm pathway, it has been shown that genes, known to influence regulation of circadian rhythms (i.e. *TIMELESS*, *CLOCK*) can be contributory variants to susceptibility to bipolar disorder (Harvey 2008; Mansour et al. 2005).

The results with respect to the performed pathway analysis showed an overrepresentation of brain-expressed genes and genes in the long-term potentiation pathway to be disrupted by CNVs more often in cases than controls (Table 23, page 106). After correcting the observed significance levels for the number of the performed tests (i.e. 12), no statistically significant results were observed. Thus, it was concluded that the tested pathways were not enriched in the CNV bipolar data.

This observation does not support the findings of the Zhang et al. study, where a statistically significant observation was noted with respect to "psychological disorders" and "learning behaviours" pathways. However, Raychaudhuri et al. argued that in general some of the performed pathway analyses have the following shortcomings: a) do not meticulously compare the rate of CNV events in the cases with the rate in controls; b) as the studies compare rates of disrupted genes, not events, it is not taken into consideration that multiple genes could contribute to one event or that single genes could be disrupted by multiple CNVs. To take the above-mentioned considerations into account, a new method for pathway analysis, integrated in PLINK, was developed (Raychaudhuri et al. 2010). The method for comparison in this PhD work has made use of this PLINK test.

One possible explanation of why the results obtained in this PhD are different with the ones presented by Zhang et al., with respect to pathway analysis could stem from the fact that it has been observed that genes with common function can cluster together on the human genome and a CNV affecting this region could therefore show to affect an entire pathway (Raychaudhuri et al. 2010). It was noted that 11 of the 16 deleted psychological disorder genes in the Zhang et al. data reside on 22q11.2 (CNV observed in two individuals). As this region has been previously implicated in

schizophrenia, these genes could have been annotated as psychological disorder genes. It was shown that when the individuals with 22q11.2 deletion were removed from analysis, the psychological disorder pathway was no longer statistically significantly associated with bipolar disorder (Raychaudhuri et al. 2010). Thus, great care should be taken when interpreting results derived using pathway analyses.

In conclusion, although, the pathways analysis did not reveal significant findings and based on the fact that only a small number of key pathways were explored, further research is necessary to elucidate which pathways have biological relevance for developing bipolar disease.

5.8. Exploratory analysis of CNVs in bipolar subphenotypes

The WTCCC bipolar disorder sample set comprised bipolar I and II disorder, bipolar disorder not otherwise specified and schizoaffective disorder, bipolar type. The issue of heterogeneity of large neuropsychiatric diagnostic categories has been previously addressed and it is logical to hypothesise that distinct phenotypic subgroups may have distinct genetic background (Craddock et al. 2010). In an attempt to reduce the heterogeneity in the sample, the different subphenotypes were analysed separately.

When bipolar I disorder cases were compared with controls with regard to CNV size, statistically significant associations were observed with respect to the overall number of deletions and the total number of CNVs (Table 26, page 120). This result was in the opposite direction than expected, i.e. controls had more CNVs than cases. These results were similar to the ones observed when the bipolar samples was analysed as a whole (Table 17, page 104). When the CNVs were split according to their size, it became clear that the significant findings in the bipolar I sample were mainly due to deletions of size ≤ 200 kb. This size range represented the smallest CNVs that the Affymetrix 500K array could detect. Even though the quality control metrics that were applied in this PhD were quite stringent, this CNV size category was the most likely to produce false-positive results.

All of the other comparisons in the separate size ranges were not significantly different. As the results were not corrected for multiple hypotheses testing, the observation that the controls exhibited more CNVs than the bipolar I disorder cases is likely to be a chance finding. The conservative interpretation of this finding should be

that bipolar I cases are not very likely to be different than controls with respect to CNVs.

Similar analysis was performed for bipolar II disorder cases and none of the comparisons were statistically significant (Table 27, page 121).

When schizoaffective disorder bipolar type cases were compared with controls (Table 28, page 121), a statistically significant enrichment in the cases was found for duplications of size ≤ 200 kb ($p = 0.01$) and for deletions in the range of 200 – 500 kb ($p = 0.03$). If multiple hypothesis correction is applied, these results will be no longer statistically significat. Therefore, the conservative interpretation of these results would be that it is likely to be a false-positive finding.

As the sample was split into categories according to the expressed phenotype, the sample size decreased. This poses the question if a sample size which for example comprises 87 individuals (this was the case with the schizoaffective disorder bipolar type cases), is large enough in terms of power to observe any CNVs that have inherently low frequencies. It is unlikely that such number of individuals is sufficient.

In summary, the data showed that large and rare CNVs did not play a particularly prominent role in one or more of the studied phenotypes. However, this does not rule out smaller CNVs or some specific variants that could increase the susceptibility to a particular type of bipolar disorder.

5.9. Specific CNVs in bipolar disorder

Although an overall increase of CNV burden in cases was not observed when compared with controls, some individual CNV were observed more frequently in cases than in controls. It has to be said that none of the comparisons showed statistically significant experiment-wise association (after multiple hypotheses correction). Therefore, the data did not exhibit a statistically increased CNV burden in cases of some specific CNVs. From this analysis, no further evidence can be presented to show if these loci are true findings or not.

Genes disrupted by particular CNVs found more often in cases than controls could point to specific pathways involved in the susceptibility to bipolar disorder. The genes disrupted by the CNVs with an uncorrected nominal significance level $\alpha \leq 0.05$ included tenascin R (*TNR*), ovostatin 2 (*OVOS2*), ankyrin repeat domain 30B

(*ANKRD30B*) and zinc finger proteins (*ZNF68, ZNF90, ZNF486* and *ZNF254*). The results with respect to this analysis can be found in Table 24 (page 107).

Tenascin R protein is exclusively expressed in the nervous system (Woodworth et al. 2004). Jäger et al. performed serological analysis of different tumour types and observed that ANKRD30B mRNA in addition to breast and testes is also expressed in brain (Jager et al. 2001). The zinc finger proteins are uncharacterised proteins and typically function as DNA binding proteins and could have a possible role in gene expression. Interestingly, a zinc finger protein (*ZNF804A*) has been associated with schizophrenia in a large meta-study (O'Donovan et al. 2008).

In conclusion, the genes affected by CNVs found relatively more often in cases than in controls included some potentially interesting candidate genes based on their expression. However, in order to show if there is a genuine association between developing bipolar disorder and these CNVs, a replication in other studies is needed. Even though the support of these loci is weak, single observations of a CNV occurrence in a disorder could ultimately make an important contribution to the understanding of disease pathogenesis, as discussed already in section 5.6 (page 144). In order to facilitate this, the full list with rare CNVs found in the cases and controls has been uploaded on the following website, which has been made publicly available- http://x004.psycm.uwcm.ac.uk/~detelina/ (Grozeva et al. 2010).

5.10. Previously implicated CNVs in psychiatric disorders

5.10.1. Bipolar disorder

In contrast to schizophrenia, there were only a handful of studies exploring the potential involvement of CNVs in developing bipolar disorder. Following up on these studies, the putative CNVs were further investigated in the WTCCC bipolar data.

Wilson et al. studied on average ~35 brains from bipolar disorder cases, schizophrenia cases and controls using a BAC array comparative genome hybridisation (aCGH) with resolution of ~1.4 Mb and observed evidence suggesting that CNVs at 1p34.3, 14q23.3 and 22q12.3 loci were associated with bipolar disorder (Wilson et al. 2006). The inherent size-overestimation of CNVs inferred by BAC

aCGH indicates the need for a more systematic approach to detect CNVs in bipolar disorder. In addition, Sutrala et al. did not identify the CNVs reported by Wilson et al. using different method in a large sample of schizophrenia patients, and most importantly, in two of the samples already analyzed by Wilson et al. and previously shown to exhibit the duplications and deletions in question. Sutrala et al. concluded that more reliable methods need to be used in order to validate the existence of CNVs prior to full scale association studies being carried out (Sutrala et al. 2007). Therefore, it was not surprising that the possible involvement of CNVs on chromosomes 1p34.3, 14q23.3 and 22q12.3 affecting the *GLUR7*, *AKAP5* and *CACNG2* genes, in the WTCCC data, were not confirmed.

A study in bipolar disorder has reported a statistically significant increase in the frequency of gains affecting the *GSK3β* gene in bipolar patients as compared to controls (Lachman et al. 2007). *GSK3β* is a candidate gene for bipolar disorder as it encodes glycogen synthase kinase, which is a target for lithium salt. Lithium is well known as an effective treatment for bipolar disorder (Malhi et al. 2009). The sample size that Lachman et al. studied comprised ~80 individuals with bipolar disorder and ~80 controls (Lachman et al. 2007). In order to refute or accept the hypothesis of the possible involvement of duplications disrupting the *GSK3β* gene, additional studies are required. Saus et al. studied CNVs overlapping the *GSK3β* gene and the possible association with mood disorders in a Spanish population, but failed to replicate the results observed by Lachman et al. (Saus 2009). Similarly to Saus et al., the data from WTCCC showed no association between CNV variants affecting *GSK3β* and bipolar disorder.

A recent whole-genome study presented an analysis of the CNV structure in a three-generation pedigree segregating affective disorder (Yang et al. 2009). Some CNVs were found to be enriched in subjects with affective disorder (located at 6q27, 9q21.11, 12p13.31 and 15q11 loci). These regions were examined in the WTCCC bipolar study, but no statistically significant associations were observed. The study by Yang et al. made use of a family-based design (i.e. one three generation Amish family consisting of 46 individuals). This could potentially explain the difference with the WTCCC results. Some of the CNV variants in the Amish family could be family-specific and therefore, not observed in a population-based bipolar case sample as the one used in this PhD work. Furthermore, the array used was Illumina HumanHap550 BeadChip, which has a higher resolution than the one used in this PhD work. The

examined CNV of interest had size ranging from 19 kb to 207 kb. With the array applied in the current analysis, some of the CNVs could not be detected at all (e.g. with size of 19 kb) while others were with size, which was at the lowest end of the QC criteria ($\geq$ 100 kb). Therefore, such CNVs could have been potentially missed in the current study.

Although deletions in the Velo-Cardio-Facial Syndrome (VCFS) region have been previously implicated in bipolar disorder, no deletions were observed in the current WTCCC analysis (Papolos et al. 1996). These data strongly suggested that deletions in this region were not involved in bipolar susceptibility. Murphy et al. could not also replicate the finding of high prevalence of bipolar spectrum disorder in VCFS (Murphy et al. 1999). In the bipolar cases no reciprocal duplications at this locus were observed. Interestingly, that was not the case with respect to controls. Eight individuals exhibited duplications and six spanned the whole VCFS region. These eight controls were part of the National Blood Service controls. It has been previously suggested that the phenotype associated with duplication at the 22q11.2 shares some characteristics with VCFS (Ensenauer et al. 2003; Ou et al. 2008; Portnoi et al. 2005). The duplication is considered to lead to a separate syndrome named microduplication 22q11.2 syndrome (MIM #608363). The clinical presentation is variable ranging from normal to mild behavioural abnormalities and to multiple defects. It has been previously argued that theoretically such reciprocal duplication should have similar frequencies to the deletion (Yobb et al. 2005). As such duplication was not observed in the 440 schizophrenia samples, there is little support that the duplication is a predisposing factor for schizophrenia. In addition, the WTCCC bipolar CNV data did not provide evidence to support the hypothesis that the duplication at the 22q11.2 locus predisposes to bipolar disorder.

Overall, previous studies investigating the possible involvement of CNVs in the susceptibility to bipolar disorder comprised relatively small sample sizes. Furthermore, some of them have only focused on specific candidate genes. Thus, extreme care is needed in determining if the observed associated variants do indeed play a role in the susceptibility to bipolar disorder. A more systematic approach is needed for the determination of CNV structure in larger bipolar disorder sample sets. The first study satisfying these criteria was published in 2008 (Zhang et al. 2008). The results were already discussed individually in section 5.5 (page 141).

5.10.2. CNVs implicated in autism and schizophrenia

There is a vast body of evidence for the potential association of CNVs and the development of autism and schizophrenia (Guilmatre et al. 2009; Hannes et al. 2009; Ingason et al. 2010; Ingason et al. 2009; International Schizophrenia Consortium 2008; Kirov 2010; Kirov et al. 2009a; McCarthy et al. 2009; Rujescu et al. 2009; Stefansson et al. 2008). Some of the observed findings have been replicated in multiple studies. Replication of particular chromosomal abnormalities in independent reports is considered to be suggestive for the potential importance of the given variants and their association with the psychiatric illness.

In this section, I will present the main findings which have been reported in autism and schizophrenia and will discuss the possible involvement of these variants in bipolar disorder. Only variants which were reported in multiple studies or had reported a strong statistical support for the possible association with the diseases were considered. Examining the variants implicated in autism and schizophrenia with respect to bipolar disorder is justified as it has been postulated that these disorders could share some of the underlying genetic liability factors (Craddock and Owen 2010; Lichtenstein et al. 2009; Moskvina et al. 2009; Rzhetsky et al. 2007). A table of the main findings with respect to these loci can be found in section 4.2.7.1 (Table 25, page 109).

The CNVs implicated in autism and schizophrenia have been examined together for involvement in the susceptibility to bipolar disorder. The reasoning behind this decision was that for most of these loci, the data show that the same CNVs are associated with both autism and schizophrenia. This gives support to the hypothesis that the two disorders share common pathogenetic mechanisms.

1q21.1

Copy number variation at 1q21.1 locus has been initially associated with schizophrenia in two large genome-wide CNV surveys (International Schizophrenia Consortium 2008; Stefansson et al. 2008). Overall, this deletion has a rare occurrence, but is a powerful risk factor for susceptibility to schizophrenia. In a study performed by the International Schizophrenia Consortium, the deletion at this locus was observed with a frequency of 0.29% in schizophrenia cases and in 0.03% of the controls (International Schizophrenia Consortium 2008). Similar results have been

independently observed in a study performed by Stefansson et al. (Stefansson et al. 2008). The deletion was detected with frequency of 0.23% in cases, and 0.02% in the large control sample set respectively (Stefansson et al. 2008). When the individuals with the deletions were closely examined, three of the cases had learning disabilities and two controls had dyslexia.

Kirov et al. combined the data with respect to the deletion from these two large studies, excluding the overlapping individuals. The observed p -value from the combined data was $p = 9.6 \times 10^{-6}, OR = 10$ $(95\%CI : 2.9 - 34.2)$ (Kirov et al. 2009a).

Interestingly, chromosomal rearrangement (deletions and duplications) at 1q21.1 have been also implicated in phenotypes including autism spectrum disorders, developmental delay, dysmorphic features, congenital heart disease, mental retardation and learning disability (Brunetti-Pierri et al. 2008; Christiansen et al. 2004; de Vries et al. 2005; Mefford et al. 2008; Sharp et al. 2006b; Weiss et al. 2008).

No deletions at this locus were observed in the case bipolar data as compared with two deletions in 2806 controls. The rate of the deletions was very low in the non-psychiatric phenotypes, analysed by WTCCC- 0.01%. Interestingly a duplication was observed in one bipolar case. The frequency of duplication at this locus was observed with similar frequencies in cases and controls (0.06% and 0.07% respectively). In the 10,259 non-psychiatric controls, the duplication was found with similar rate-0.06%.

The general assumption is that a given rearrangement is more likely to be pathogenic if it has appeared *de novo* in an affected individual (Stefansson et al. 2008). To test if this duplication has arisen *de novo* or has been inherited from a parent, the proband and her parents were genotyped using an Illumina array. It was observed that the duplication was inherited from the father. Interestingly, the father also suffered from bipolar I disorder. Thus, it can be argued that the duplication in this particular family could play a role (potentially acting in concert with other factors) in developing the disorder. Of course, the hypothesis of the possible involvement of the 1q21.1 duplication as a susceptibility factor in some families with segregating bipolar disorder can only be accepted or refuted when there are more data from other whole-genome CNV studies in bipolar disorder.

NRXN1, 2p16.3

Deletions affecting the *NRXN1* gene were found to be associated with schizophrenia by Kirov et al., who observed the deletion in two siblings with schizophrenia. When 372 controls were screened, the deletion was not detected (Kirov et al. 2008). Rujescu et al. analysed large sample comprising 2977 schizophrenia cases and ~33,000 controls in order to determine the magnitude of the association, the prevalence and the odds ratio of the *NRXN1* gene in schizophrenia cases. When all CNVs at this locus were taken into account, no statistically significant association with schizophrenia was observed. Nevertheless, the CNV in cases were found three times more often than control (0.47% in cases, 0.15% in controls). Furthermore, when CNVs disrupting exons were analysed, it was observed that such CNVs are statistically significantly enriched in schizophrenia cases ($p = 0.0027$; $OR = 8.97$, $95\%CI : 1.8 - 51.9$) (Rujescu et al. 2009). Other studies have also reported schizophrenia cases with CNVs affecting the *NRXN1* gene (Ikeda et al. 2010; Need et al. 2009; Walsh et al. 2008). Interestingly, there are data that this deletion, disrupting the *NRXN1* gene, appears to be also increased in autism (Glessner et al. 2009; Kim et al. 2008; Kirov et al. 2009b; Szatmari et al. 2007; Weiss et al. 2008).

No deletions (or duplications) were observed in the bipolar cases. This suggests that deletions affecting the *NRXN1* gene could be a specific factor influencing the susceptibility to schizophrenia and autism, but not bipolar disorder.

15q11.2

Stefansson et al. observed an association between deletions at this locus and schizophrenia (frequency in cases and controls- 0.55% and 0.19% respectively) (Stefansson et al. 2008). Previously the same deletion has been observed in a case with mental retardation, developmental delay and speech impairment (Murthy et al. 2007). Kirov et al. combined data from Stefansson et al. study, data from the International Schizophrenia Consortium and results from their own study with respect to this region and observed the deletion with frequency of 0.6% in cases and 0.2% in controls ($p = 4.46 \times 10^{-8}$; $OR = 2.8$; $95\%CI : 2.0 - 3.9$) (Kirov et al. 2009a).

In the WTCCC bipolar cases the frequency of the deletion in this locus was 0.18% compared with 0.49% in the controls. This result was in the opposite direction to previous findings in schizophrenia. A lower rate in cases when compared with

controls has been also observed by Ikeda et al., in a Japanese whole-genome CNV study (0.2% in cases; 0.58% in controls) (Ikeda et al. 2010). The relatively high frequency of this deletion observed in the WTCCC controls and in the analysed controls by Ikeda et al. suggests that this CNV may be less pathogenic than initially regarded.

However, this deletion has been observed in 0.8% of 1105 children with unexplained intellectual disability. When compared to publicly available control data, the deletions were found significantly enriched in such cases ($p = 0.003$) (Mefford et al. 2009). Additionally, data presented by Doornbos et al. have also suggested that the deletion is likely to be pathogenic. An increased rate of the deletion has been detected in patients with mental retardation and/or multiple congenital abnormalities (0.57%) (Doornbos et al. 2009).

It is of interest that in the 10,259 non-psychiatric controls (Table 32, page 128) deletions at 15q11.2 locus were observed with higher frequency than the rate in previously analysed controls (0.39% vs. ~0.2%). Furthermore, the new frequency estimations were in line with the ones observed in the 2806 controls analysed in this PhD (0.49%). Therefore, it could be argued that this CNV does not increase the susceptibility only to schizophrenia and other neurodevelopmental disorders, but also predisposes to other non-psychiatric disorders, such as the ones included in the WTCCC study.

15q13.3

Associations between deletion at this locus and schizophrenia have been independently observed in two large studies (International Schizophrenia Consortium 2008; Stefansson et al. 2008). The International Schizophrenia Consortium has observed the 1.3 Mb deletion in 0.27% of the cases and in 0% of the controls ($p = 0.046, OR = 17.9$) (International Schizophrenia Consortium 2008). In Stefansson et al. study the deletion was observed in 0.17% of cases and in 0.02% of controls $p = 5.3 \times 10^{-4}, OR = 11.54$) (Stefansson et al. 2008).

Kirov et al. combined the data from these two studies with their own Cardiff data, excluding the overlapping individuals between the studies. The rate of the deletion in the three studies was 0.2% in the cases. In controls, the observed frequency was 10 times lower (0.02%). The association between the deletion and

susceptibility to schizophrenia was statistically significant ($p = 2.81 \times 10^{-8}$; $OR = 11.4, 95\%CI : 4.8 - 27$) (Kirov et al. 2009a).

Similarly to the other loci previously discussed, the deletion has not been exclusively associated with schizophrenia. A recurrent microdeletion syndrome due to deletions at this locus was reported by Sharp et al. (Sharp et al. 2008). The syndrome is characterised with mental retardation, seizures and variable facial dysmorphic features. Notably, one of the genes in the critical region is *CHRNA7* (cholinergic receptor), encoding a synaptic ion channel protein that mediates neuronal signal transmission (http://www.ncbi.nlm.nih.gov/gene/1139). Linkage studies have implicated this gene in a susceptibility to juvenile epilepsy and schizophrenia (Elmslie et al. 1997; Freedman et al. 1997). *CHRNA7* represents a plausible candidate gene as haploinsufficiency could cause the seizures observed in epilepsy cases with the deletion (Sharp et al. 2008). With respect to common epilepsy, it has been suggested that deletions at this locus constitute the most prevalent risk factor ($p = 5.32 \times 10^{-8}$) (Helbig et al. 2009). Subsequently, this finding has been replicated (Dibbens et al. 2009).

The frequency of the deletion has been observed to be 0.17% in a large sample of cases with mental retardation/developmental delay and autism spectrum disorders (Miller et al. 2009).

Van Bon et al. characterised 18 probands with deletions at the 15q13.3. In addition, four patients with a reciprocal duplication have also been assessed in an attempt to ascertain the clinical significance of the duplication. The deletion has been characterised by a broad intra- and inter-familial variability ranging from cardiac defects, mild developmental delay to a complete lack of learning problems (i.e. some of the carriers have been functioning at normal cognitive level). The observation that the deletion was detected in some healthy individuals suggests that it does not lead to mental retardation *per se* as the only causative factor. Therefore, it was concluded that the deletion plays a significant role in the pathogenesis of different conditions affecting the brain (including schizophrenia), but the broad phenotypic variability have also suggested that the outcome is likely to be determined in combination with other factors, such as other genetic, epigenetic or environmental factors. It was also argued that there was a possibility that some individuals were able to compensate this impairment, originally caused by the CNV. Evidence comes from the observation that

some individuals have shown learning problems during childhood which have been overcome in adulthood and they have exhibited normal functioning at the time of ascertainment (van Bon et al. 2009).

In contrast to the deletion, no distinct phenotype was observed in people with a reciprocal duplication (van Bon et al. 2009). It is of note that one of the cases with duplication, suffered from bipolar II disorder, pervasive developmental disorder not otherwise specified and has shown autistic features. He had positive family history of psychiatric disorders: the mother had personality disorder and recurrent depression, when the father was known to have schizophrenia. Notably, *CHRNA7*, one of the affected genes by the duplication has been previously shown to be associated with bipolar disorder (Hong et al. 2004; Turecki et al. 2001). Patients with the duplications were not found to share recognisable phenotype, but it is of note that psychiatric disease was noted in two out of the four studied individuals (van Bon et al. 2009).

In the WTCCC bipolar dataset no deletions were observed. Duplications were found with a frequency of 0.12% (two of 1697 cases). In the controls neither deletions nor duplications were detected. Even although this rate was not significantly different between the cases and controls, this CNV deserves further attention as it was observed in a case diagnosed with bipolar disorder, as shown by van Bon et al. (van Bon et al. 2009). In addition, in the 10,259 non-psychiatric control group, four duplications were observed (frequency-0.04%), which makes the duplication observed three times more often in cases than in controls.

The deletion at 15q13.3 has exhibited an association with neuropsychiatric conditions in multiple studies, which lends support to the theory that this CNV could potentially be pathogenic. There is not such strong evidence with respect to the reciprocal duplication. Nevertheless, the duplication has been found with increased frequencies in cases with schizophrenia as compared to the frequency in controls (International Schizophrenia Consortium 2008; Stefansson et al. 2008). Furthermore, duplications have been observed in cases with autism, language delay, anxiety spectrum disorder or obsessive compulsive disorder (Miller et al. 2009).

The potential association between developing bipolar disorder and duplication at 15q13.3 awaits investigation in independent studies.

16p11.2

Recurrent deletions and duplications at the 16p11.2 locus have been implicated in the susceptibility to autism spectrum disorders (Fernandez et al. 2009; Kumar et al. 2008; Marshall et al. 2008; Weiss et al. 2008). Weiss et al. investigated families with a proband affected with autism and observed a recurrent deletion (along with a reciprocal duplication), statistically associated with autism ($p = 1.1 \times 10^{-4}$) (Weiss et al. 2008). In the same study, a replication was performed using large number of cases and controls. The deletion was found to be present in one of 648 cases with schizophrenia, 1 out of 420 cases with bipolar disorder, in 1 of 203 attention-deficit hyperactivity disorder cases and in 1 of 3000 patients with panic disorder, anxiety or addiction. Overall, the frequency of autism deletion carriers was 1%, 0.1% among patients with a psychiatric or language disorder and 0.01% in the general population. The reciprocal duplication was observed in two subjects with bipolar disorder (0.48%) and five unscreened controls in the Icelandic control population (total number controls-18,834). The overall frequency was as follows: 0.5 % in subjects with psychiatric or language disorder and in 0.03 % of the general population (Weiss et al. 2008).

A deletion at 16p11.2 is thought to be one of the most frequent CNV associated with autism (the other one is a duplication of the Prader-Willie/Angelman region) found in $\sim 1\%$ of the cases (Glessner et al. 2009; Kumar et al. 2008; Weiss et al. 2008).

To add to the evidence of the possible involvement of this locus in developing neuropsychiatric disorders, two individuals with childhood onset schizophrenia have also been found to carry the duplication (Walsh et al. 2008). Furthermore, a duplication at this locus has been significantly associated with a substantial risk for developing schizophrenia ($p = 4.8 \times 10^{-7}; OR = 8.4, 95\% CI: 2.8 - 25.4$) (McCarthy et al. 2009). The same duplication has also been associated with bipolar disorder ($p = 0.017$) and autism ($p = 1.9 \times 10^{-7}$). Conversely, the reciprocal deletion has only been associated with autism and developmental disorders ($p = 2.3 \times 10^{-13}$). McCarthy et al. examined these findings further. The head circumference has been found to be increased in patients with the deletion when compared with duplication carriers ($p = 0.0007$). The finding that autism features an increased head circumference is not

new. Several studies have reported an increased head circumference in patients with autism (Fidler et al. 2000; Lainhart et al. 2006).

These findings have suggested that mutations at this locus confer risk not only to schizophrenia, but also to other neuropsychiatric conditions (Kumar et al. 2008; Marshall et al. 2008; McCarthy et al. 2009). McCarthy et al. studied the duplication in families, where the proband was a carrier and observed the duplication in some relatives exhibiting a range of neuropsychiatric conditions including schizophrenia, bipolar disorder, depression and psychosis signs not otherwise specified. Healthy carriers were also observed. Such observations are consistent with the theory that variants at this locus are associated with incomplete penetrance and variable expression. The findings of the possible involvement of rare CNVs in schizophrenia have shown marked genetic heterogeneity. Although 16p11.2 locus accounts for only a small fraction of the schizophrenia cases, people who harbour the duplication are at a substantial risk of developing schizophrenia ($OR = 8.4$, $95\%CI : 2.8 - 25.4$) (McCarthy et al. 2009).

With respect to variants at 16p11.2, three duplications were observed in the WTCCC bipolar cases (0.2% frequency) as compared to one in the controls (0.04% frequency). This locus, showed a trend for an overrepresentation in bipolar cases. The rate in the cases was increased 5-fold when compared with the baseline rate (i.e. the rate in the control population), although it did not reach a statistical significance. Nevertheless, a statistical significance has been noted when a meta-analysis was performed as part of the McCarthy et al. study. The meta-analysis comprised 4822 bipolar cases and 25,225 controls (including the results from the current PhD analysis). A significant difference between the cases and controls was detected ($p = 0.017$) (McCarthy et al. 2009). Interestingly, in schizophrenia the statistical association with the duplication is much stronger- $p = 4.8 \times 10^{-7}$. Furthermore, the effect size is much more pronounced in schizophrenia than bipolar suggesting that the effect of this duplication is weaker in bipolar disorder. The finding that duplications at 16p11.2 locus may be involved in bipolar disorder was further supported by the analysis of the 10,259 non-psychiatric controls. The duplication in the bipolar cases had a frequency of 0.18% whereas the frequency in the controls was 0.04% ($p = 0.06$).

Parent DNA was available for one of the three probands found to be a duplication carrier. The trio was genotyped in order to examine the transmission and it was found that the duplication had arisen *de novo*. Notably, the father also suffered from bipolar I disorder. Thus, it can be argued that in this family, there are more factors that could potentially play a role in developing the disease, such as other genetic factors shared in the family, shared environmental factors, epigenetic factors, etc.

16p13.1

Initially CNVs at this locus were implicated in autism (Sharp et al. 2006b; Ullmann et al. 2007) and mental retardation (Freitag 2007; Hannes et al. 2009; Mefford et al. 2009). A 1.5 Mb duplication at 16p13.1 locus has been found associated with autism while the reciprocal deletion- with mental retardation and other clinical abnormalities. Nevertheless, the duplications and the reciprocal deletions have been found to be present in mildly affected and in completely healthy individuals suggesting that the CNVs at this locus predispose to autism/mental retardation, but developing the disease could be dependent on other factors as well (Ullmann et al. 2007).

CNVs at 16p13.1 have also been implicated in schizophrenia (Ikeda et al. 2010; Kirov et al. 2009a). Kirov et al. observed the duplication with frequency in cases of 0.6% and in controls-0.2%. Furthermore, Ingason et al. examined a large sample set of schizophrenia cases and controls (4345 and 35,079 respectively) for duplications or deletions at this locus (Ingason et al. 2009). A 3-fold excess of duplications and deletions was observed in the schizophrenia cases when compared with controls. The duplications were present in 0.30% of cases vs. 0.09% in controls ($p = 0.007$) and deletions in 0.12% of cases vs. 0.04% in controls ($p > 0.05$).

In the WTCCC bipolar study, two duplications (0.12% frequency) were observed in cases compared to five duplications in the controls (0.18% frequency). In the 10,259 controls the rate of the duplication was even higher-0.24%. Therefore, it could be concluded that there was no trend for overrepresentation of the duplication in the bipolar cases at this locus. Similar were the results with respect to the reciprocal deletion, which was not observed in any cases and was detected in one control.

17p12

Kirov et al. observed a 1.4 Mb deletion at 17p12 locus that was not detected in any of the studied controls. When these data were combined with data from two large genome-wide surveys on schizophrenia, the combined statistical significance level was $p = 5 \times 10^{-5}$. The deletion was found to be 10 times more common in schizophrenia cases than in controls (0.15% in cases vs. 0.015% in controls) (Kirov et al. 2009a).

It is of note that deletions at this locus cause hereditary neuropathy with pressure palsies (HNPP, MIM #162500) (Chance et al. 1993). The reciprocal duplication has been found to cause Charcot-Marie-Tooth disease Type 1A (CMT1A, MIM #118220) (Lupski et al. 1991). These neurological phenotypes are known to be caused by the deletion/duplication of peripheral myelin protein 22 (*PMP22*) (Chance et al. 1993; Lupski et al. 1991). Nevertheless, there has been a previous report of a patient with schizophrenia having the deletion, without manifestation of neurological symptoms (Ozeki et al. 2008).

In the WTCCC bipolar data, one duplication at 17p12 was detected (frequency-0.06%). In the controls neither deletions nor duplications were observed. In the analysed 10,259 non-psychiatric controls the frequency of the duplication was 0.03%. At present, the suggestion of a possible involvement of the duplication at this locus in bipolar disorder and especially the duplication of the *PMP22* gene, is premature as the prevalence rate of Charcot-Marie-Tooth disease type 1A is 1 in 2500 individuals (Lupski et al. 1991). Therefore, it is difficult to determine whether the duplication is a genuine rare variant predisposing to bipolar disorder or a chance occurrence.

22q11.2

This locus was discussed already in section 5.10.1 (page 150), where loci previously associated with bipolar disorder were described. In summary, in the WTCCC bipolar cases no deletions were observed. The presented herein data did not provide evidence to support the hypothesis that deletions at the 22q11.2 locus predispose to bipolar disorder.

Summary from follow up of the autism/schizophrenia loci in WTCCC bipolar data

Several specific deletions and duplications have been implicated in schizophrenia in more than one study (e.g. 1q21.1, *NRXN1*, 15q11.2, 15q13.3, 16p11.2, 16p13.3 and 22q11.2). Furthermore, for most of these loci a plethora of evidence for the possible involvement in autism or mental retardation has been observed.

Analysing these loci in the bipolar cases has provided some suggestive evidence for their potential involvement in the pathogenesis of the disorder, namely: 1q21.1, 15q13.3 and 16p11.2. It should be noted, however, that none of these were found to be significantly overrepresented in bipolar cases when compared with controls. This suggests that these findings should be treated with caution and should await further replication in other whole-genome CNV studies in large bipolar sample sets. Interestingly, for three of the loci implicated in schizophrenia (1q21.1, 15q13.3 and 17p12) the duplication was observed in bipolar cases, opposed to the deletion in schizophrenia cases.

There has been previous evidence that duplications as compared to deletions are not as pathogenic. The clinical presentation associated with deletions is likely to be more uniform, severe and commonly includes dysmorphology which facilitates diagnosis (Miller and Therman 2001).

Interestingly, Need et al. observed an excess of deletions in schizophrenia patients (Need et al. 2009). For duplications, on the other hand, evidence is emerging for more subtle presentations, creating the potential for duplication carriers to go undetected (Cook and Scherer 2008). If deletions are indeed more likely to be pathogenic than duplications, it could be speculated that deletions could contribute to the chronic course of schizophrenia. In contrast, bipolar disorder is characterised with an episodic nature of the illness and duplications could confer to the more cyclic nature of the phenotype.

5.11. CNVs in bipolar disorder and schizophrenia

5.11.1. Direct comparison between bipolar disorder and schizophrenia with respect to CNVs

There is a plethora of prior data that bipolar disorder and schizophrenia share underlying genetic components that predispose to developing the disorders (Gottesman et al. 2010; Lichtenstein et al. 2009; Rzhetsky et al. 2007). Therefore, it was sought to determine if the similarities in the genetic overlap stretch to CNVs as well.

As the same methodology as in the analysis of the bipolar cases was applied for the analysis of the schizophrenia cases, a direct comparison between the two was possible. Genotyping in the same pipeline of the two cohorts in the WTCCC context, and analysing them at the same time using identical methods ruled out any batch effects. Thus, I was able to test the hypothesis if bipolar disorder resembles schizophrenia with respect to CNV structure. In the previous section, it was shown that there are differences with respect to the involvement in the susceptibility of specific CNV loci. Herein, results from comparison between bipolar disorder and schizophrenia with respect to CNV burden will be presented.

When bipolar and schizophrenia cases were compared directly against each other with respect to CNVs ≥ 1 Mb, a statistically significant excess of deletions was observed in the schizophrenia cases ($p = 0.0009$) and total number of CNVs ($p = 0.0006$). These p -values were not corrected for multiple hypotheses testing, but nevertheless they would survive a Bonferroni correction for the number of the performed tests. There was also a trend towards an excess of large duplications (≥ 1 Mb) ($p = 0.053$) in schizophrenia cases compared with bipolar cases. The full results with respect to this comparison were presented in section 4.4 (page 125). The data exhibited a statistically significant difference between bipolar disorder and schizophrenia with respect to global burden of rare and large CNVs (≥ 1 Mb).

The bipolar disorder sample set was nearly four times larger than the studied schizophrenia cohort (1697 bipolar cases versus 440 schizophrenia cases). This suggests that if there were an effect similar to the one detected in the schizophrenia sample, it should have also been detected in the bipolar dataset. In conclusion, the

comparison between the bipolar and schizophrenia cohorts suggests that the disorders differ with respect to the CNV burden.

5.11.2. Comparison between CNVs in bipolar disorder, schizophrenia and WTCCC non-psychiatric diseases

Recently, data from one of the largest genetic studies became publicly available (Wellcome Trust Case Control Consortium 2007). All the phenotypes that took part in Wellcome Trust Case Control Consortium (WTCCC) study were analysed with respect to CNVs. This analysis was done by WTCCC based on the SNP genotyping data of the Affymetrix 500K array. This provided an opportunity to compare the copy number variation structure in bipolar disorder and schizophrenia with a very large sample set of >10,000 people from the general population in the UK, who are not affected with neuropsychiatric disorders.

As the analysed bipolar cases and controls in this PhD were part of the WTCCC study, the first question was how the two studies relate to each other. This presented an opportunity to examine the correspondence rate between the two studies, as the analysed samples and initial raw genotyping data used for inferring of the CNVs were the same.

When the calls for the large CNVs implicated in schizophrenia (that is, 1q21.1, 15q11.2, 16p11.2, etc.) were examined in the WTCCC data and in the bipolar cases and controls (analysed in this PhD work), 100% correspondence between the two studies was observed. This provided confidence that results, even though based on different calling algorithms, were comparable without introducing bias.

The frequency of the schizophrenia-associated loci in over 10,000 people from the general population affected with non-psychiatric diseases was examined and compared with the frequency in the bipolar cases (section 4.5, Table 32, page 128). Only one region showed marginal trend for an overrepresentation in cases-duplications at 16p11.2 ($p = 0.06$). This finding was not surprising, as this locus has been tested in a large meta-analysis in bipolar disorder as discussed previously 5.10.2 (page 153).

To evaluate if there is a correspondence between the rates in the original 2806 controls analysed in this PhD and the rates in the new large sample set of non-

psychiatric controls, a comparison between the two was performed. The rate was almost the same, which was indicated by the non-significant association between the two groups. This observation has shown that there was no differential bias between the two studies. In addition, it was shown that the previously inferred rate of CNVs in the 2806 controls could be relied upon.

The rate of the CNVs previously implicated in schizophrenia was compared between the bipolar cases and the large set of non-psychiatric controls. As observed previously, no statistically significant overrepresentation was observed in bipolar cases for any of the loci. Therefore, a replication of the main finding in this PhD was noted, using previously not analysed large control sample set. Furthermore, the bipolar cases were analysed independently by the Wellcome Trust and the results, confirmed the findings already presented in the PhD work.

When the rate of the CNVs previously associated with schizophrenia was compared between a large set of schizophrenia cases and the 10,259 non-psychiatric controls, most of the CNVs were statistically significantly overrepresented in the schizophrenia cases. The loci which were confirmed to be associated with schizophrenia were the following: 1q21.1, 15q11.2 (confirmed with $p = 0.05$, therefore the results should be treated with caution), 15q13.3, 16p11.2 and 22q11.2. Surprisingly, the evidence was not that convincing for CNVs at the *NRXN1* gene, 16p13.1 and 17p12. The CNVs at these loci were detected at rates that are closer to the frequencies among schizophrenia patients, and higher than in previous controls. Therefore, further large samples will be required to establish the exact prevalence of these CNVs and their role in psychiatric and non-psychiatric phenotypes. It could be speculated that some of these CNVs do not increase the risk specifically to develop schizophrenia and other neurodevelopmental disorders, but also predispose to other disorders, such as diabetes and heart disease (included in the WTCCC study), or to non-specific changes in the brain that lead to increased help-seeking behaviour and increase the probability of being sampled in a disease study. This can explain their relatively high rate in unscreened members of the general population who suffer with non-psychiatric medical disorders. It is also possible that the current frequency results in controls could be more representative of the true population prevalence, and the odds ratios reported for some of these loci might have to be scaled down. It should be acknowledged, that the frequencies of all of these loci were still higher in the

schizophrenia cases when compared with the 10,259 controls. In addition, when the cumulative rate of the CNVs, previously implicated in schizophrenia was compared between schizophrenia cases and non-psychiatric controls, a highly statistically significant difference was noted- $p = 5 \times 10^{-11}$. Bipolar cases were also compared with the large set of controls and no difference was observed ($p = 0.19$). This further replicates the findings already presented in this PhD and provides further support to the observation that bipolar disorder and schizophrenia are different with respect to involvement of CNVs in the susceptibility.

When the burden of very large CNVs (≥ 1 Mb) was compared between the diseases, bipolar and rheumatoid arthritis cases had the lowest rate. These differences were statistically significant when bipolar cases were compared with the large set of controls with respect to deletions and total number of CNVs ≥ 1 Mb ($p = 0.02$ and $p = 0.003$ respectively). The differences originate from the fact that the cases have lower rate of such CNVs than controls. When the schizophrenia cases were compared with the controls, statistically significant differences were also observed with respect to all tested CNVs ≥ 1 Mb (deletions- $p = 0.001$, duplications- $p = 0.02$ and total number CNVs- $p = 0.0001$). In contrast to bipolar disorder, the statistical significance indicates that the schizophrenia cases had an increased rate of these CNVs than the controls.

It is of interest that the cumulative burden of all the tested CNVs, previously implicated in schizophrenia and all the CNVs ≥ 1 Mb in bipolar disorder was in fact the lowest compared to any other cohort in the WTCCC study (Figure 42 and Figure 43). This was already noted by the comparison with the 1958 cohort and NBS controls from the WTCCC and was in line with the aforementioned results in this PhD.

Based on the fact that the rate of CNVs ≥ 1 Mb is the lowest in bipolar disorder it could be speculated that this trend is a genuine finding. As the rate of such very large CNVs (≥ 1 Mb) is the highest in schizophrenia, it could be also suggested that the presence of such CNVs makes somebody who is already susceptible to psychotic disorder develop schizophrenia rather than bipolar disorder.

Figure 42 Rate of deletions, duplications and total number of CNVs ≥1 Mb in the cohorts from the WTCCC study

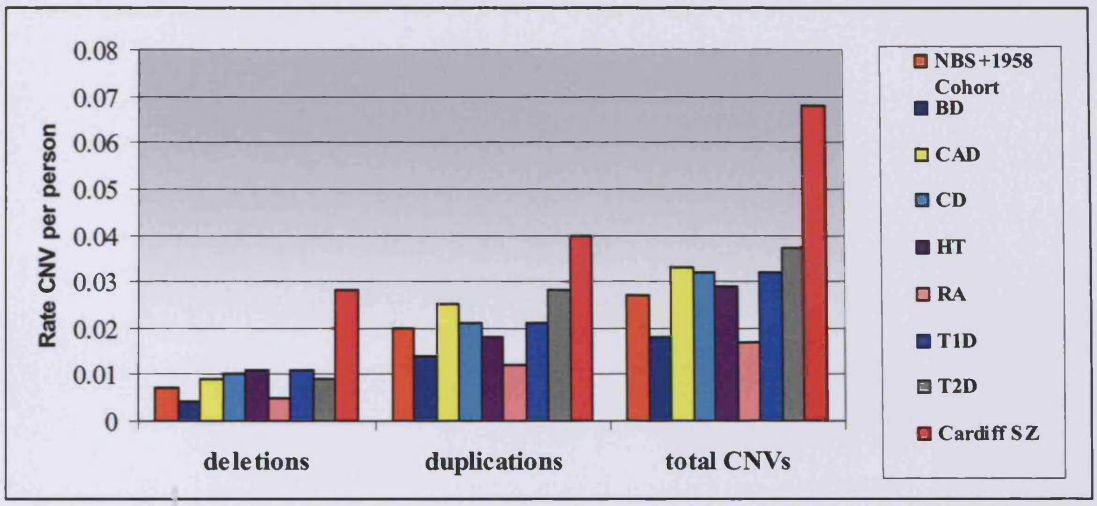
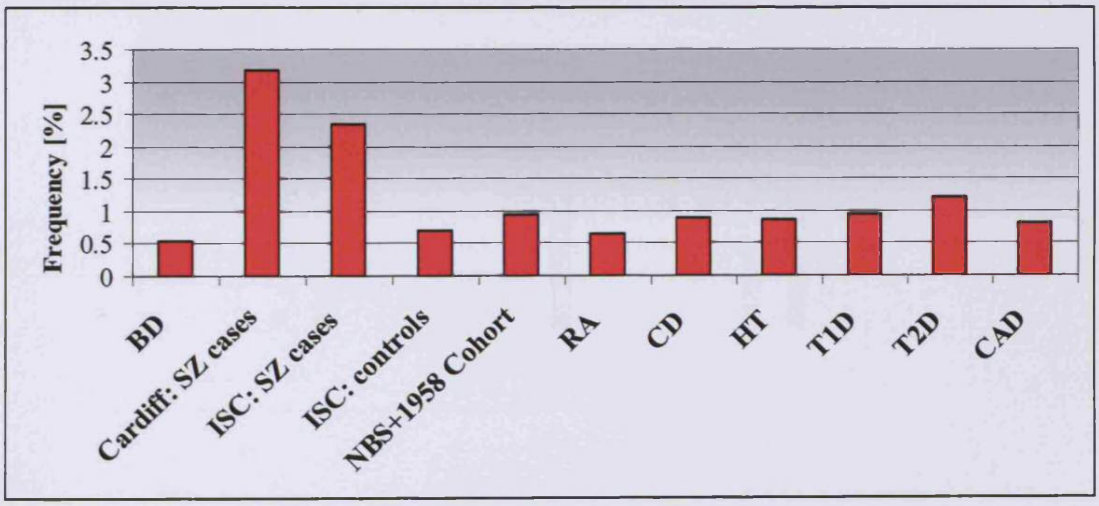


Figure 43 Rate of cumulative burden of CNVs, previously implicated in schizophrenia in the cohorts from the WTCCC study



burden based on loci 1q21.1, *NRXN1* gene, 15q11.2, 15q13.3, 16p11.2, 16p13.1, 17p12 and 22q11.2; ISC SZ cases- schizophrenia cases that participated in the International Schizophrenia Consortium study

The evaluation of the data, independently analysed by WTCCC and the inclusion of additional >10,000 controls in the analysis, provides further support to the observation, already presented in the PhD, that in bipolar disorder, unlike schizophrenia, the large and rare CNVs do not influence the liability to the condition.

5.11.3. Are bipolar disorder and schizophrenia different with respect to CNV burden?

Bipolar disorder and schizophrenia were compared with respect to the possible involvement of CNVs in the pathophysiology of the illnesses as previous evidence has shown an existence of some shared genetic susceptibility factors between the two disorders (e.g. *ZNF804A, DISC1, CACNA1C* genes) (Blackwood et al. 2001; Green et al. 2009; Moskvina et al. 2009; O'Donovan et al. 2008).

Further evidence for the possible common underlying genetic variants between schizophrenia and bipolar disorder comes from a large Swedish population-based study (> 2 million families). A key finding has been an increased risk for both schizophrenia and bipolar disorder in first-degree relatives of probands with either of the disorders. Furthermore, the comorbidity between the two disorders has been estimated to be mainly due to genetic effects common to both disorders (63%) (Lichtenstein et al. 2009). Another large population study (based on 1.5 million patient records) has observed genetic overlap between not only bipolar disorder and schizophrenia, but also between bipolar disorder, schizophrenia and autism (Rzhetsky et al. 2007).

Until now the hypothesis of a possible overlap with respect to copy number variants in bipolar disorder and schizophrenia has not been tested. Having performed studies of CNVs in both disorders using samples collected from the same population, which were genotyped and analysed for CNVs in the same pipeline, allowed such a direct comparison to be performed.

Overall, the main finding from the comparison was that bipolar disorder and schizophrenia differed with respect to the involvement of CNVs in the susceptibility of development of the disorders (the differences are mainly in respect to CNVs >1 Mb, and for CNVs previously implicated in schizophrenia). In contrast to the observations of the possible involvement of some common SNPs and genes, the CNV findings suggest that copy number variation could have a relatively specific influence on the susceptibility to schizophrenia, but not bipolar disorder. Evidence for this comes not only from the comparison between bipolar disorder and schizophrenia with respect to the overall CNV burden (in section 4.4, Table 31, page 126), but also from

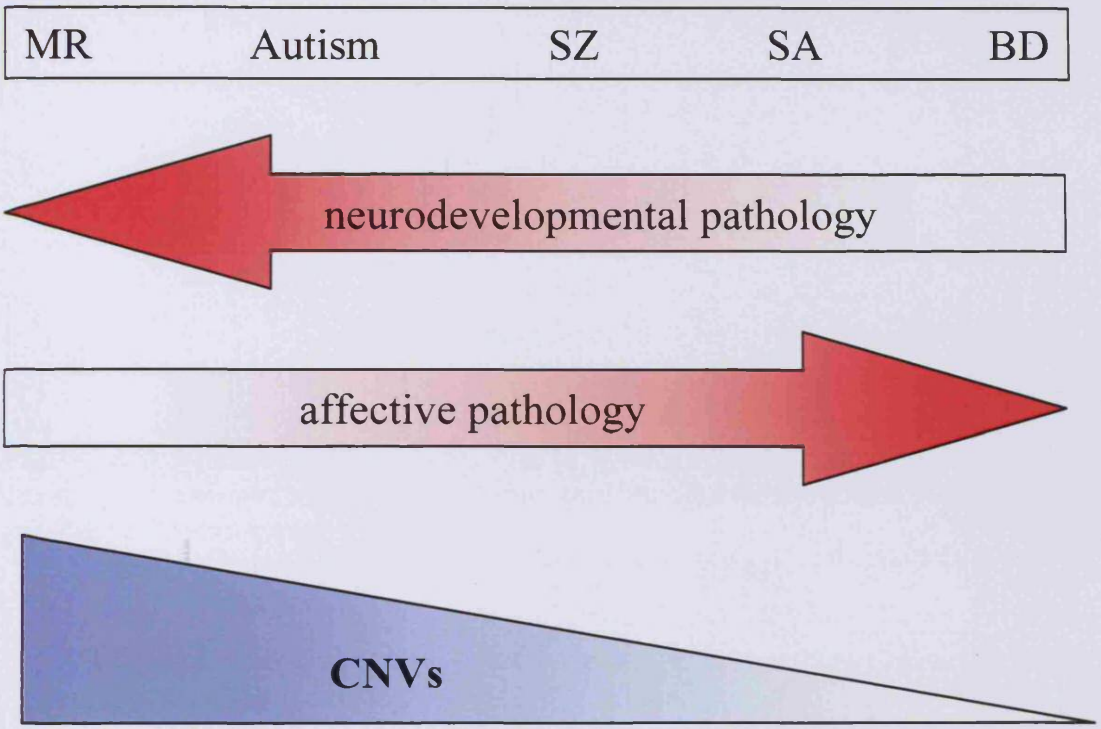
the investigation of the specific CNVs implicated in schizophrenia that were not observed in bipolar cases (section 108, Table 25, page 109).

Similar to these findings, the CNV studies performed in bipolar disorder to date have observed that CNVs appear to contribute to a lesser extent to the susceptibility to bipolar disorder than to schizophrenia (McQuillin et al. 2010; Priebe et al. 2009; Purcell 2009; Zhang et al. 2008).

The current PhD study has also shown that the variants influencing bipolar disorder are less likely to be deletions (specifically, 1q21.1, 15q13.3 and 17p12) and have a smaller effect as compared to schizophrenia (e.g. 16p11.2).

If the assumption that larger CNVs especially deletions, are more likely to affect brain development is true, then the findings presented herein are consistent with schizophrenia having a stronger neurodevelopmental component than bipolar disorder (Craddock and Owen 2010; Murray et al. 2004). Based on this evidence, Craddock and Owen suggested that there is a gradient of decreasing neurodevelopmental impairment between syndromes such as autism and mental retardation on one hand, and bipolar disorder on the other (Craddock and Owen 2010). The gradient is presented at Figure 44. Evidence for this continuum does not solely come from genetic studies. Overlapping symptoms like depressive symptoms or psychoses, the "intermediate" phenotype of schizoaffective disorder and the co-occurrence of bipolar disorder and schizophrenia in families, lends support to the theory of a clinical continuum (Alaerts and Del-Favero 2009).

Figure 44 Suggested model of the potential gradient in the neurodevelopmental impairment between mental retardation and bipolar disorder and the potential gradient of involvement of CNVs



adapted from (Craddock and Owen 2010); MR-mental retardation, SZ-schizophrenia, SA-schizoaffective disorder

It is possible that CNVs predispose to brain dysfunction that affect intellectual functioning and personality development and therefore could modify expression of the phenotype in those who have a tendency to develop psychosis (Craddock et al. 2009). It is of note, that most of the CNVs robustly associated with schizophrenia, have also been implicated in neurodevelopmental disorders with childhood onset in which cognitive impairment is common (that is, mental retardation and autism spectrum disorders). Guilmatre et al. estimated the collective frequency of a set of recurrent CNVs in patients suffering from schizophrenia, autism and mental retardation (Guilmatre et al. 2009). 28 loci previously associated with schizophrenia, autism and mental retardation were investigated. It was observed that recurrent CNVs were found in 39.3% of the selected loci and that the collective frequency of CNVs is significantly associated with schizophrenia, autism and mental retardation ($p = 0.01$, $p < 0.001$ and $p = 0.001$ respectively). The main finding from the study suggested that most of the studied CNVs were present in the three examined illnesses, which supports the existence of shared biological pathways for the susceptibility to the three

conditions. Guilmatre et al. concluded that none of the recurrent rearrangements were limited to one disease (Guilmatre et al. 2009).

The observation that most of the specific CNVs, found in schizophrenia, are also detected in mental retardation and autism spectrum disorders lends support to the hypothesis that schizophrenia and these disorders share overlapping biological pathways. Why one CNV leads to developing schizophrenia in one patient and autism in other, could be explained by the fact that expressivity could be influenced by additional genetic events, epigenetic effects, environmental influence, etc.

In contrast, in bipolar disorder the relative paucity of these variants, lends support to the hypothesis that large and rare CNVs have less prominent role in the development of the disorder (Figure 44, page 171). Studies of CNVs have found an overlap between autism and schizophrenia, whereas common single nucleotide variants have demonstrated overlap between schizophrenia and bipolar disorder. It could also be speculated that a subset of schizophrenia cases share common biological underpinnings with childhood neurodevelopmental disorder whereas other subset is more closely related to bipolar disorder. Nevertheless, this finding further supports the view that schizophrenia has a more prominent neurodevelopmental component than bipolar disorder.

It has been shown that cognitive impairment among schizophrenia patients is present from early childhood, before the disease onset which supports the hypothesis of a stronger neurodevelopmental component that could play part in developing schizophrenia. Schizophrenia and bipolar disorder differ for the fact that schizophrenia has been characterised with earlier onset, poor response to medication, frequent relapse and chronic course of illness (Williams et al. 2009). Furthermore, minor physical anomalies are more prevalent in schizophrenia patients than in healthy controls and are considered as markers of an aberrant neurodevelopment (John et al. 2008). Andreasen has also shown that schizophrenia is associated with cognitive and social impairment (Andreasen 1995). Patients with schizophrenia perform one or two standard deviations below when compared with healthy controls on cognitive tasks involving memory, attention, problem solving and social cognition. Interestingly, these deficits have been present prior to the start of treatment with antipsychotic drug, suggesting that the cognitive impairment is not caused by the treatment (Keefe and Eesley 2009). It has been noted that to some extent first-degree relatives have also exhibited cognitive impairment.

In contrast to the diminished cognitive functioning in schizophrenia, cognitive impairment is less typical for bipolar disorder patients and it has been observed after the disease onset. Unlike schizophrenia, bipolar disorder is generally not considered to have a neurodevelopmental component (Murray et al. 2004; Walker et al. 2002). This theory has been supported by a number of longitudinal studies which have described no differences in the intelligence of children who will later develop bipolar disorder as compared to children who will not develop the disorder (reviewed in (Barnett et al. 2006)). In bipolar disorder, like schizophrenia, cognitive impairment has also been noted, but it is found after the onset of the illness and limited to acute episodes (reviewed in (Burdick et al. 2009)). Nevertheless, there are studies that report some evidence for neurodevelopmental basis of bipolar disorder, but the evidence is far from conclusive as it is in schizophrenia.

In summary, the analysis from this PhD showed that large rare CNVs were not involved in bipolar disorder as compared to the extent they were involved in schizophrenia. Therefore, CNVs could have a specific effect on schizophrenia, but not in bipolar disorder. The genetic liability factors for developing bipolar illness could be smaller deletion/duplications, SNPs, point mutations, environmental factors and pure random factors that could work together to influence the expression of bipolar disorder.

6. Future directions

Studying psychiatric disorders in general and bipolar disorder in particular are challenging due to their inherent genetic heterogeneity, incomplete penetrance of psychiatric traits and non-Mendlian inheritance of complex traits.

Nevertheless, there is unequivocal evidence that genetic factors play an important role in developing bipolar disorder. The research that I present herein focused on investigating a particular type of genetic factors. As we are still in the discovery phase of investigation, in order to study the full variety of genomic variation, sequencing approaches will be the most powerful methods. They would be particularly potent at identifying small changes along with inversions and translocations, which to date have received little attention. As technology is constantly improving, it seems plausible that the cost of sequencing will decrease up to a point when processing of many samples (in the range of thousands) would be feasible. The high-throughput DNA sequencing promises to improve power to identify new mutations associated with susceptibility to developing disease.

In addition, increasing the sample size of investigated bipolar individuals will help to tease out genetic variants with low-penetrance effect. Furthermore, if well phenotypically characterised samples are readily available, this could potentially help to unravel genetic variants that could influence specific phenotypes.

My belief is that in the next couple of years much more will become known with respect to possible causal genomic factors that engender risk for developing common and complex disorders. The results are awaited with great anticipation as the findings will bring enormous benefit to the people who suffer from such disorders.

7. Conclusions

The CNV structure was successfully inferred in a large cohort of bipolar cases and healthy controls using SNP genotyping data from Affymetrix 500K arrays. It was observed that considerable care is needed in analysing copy number data based on SNP genotyping data, as CNV analyses are susceptible to a range of artefacts which could potentially lead to false-positive results. Despite the important technical challenges and potential artefacts, it was demonstrated that high-confidence and robust CNV calls are possible to be inferred.

Furthermore, it was determined that large and rare copy number variants were not associated with developing bipolar disorder. Some specific CNVs were found more often in cases than in controls, but this observation did not withstand correction for multiple hypotheses testing.

Analysing schizophrenia and bipolar cases using identical methods and processing the data in the same pipeline, allowed direct comparisons between the disorders. The main observation was that in bipolar illness, very large CNVs were not involved to the extent that they have been shown to be involved in schizophrenia. Furthermore, some specific duplications (at loci already implicated in schizophrenia, e.g. 1q21.1, 16p11.2, 16p13.1, 17p12 and 15q13.3) were observed in cases with bipolar illness, but the evidence for their involvement was suggestive and needs replication in other bipolar studies. The relative paucity of large CNVs in bipolar disorder suggests a lesser neurodevelopmental component than schizophrenia, given that large aberrations have been associated with severe neurodevelopmental phenotypes, such as mental retardation and autism, and are often accompanied by a variety of more pronounced physical anomalies, where bipolar disorder is characterised by relapse and remit.

The strengths of this PhD work comprise a large sample size, rigorous quality control and the ability to directly compare bipolar disorder with schizophrenia. Some important limitations are inherent in all CNV studies. The first is with respect to the power to detect very small increases in CNV burden. Second, the Affymetrix 500K array has less resolution for CNV detection than later platforms. This has much less impact on large CNVs (the focus of this PhD work), but due to this limitation, smaller CNVs (≤ 100 kb) were not examined. It has been previously shown that the number

of CNVs exponentially increases with decrease of CNV size (McCarroll et al. 2008). In future studies, it will be important to use platforms with much higher resolution. This will potentially allow CNVs to be called at similar levels of accuracy as SNPs in order to provide further information about a wider spectrum of CNV sizes.

While additional studies are required to further explore the hypothesis of the potential involvement of CNVs in the susceptibility to bipolar disorder, the analysis presented in this PhD thesis, provides one of the first attempts to shed some light on the potential involvement (or lack of involvement) of large deletions and duplications in the aetiology of bipolar disorder.

Appendix

Table 38 Genes in CNVs in cases only

Chr	Start bp	End bp	CNV	NCBI genes	Number of observations
1	45592076	45881891	dup	TESK2,LOC126661,MMACHC,PRDX1,AKR1A1,NASP,CCDC17,GPBP1L1	1
1	46131935	46268021	dup	MAST2	1
1	65071676	65413912	dup	JAK1,MIRN101-1,AK3L1,AK3L2	1
1	86324452	86589985	del	COL24A1,ODF2L	1
1	88847985	89032679	dup	PKN2	1
1	107551303	107823281	dup	NTNG1	1
1	143576984	143916898	del	PDE4DIP,FLJ21272,SEC22B	1
1	152936201	153149168	del	KCNN3	1
1	166729757	167024066	del	XCL2,XCL1,DPT	1
1	173980586	174202365	dup	LOC100128153,RFWD2	1
1	217365162	217613673	del	LOC643723,LYPLAL1	1
1	221702789	222127692	dup	LOC644151,LOC653428,CAPN8,CAPN2,TP53BP2	1
1	231361905	231514154	del	PCNXL2	1
1	244896239	245307486	dup	C1orf71,SCCPDH,LOC100130097,LOC149134,AHCTF1,ZNF695,ZNF670	1
2	155674	306842	dup	LOC727818,SH3YL1,ACP1,FAM150B	1
2	9950762	10157626	dup	TAF1B,GRHL1,UNQ5830,KLF11,LOC100131506,CYS1	1
2	46875500	47239348	dup	LOC388948,MCFD2,TTC7A,LOC100129286,C2orf61	1
2	49008203	49120697	del	FSHR	1
2	49149900	49602914	del	FSHR	1
2	53738562	53909774	dup	ASB3,CHAC2,C2orf30	1
2	67026445	67574931	del	LOC644838,ETAA1	1
2	80501388	81016718	del	CTNNA2	1
2	105308220	105583762	del	TGFBRAP1,C2orf49,FHL2,LOC728966	1
2	108755278	109814503	del	RANBP2,CCDC138,EDAR,SH3MD4,LOC100132457,LOC729164,SEPT10,ANKRD57,LOC100131577	1
2	137213276	137516286	del	THSD7B	1
2	175063486	175545089	del	LOC100130325,WIPF1,LOC100133109,CHRNA1,CHN1	1
2	211762892	212453633	dup	ERBB4	1
2	230634380	230823051	del	SLC16A14,SP110,SP140	1
3	96226	577858	del	LOC642891,CHL1	1
3	6731867	7502372	dup	GRM7	2
3	56538380	56738659	dup	CCDC66,C3orf63,ARHGEF3	1
3	65782031	66651118	dup	MAGI1,SLC25A26,LRIG1	1
3	108045770	108523998	del	LOC344595	1
3	109343496	109735186	dup	IFT57,HHLA2,MYH15	1
3	126612179	126918638	dup	SNX4,OSBPL11	1
3	159972476	160172525	del	MFSD1	1
3	169645563	170273620	dup	LOC389174,MIRN551B,LOC253820,C3orf50	1
3	175127179	175441056	dup	NLGN1	1
3	191379157	191640477	del	CLDN1,CLDN16,UNQ846	1
4	5108518	5215976	del	STK32B	1
4	74735392	74864580	dup	IL8	1
4	78951173	79103165	dup	CNOT6L,MRPL1	1
4	95781900	96255099	dup	PDLIM5,LOC728442,BMPR1B	1
4	100189816	100342799	del	METAP1,ADH5,ADH4	1

4	113529096	113893894	dup	ALPK1,NEUROG2,LOC91431,C4orf21,LARP7,MIRN367,MIRN302D,MIRN302A,MIRN302C,MIRN302B	1
4	114884938	115470791	dup	CAMK2D,ARSJ	1
4	120432331	120685299	dup	USP53,LOC401152,FABP2,LOC729249,FLJ14186,LOC645513,PDE5A	1
4	144067928	144176340	del	LOC729675	1
4	154588412	155163845	dup	KIAA0922,TLR2,RNF175,SFRP2	1
4	162419238	162568359	del	FSTL5	1
4	177932828	178065111	dup	VEGFC	1
4	178467820	178584108	dup	NEIL3	1
4	184350807	184681169	dup	WWC2,CLDN22,LOC100132463,LOC100131811,CDKN2AIP,ING2	1
5	1695155	1883374	dup	MRPL36,NDUFS6	1
5	2409789	2732511	dup	LOC100133292	1
5	19203456	19757247	dup	CDH18	1
5	20009671	20700523	dup	CDH18	1
5	22346817	23379949	dup	CDH12	1
5	37541744	37660208	dup	WDR70	1
5	58867115	59095265	del	PDE4D,MIRN582	1
5	60152290	60767215	dup	ELOVL7,ERCC8,NDUFA12L,DKFZP686E2158,ZSWIM6,LOC728153	1
5	95457099	95693430	del	LOC441097	1
5	99678948	100293602	dup	LOC100133050,LOC100132020,TMEM157,ST8SIA4	1
5	110682996	110894149	del	CAMK4,STARD4	1
5	126925062	127404930	del	CTXN3,LOC728586	1
5	130387114	130720739	dup	HINT1,LYRM7,CDC42SE2	1
5	180460061	180624927	dup	TRNAL-AAG,TRNAV9,OR2V2,TRNAV-AAC,TRNAV-CAC,TRNAP-UGG,TRNAT1,TRIM7,TRNAA-UGC,TRNAK-CUU,TRIM41,GNB2L1,SNORD96A,SNORD95,TRIM52	1
6	119769	496352	dup	FLJ43763,DUSP22,IRF4,EXOC2,LOC727827,LOC642335	1
6	63009959	63255396	dup	KHDRBS2	1
6	68334871	68703877	del	LOC100128757	1
6	74168491	74593230	dup	DDX43,C6orf150,MTO1,EEF1A1,SLC17A5,CD109	1
6	91481484	94497325	dup	EPHA7	1
6	106604408	106937568	dup	PRDM1,ATG5	2
6	125557330	125906207	dup	TPD52L1,HDDC2	1
6	127357460	127691043	del	RSPO3,RNF146,ECHDC1	1
7	12695794	13112747	dup	ARL4A	1
7	16058262	16586160	dup	LOC729920,SOSTDC1,LOC100129771,LOC100129335	1
7	49540495	49894866	dup	VWC2,LOC100128734	1
7	63200647	63487624	del	LOC730291,ZNF679,LOC728927	1
7	141274408	141407299	del	CLEC5A,TAS2R38,MGAM	1
7	145582575	145743361	del	CNTNAP2	1
7	157369323	157513462	del	PTPRN2	1
8	1987919	2374301	del	MYOM2	1
8	18566441	19801599	dup	PSD3,LOC100131275,LOC100128993,SH2D4A,ChGn,LOC100130604,INTS10	2
8	82552799	82735204	dup	FABP4,LOC646486,LOC100129523,IMPA1	1
8	86461282	86722738	dup	CA1,CA3,LOC100132709,CA2	1
8	87220440	87426294	dup	ATP6V0D2,LOC100128962,SLC7A13,WWP1	1
8	124096689	126212968	del	DERL1,WDR67,TRNAM-CAU,FAM83A,LOC100131726,C8orf76,ZHX1,ATAD2,MIRN548D1,C8orf32,FBXO32,C8ORFK36,ANXA13,FAM91A1,FER1L6,TMEM65,TRMT12,RNF139,TATDN1,NDUFB9,MTSS1,ZNF572,SQLE,KIAA0196,NSMCE2	1

8	138893819	139296115	dup	FLJ45872,FAM135B	1
8	145891814	146125907	dup	ZNF251,ZNF34,RPL8,ZNF517,LOC100130027,LOC100129596,ZNF7,COMMD5,ZNF250	1
9	9423671	9550338	del	PTPRD	1
9	18416823	18776181	del	ADAMTSL1	1
9	70797218	71143539	dup	PIP5K1B,PRKACG,FXN,LOC100131414,TJP2,C9orf61	2
9	71528430	72053813	dup	PTAR1,C9orf135,MAMDC2	1
9	104723219	104885068	del	CYLC2	1
9	118549097	118655895	del	ASTN2	1
9	131966512	132099642	dup	FREQ	1
10	1412160	6132099	dup	ADARB2,LOC642384,C10orf109,LOC100129465,LOC728209,LOC727878,PFKP,LOC100133296,PITRM1,KLF6,LOC100130652,LOC727894,LOC728544,LOC100128356,LOC338588,AKR1CL2,tAKR,AKR1C1,AKR1C2,AKR1C3,AKR1CL1,AKR1C4,UCN3,TUBAL3,NET1,CALML5,CALML3,ASB13,C10orf18,GDI2,TRNAV-UAC,ANKRD16,FBXO18,IL15RA,IL2RA	1
10	29684578	29911570	dup	SVIL,MIRN604	1
10	41956473	43152050	dup	LOC100131936,LOC653097,LOC100132349,CCNYL2,MGC16291,ZNF37B,LOC100128934,ZNF33B,LOC100129622,LOC283028,BMS1,RET,GALNACT-2,RASGEF1A	1
10	47063957	51229942	dup	LOC728684,ANXA8L2,LOC728449,FAM21B,ASAH2C,LOC727950,LOC100132209,CTGLF6,LOC100133093,ANXA8,LOC653110,ZNF488,RBP3,GDF2,GDF10,PTPN20B,FRMPD2L1,LOC644021,LOC644054,LOC100133265,FRMPD2,MAPK8,ARHGAP22,C10orf64,LOC642343,WDFY4,LRRC18,C10orf72,C10orf73,LOC728883,C10orf128,LOC100132730,C10orf71,LOC100130757,DRGX,LOC100128032,ERCC6,PGBD3,CHAT,SLC18A3,C10orf53,OGDHL,LOC727726,FAM21D,LOC728955,CTGLF5,TIMM23B,LOC100133089,CTGLF4,MSMB	1
10	51462803	51815423	dup	FAM21A,ASAH2,SGMS1	1
10	57761600	58753290	del	ZWINT	2
10	65228767	68139445	dup	CTNNA3	2
10	81567623	81966905	dup	LOC642361,FAM22E,SFTPD,LOC642521,LOC642538,C10orf57,PLAC9,ANXA11	1
10	83980248	84561902	del	NRG3	1
10	93824133	94106871	del	CPEB3,LOC100130772,MARCH5	1
10	98850750	99028923	dup	SLIT1,ARHGAP19	1
10	112373593	112648954	dup	RBM20,LOC282997,LOC100132573,PDCD4,LOC92482	1
11	201447	373554	del	RIC8A,SIRT3,PSMD13,NLRP6,ATHL1,IFITM5,IFITM2,IFITM1,IFITM3,B4GALNT4	1
11	201447	430343	dup	RIC8A,SIRT3,PSMD13,NLRP6,ATHL1,IFITM5,IFITM2,IFITM1,IFITM3,B4GALNT4,PKP3,SIGIRR,TMEM16J	1
11	4120756	4232709	dup	LOC196120,LOC390031	1
11	54596906	55739642	dup	TRIM48,OR4A16,OR4A15,OR4C15,OR4C16,OR4C11,OR4P4,OR4S2,OR4C6,OR5D13,OR5D14,OR5L1,OR5D18,OR5L2,OR5D16,SPRYD5,OR5W2,OR5I1,OR10AG1,OR5F1,OR5AS1,OR8I2,OR8H2,OR8H3,OR8J3,OR8K5,OR5J2	2
11	54937013	55112298	del	OR4C15,OR4C16	1
11	68304836	68542057	dup	CPT1A,MRPL21,IGHMBP2,MRGPRD,MRGPRF	1
11	69837338	70074717	dup	PPFIA1,CTTN,SHANK2	1
11	121963463	122151248	dup	UBASH3B	1
11	128156738	128275086	dup	FLI1,KCNJ1,KCNJ5,C11orf45	1
12	971525	1119511	dup	LOC100130219,ERC1	1
12	2022432	2541278	dup	CACNA1C	1
12	5094169	5260067	del	LOC387826	1

12	24357971	24519700	dup	SOX5	1
12	38872898	39028521	del	LRRK2	1
12	52503589	52628593	del	HOXC13	1
12	71297905	71840265	dup	TRHDE	1
12	77701054	77869712	del	LOC100129021	1
12	86803928	87584968	del	C12orf50,C12orf29,CEP290,TMTC3,KITLG	1
12	95329061	95526343	del	C12orf55	1
12	97551176	97756322	dup	IKIP,APAF1,ANKS1B	1
12	100749452	101055184	del	DRAM,CCDC53,NUP37,C12orf48	1
12	121217208	121868133	dup	LRRC43,IL31,B3GNT4,DIABLO,VPS33A,CLIP1,TRNAD-GUC,ZCCHC8,RSRC2,KNTC1,GPR109A,GPR109B,GPR81, DENR,CCDC62	1
12	127358570	127647495	dup	TMEM132C	1
13	30803641	30977080	dup	B3GALTL	1
13	31177504	31439521	dup	RXFP2	1
13	48259408	48519957	del	LOC647131,FNDC3A	1
13	82905273	83506735	dup	SLITRK1	1
14	21604337	21840943	dup	TRA@,TRAV22,TRAV23DV6,TRDV1,TRAV24,TRAV25,TRAV26-1,TRAV8-7,TRAV27,TRAV29DV5,TRAV30,TRAV26-2,TRAV34,TRAV35,TRAV36DV7,TRAV38-1,TRAV38-2DV8	4
14	33465639	33586331	dup	EGLN3	1
14	50241969	50382223	del	NIN	1
14	52385742	52525010	dup	FERMT2	1
15	58226926	58465642	dup	ANXA2	1
15	74654705	74895929	del	SCAPER	1
15	82306210	82555245	dup	ADAMTSL3	1
15	99212626	100060877	del	LOC145757,ALDH1A3,LRRK1,CHSY1,SELS,SNRPA1,PCSK6,TM2D3,TARSL2	1
16	5059873	5217562	del	ALG1,FAM86A	1
16	6490401	6854359	dup	A2BP1,LOC100131413	1
16	7134152	7356599	dup	A2BP1	1
16	8760561	8866827	del	ABAT,TMEM186,PMM2,LOC100132944,LOC100130283,CARHSP1,LOC100129895	1
16	25870606	26088563	dup	HS3ST4	1
16	26422573	27239475	dup	C16orf82,JMJD5,NSMCE1,IL4R	1
16	45086927	45948101	dup	FLJ43980,LOC100128802,SHCBP1,VPS35,ORC6L,MLCK,LOC388272,GPT2,DNAJA2,NETO2,LOC100127930,ITFG1	1
16	76340176	76766570	del	KIAA1576,CLEC3A,WWOX	1
16	76901619	77102849	del	WWOX,LOC645947	1
16	87678937	87804936	del	ACSF3,C16orf81,LOC400558,CDH15,LOC146429	1
17	9975140	10275516	dup	GAS7,LOC100129677,MYH13,LOC100128560,MYH8	1
17	14093529	15357533	dup	HS3ST3B1,LOC100131109,LOC388339,FLJ45831,PMP22,TEKT3,CDRT4,FAM18B2	1
17	16659493	17062349	dup	LOC100129981,LOC96597,LOC100129535,TNFRSF13B,LOC100128283,M-RIP,LOC201164,FLCN	1
17	31923810	33308219	dup	ZNHIT3,MYO19,PIGW,GGNBP2,MGC4172,MRM1,LOC727862,LHX1,AATF,ACACA,C17orf78,TADA2L,DUSP14,AP1GBP1,LOC100131822,DDX52,HNF1B,LOC284100	1
17	78476537	78599918	dup	TBCD,B3GNTL1	1
18	12246238	12506082	dup	CIDEA,TUBB6,AFG3L2,SLMO1,SPIRE1	1
18	14597916	15092421	dup	ANKRD30B,LOC647983,LOC100131500	3
18	28682047	29322188	dup	C18orf34	1
18	45714083	45940515	del	MYO5B	1

19	7035307	7181438	del	ZNF557,LOC100131165,INSR,LOC100128567	1
19	19969224	20657868	dup	ZNF682,ZNF90,LOC729903,ZNF486,FLJ44894,ZNF626	5
19	23605608	24295825	dup	ZNF675,ZNF681,LOC730087,ZNF254	3
19	44819692	45042569	del	LOC148003,LOC400696,LGALS14,CLC,DYRK1B,FBL	1
19	59886999	60014905	del	KIR3DL3,KIR2DL3,KIR2DL1,KIR2DL4	1
20	116466	352407	del	DEFB128,DEFB129,DEFB32,C20orf96,ZCCHC3,SOX12,NRSN2,TRIB3,RBCK1	1
20	25794494	26236535	dup	C20orf91,LOC284801,MIRN663	1
22	24873287	25340497	dup	SEZ6L,FLJ38343,ASPHD2,LOC100128401,HPS4,SRRD,TFIP11,TPST2,LOC100130561,CRYBB1	1
23	142664	2770060	del	PLCXD1,GTPBP6,PPP2R3B,SHOX,LOC442442,CRLF2,CSF2RA,IL3RA,SLC25A6,LOC729629,CXYorf2,ASMTL,P2RY8,SFRS17A,ASMT,DHRSX,ZBED1,LOC100130595,LOC401577,CD99,XG,GYG2	5
23	22008574	22149641	dup	PHEX	1
23	35570709	35917535	dup	MAGEB16,CXorf22	1
23	37739464	37857804	dup	SYTL5	1
23	56824308	57065815	dup	SPIN3	1
23	86509912	86714253	dup	KLHL4	1
23	139671186	139903522	dup	CDR1	1
23	143459775	146067963	del	SPANXN1,SLITRK2,LOC100129095,CXorf1,MIRN890,MIRN888,MIRN892A,MIRN892B,MIRN891B,MIRN891A,LOC100128265,LOC100133053,LOC100129239,LOC100132556	1
23	144046112	144430381	dup	SPANXN1	1
23	151732760	151922021	dup	CETN2,NSDHL,ZNF185,PNMA5	1

Bibliography

Abrahams, B. S. and Geschwind, D. H. 2008. Advances in autism genetics: on the threshold of a new neurobiology. *Nat Rev Genet* 9(5), pp. 341-355.

Affymetrix Inc 2007. Affymetrix Genotyping Console 2.1 User Manual.

Akiskal, H. S. 2009. *Kaplan & Sadock's Comprehensive Textbook of Psychiatry*. 9th ed. Philadelphia Wolters Kluwer, Lippincott Williams&Wilkins.

Alaerts, M. and Del-Favero, J. 2009. Searching genetic risk factors for schizophrenia and bipolar disorder: learn from the past and back to the future. *Hum Mutat* 30(8), pp. 1139-1152.

Allen, M. G. et al. 1974. Affective illness in veteran twins: a diagnostic review. *Am J Psychiatry* 131(11), pp. 1234-1239.

Altman, D. G. 1991. *Practical statistics for medical research* 1st ed. Chapman&Hall.

American Psychiatric Association 1994. *Diagnostic and Statistical Manual for Mental Disorders*. Washington DC: American Psychiatric Association.

Andreasen, N. C. 1995. Symptoms, signs, and diagnosis of schizophrenia. *Lancet* 346(8973), pp. 477-481.

Antonarakis, S. E. and Beckmann, J. S. 2006. Mendelian disorders deserve more attention. *Nat Rev Genet* 7(4), pp. 277-282.

Armengol, L. et al. 2009. Identification of copy number variants defining genomic differences among major human groups. *PLoS One* 4(9), p. e7230.

Baker, M. 2010. Genomics: The search for association. *Nature* 467(7319), pp. 1135-1138.

Barnes, C. et al. 2008. A robust statistical method for case-control association testing with copy number variation. *Nat Genet* 40(10), pp. 1245-1252.

Barnett, J. H. et al. 2006. Cognitive reserve in neuropsychiatry. *Psychol Med* 36(8), pp. 1053-1064.

Barnett, J. H. and Smoller, J. W. 2009. The genetics of bipolar disorder. *Neuroscience* 164(1), pp. 331-343.

Barrett, J. C. and Cardon, L. R. 2006. Evaluating coverage of genome-wide association studies. *Nat Genet* 38(6), pp. 659-662.

Bassett, A. S. and Chow, E. W. 2008. Schizophrenia and 22q11.2 deletion syndrome. *Curr Psychiatry Rep* 10(2), pp. 148-157.

Baum, A. E. et al. 2008a. A genome-wide association study implicates diacylglycerol kinase eta (*DGKH*) and several other genes in the etiology of bipolar disorder. *Mol Psychiatry* 13(2), pp. 197-207.

Baum, A. E. et al. 2008b. Meta-analysis of two genome-wide association studies of bipolar disorder reveals important points of agreement. *Mol Psychiatry* 13(5), pp. 466-467.

Bearden, C. E. et al. 2001. The neurocognitive phenotype of the 22q11.2 deletion syndrome: selective deficit in visual-spatial memory. *J Clin Exp Neuropsychol* 23(4), pp. 447-464.

Berrettini, W. H. 2000. Are schizophrenic and bipolar disorders related? A review of family and molecular studies. *Biol Psychiatry* 48(6), pp. 531-538.

Bertelsen, A. et al. 1977. A Danish twin study of manic-depressive disorders. *Br J Psychiatry* 130, pp. 330-351.

Bigos, K. L. et al. 2010. Genetic variation in CACNA1C affects brain circuitries related to mental illness. *Arch Gen Psychiatry* 67(9), pp. 939-945.

Bijlsma, E. K. et al. 2009. Extending the phenotype of recurrent rearrangements of 16p11.2: deletions in mentally retarded patients without autism and in normal individuals. *Eur J Med Genet* 52(2-3), pp. 77-87.

Blackwood, D. H. et al. 2001. Schizophrenia and affective disorders--cosegregation with a translocation at chromosome 1q42 that directly disrupts brain-expressed genes: clinical and P300 findings in a family. *Am J Hum Genet* 69(2), pp. 428-433.

Blackwood, D. H. et al. 1996. A locus for bipolar affective disorder on chromosome 4p. *Nat Genet* 12(4), pp. 427-430.

Bodmer, W. and Bonilla, C. 2008. Common and rare variants in multifactorial susceptibility to common diseases. *Nat Genet* 40(6), pp. 695-701.

Bramon, E. and Sham, P. C. 2001. The common genetic liability between schizophrenia and bipolar disorder: a review. *Curr Psychiatry Rep* 3(4), pp. 332-337.

Branzei, D. and Foiani, M. 2007. Template switching: from replication fork repair to genome rearrangements. *Cell* 131(7), pp. 1228-1230.

Brunetti-Pierri, N. et al. 2008. Recurrent reciprocal 1q21.1 deletions and duplications associated with microcephaly or macrocephaly and developmental and behavioral abnormalities. *Nat Genet* 40(12), pp. 1466-1471.

Bucher, K. D. et al. 1981. The transmission of manic depressive illness--II. Segregation analysis of three sets of family data. *J Psychiatr Res* 16(1), pp. 65-78.

Buetow, K. H. et al. 2001. High-throughput development and characterization of a genomewide collection of gene-based single nucleotide polymorphism markers by

chip-based matrix-assisted laser desorption/ionization time-of-flight mass spectrometry. *Proc Natl Acad Sci U S A* 98(2), pp. 581-584.

Burdick, K. E. et al. 2009. The role of general intelligence as an intermediate phenotype for neuropsychiatric disorders. *Cogn Neuropsychiatry* 14(4-5), pp. 299-311.

Burmeister, M. et al. 2008. Psychiatric genetics: progress amid controversy. *Nat Rev Genet* 9(7), pp. 527-540.

Cardno, A. G. et al. 1999. Heritability estimates for psychotic disorders: the Maudsley twin psychosis series. *Arch Gen Psychiatry* 56(2), pp. 162-168.

Carlson, C. et al. 1997. Molecular analysis of velo-cardio-facial syndrome patients with psychiatric disorders. *Am J Hum Genet* 60(4), pp. 851-859.

Carter, N. P. 2007. Methods and strategies for analyzing copy number variation using DNA microarrays. *Nat Genet* 39(7 Suppl), pp. S16-21.

Chakravarti, A. 1999. Population genetics--making sense out of sequence. *Nat Genet* 21(1 Suppl), pp. 56-60.

Chance, P. F. et al. 1993. DNA deletion associated with hereditary neuropathy with liability to pressure palsies. *Cell* 72(1), pp. 143-151.

Chen, P. L. et al. 2009. Fine mapping on chromosome 10q22-q23 implicates Neuregulin 3 in schizophrenia. *Am J Hum Genet* 84(1), pp. 21-34.

Christian, S. L. et al. 2008. Novel submicroscopic chromosomal abnormalities detected in autism spectrum disorder. *Biol Psychiatry* 63(12), pp. 1111-1117.

Christiansen, J. et al. 2004. Chromosome 1q21.1 contiguous gene deletion is associated with congenital heart disease. *Circ Res* 94(11), pp. 1429-1435.

Chubb, J. E. et al. 2008. The DISC locus in psychiatric illness. *Mol Psychiatry* 13(1), pp. 36-64.

Chubykin, A. A. et al. 2005. Dissection of synapse induction by neuroligins: effect of a neuroligin mutation associated with autism. *J Biol Chem* 280(23), pp. 22365-22374.

Cichon, S. et al. 2009. Genomewide association studies: history, rationale, and prospects for psychiatric disorders. *Am J Psychiatry* 166(5), pp. 540-556.

Cohen, J. 2007. Genomics. DNA duplications and deletions help determine health. *Science* 317(5843), pp. 1315-1317.

Cook, E. H., Jr. and Scherer, S. W. 2008. Copy-number variations associated with neuropsychiatric conditions. *Nature* 455(7215), pp. 919-923.

Cooke, S. F. and Bliss, T. V. 2006. Plasticity in the human central nervous system. *Brain* 129(Pt 7), pp. 1659-1673.

Cooper, G. M. et al. 2007. Mutational and selective effects on copy-number variants in the human genome. *Nat Genet* 39(7 Suppl), pp. S22-29.

Cooper, G. M. et al. 2008. Systematic assessment of copy number variant detection via genome-wide SNP genotyping. *Nat Genet* 40(10), pp. 1199-1203.

Craddock, N. et al. 2001. Association studies of bipolar disorder. *Bipolar Disord* 3(6), pp. 284-298.

Craddock, N. and Jones, I. 1999. Genetics of bipolar disorder. *J Med Genet* 36(8), pp. 585-594.

Craddock, N. et al. 2004. The Bipolar Affective Disorder Dimension Scale (BADDS)--a dimensional scale for rating lifetime psychopathology in bipolar spectrum disorders. *BMC Psychiatry* 4, p. 19.

Craddock, N. et al. 2010. Strong genetic evidence for a selective influence of GABAA receptors on a component of the bipolar disorder phenotype. *Mol Psychiatry* 15(2), pp. 146-153.

Craddock, N. et al. 2005. The genetics of schizophrenia and bipolar disorder: dissecting psychosis. *J Med Genet* 42(3), pp. 193-204.

Craddock, N. et al. 2009. Psychosis genetics: modeling the relationship between schizophrenia, bipolar disorder and mixed (or "schizoaffective") psychoses. *Schizophrenia Bulletin* (in press).

Craddock, N. and Owen, M. J. 2005. The beginning of the end for the Kraepelinian dichotomy. *Br J Psychiatry* 186, pp. 364-366.

Craddock, N. and Owen, M. J. 2010. The Kraepelinian dichotomy - going, going... but still not gone. *Br J Psychiatry* 196(2), pp. 92-95.

Craddock, N. and Sklar, P. 2009. Genetics of bipolar disorder: successful start to a long journey. *Trends Genet* 25(2), pp. 99-105.

de Kovel, C. G. et al. 2010. Recurrent microdeletions at 15q11.2 and 16p13.11 predispose to idiopathic generalized epilepsies. *Brain* 133(Pt 1), pp. 23-32.

de Vries, B. B. et al. 2005. Diagnostic genome profiling in mental retardation. *Am J Hum Genet* 77(4), pp. 606-616.

De Witte, A. et al. 2006. 60-mer Oligo-Based Comparative Genomic Hybridization. [Online]. Available at: www.chem.agilent.com.

Dean, C. and Dresbach, T. 2006. Neuroligins and neurexins: linking cell adhesion, synapse formation and cognitive function. *Trends Neurosci* 29(1), pp. 21-29.

Dibbens, L. M. et al. 2009. Familial and sporadic 15q13.3 microdeletions in idiopathic generalized epilepsy: precedent for disorders with complex inheritance. *Hum Mol Genet* 18(19), pp. 3626-3631.

Doornbos, M. et al. 2009. Nine patients with a microdeletion 15q11.2 between breakpoints 1 and 2 of the Prader-Willi critical region, possibly associated with behavioural disturbances. *Eur J Med Genet* 52(2-3), pp. 108-115.

Durand, C. M. et al. 2007. Mutations in the gene encoding the synaptic scaffolding protein SHANK3 are associated with autism spectrum disorders. *Nat Genet* 39(1), pp. 25-27.

Elmslie, F. V. et al. 1997. Genetic mapping of a major susceptibility locus for juvenile myoclonic epilepsy on chromosome 15q. *Hum Mol Genet* 6(8), pp. 1329-1334.

Ensenauer, R. E. et al. 2003. Microduplication 22q11.2, an emerging syndrome: clinical, cytogenetic, and molecular analysis of thirteen patients. *Am J Hum Genet* 73(5), pp. 1027-1040.

Erk, S. et al. 2010. Brain function in carriers of a genome-wide supported bipolar disorder variant. *Arch Gen Psychiatry* 67(8), pp. 803-811.

Estivill, X. and Armengol, L. 2007. Copy number variants and common disorders: filling the gaps and exploring complexity in genome-wide association studies. *PLoS Genet* 3(10), pp. 1787-1799.

Falconer, D. S. and Mackay, T. F. C. 1996. *Introduction to Quantitative Genetics*. 4th ed. Harlow: Addison Wesley Longman.

Falk, R. E. and Casas, K. A. 2007. Chromosome 2q37 deletion: clinical and molecular aspects. *Am J Med Genet C Semin Med Genet* 145C(4), pp. 357-371.

Fernandez, B. A. et al. 2009. Phenotypic Spectrum Associated with De Novo and Inherited Deletions and Duplications at 16p11.2 in Individuals Ascertained for Diagnosis of Autism Spectrum Disorder. *J Med Genet.*

Ferreira, M. A. et al. 2008. Collaborative genome-wide association analysis supports a role for *ANK3* and *CACNA1C* in bipolar disorder. *Nat Genet.*

Feuk, L. et al. 2006a. Structural variation in the human genome. *Nat Rev Genet* 7(2), pp. 85-97.

Feuk, L. et al. 2006b. Structural variants: changing the landscape of chromosomes and design of disease studies. *Hum Mol Genet* 15 Spec No 1, pp. R57-66.

Fidler, D. J. et al. 2000. Macrocephaly in autism and other pervasive developmental disorders. *Dev Med Child Neurol* 42(11), pp. 737-740.

Field, A. 2009. *Discovering statistics using SPSS*. 3rd ed. London: Sage Publications Ltd.

Flicek, P. et al. 2010. Ensembl's 10th year. *Nucleic Acids Res* 38(Database issue), pp. D557-562.

Flint, J. et al. 1986. High frequencies of alpha-thalassaemia are the result of natural selection by malaria. *Nature* 321(6072), pp. 744-750.

Frazer, K. A. et al. 2007. A second generation human haplotype map of over 3.1 million SNPs. *Nature* 449(7164), pp. 851-861.

Freedman, R. et al. 1997. Linkage of a neurophysiological deficit in schizophrenia to a chromosome 15 locus. *Proc Natl Acad Sci U S A* 94(2), pp. 587-592.

Freeman, J. L. et al. 2006. Copy number variation: new insights in genome diversity. *Genome Res* 16(8), pp. 949-961.

Freimer, N. B. et al. 1996. An approach to investigating linkage for bipolar disorder using large Costa Rican pedigrees. *Am J Med Genet* 67(3), pp. 254-263.

Freitag, C. M. 2007. The genetics of autistic disorders and its clinical relevance: a review of the literature. *Mol Psychiatry* 12(1), pp. 2-22.

Friedman, J. et al. 2009. Detection of pathogenic copy number variants in children with idiopathic intellectual disability using 500 K SNP array genomic hybridization. *BMC Genomics* 10, p. 526.

Friedman, J. I. et al. 2008. *CNTNAP2* gene dosage variation is associated with schizophrenia and epilepsy. *Mol Psychiatry* 13(3), pp. 261-266.

Friedman, J. M. et al. 2006. Oligonucleotide microarray analysis of genomic imbalance in children with mental retardation. *Am J Hum Genet* 79(3), pp. 500-513.

Gardina, P. J. et al. 2008. Ploidy status and copy number aberrations in primary glioblastomas defined by integrated analysis of allelic ratios, signal ratios and loss of heterozygosity using 500K SNP Mapping Arrays. *BMC Genomics* 9, p. 489.

Gershon, E. S. 2000. Bipolar illness and schizophrenia as oligogenic diseases: implications for the future. *Biol Psychiatry* 47(3), pp. 240-244.

Glessner, J. T. et al. 2009. Autism genome-wide copy number variation reveals ubiquitin and neuronal genes. *Nature* 459(7246), pp. 569-573.

Goldin, L. R. et al. 1983. Segregation and linkage analyses in families of patients with bipolar, unipolar, and schizoaffective mood disorders. *Am J Hum Genet* 35(2), pp. 274-287.

Gonzalez, E. et al. 2005. The influence of CCL3L1 gene-containing segmental duplications on HIV-1/AIDS susceptibility. *Science* 307(5714), pp. 1434-1440.

Goodwin, F. K. and Jamison, K. R. 2007. *Manic-Depressive illness Bipolar Disorders and Recurrent Depression* 2nd ed. New York: Oxford University Press, Inc.

Gorlov, I. P. et al. 2008. Shifting paradigm of association studies: value of rare single-nucleotide polymorphisms. *Am J Hum Genet* 82(1), pp. 100-112.

Gottesman, II et al. 2010. Severe mental disorders in offspring with 2 psychiatrically ill parents. *Arch Gen Psychiatry* 67(3), pp. 252-257.

Green, E. K. et al. 2009. The bipolar disorder risk allele at CACNA1C also confers risk of recurrent major depression and of schizophrenia. *Mol Psychiatry*.

Grozeva, D. et al. 2010. Rare copy number variants: a point of rarity in genetic risk for bipolar disorder and schizophrenia. *Arch Gen Psychiatry* 67(4), pp. 318-327.

Gu, W. et al. 2008. Mechanisms for human genomic rearrangements. *Pathogenetics* 1(1), p. 4.

Guilmatre, A. et al. 2009. Recurrent rearrangements in synaptic and neurodevelopmental genes and shared biologic pathways in schizophrenia, autism, and mental retardation. *Arch Gen Psychiatry* 66(9), pp. 947-956.

Hannes, F. D. et al. 2009. Recurrent reciprocal deletions and duplications of 16p13.11: the deletion is a risk factor for MR/MCA while the duplication may be a rare benign variant. *J Med Genet* 46(4), pp. 223-232.

Harvey, A. G. 2008. Sleep and circadian rhythms in bipolar disorder: seeking synchrony, harmony, and regulation. *Am J Psychiatry* 165(7), pp. 820-829.

Hayden, E. P. and Nurnberger, J. I., Jr. 2006. Molecular genetics of bipolar disorder. *Genes Brain Behav* 5(1), pp. 85-95.

Hehir-Kwa, J. Y. et al. 2007. Genome-wide copy number profiling on high-density bacterial artificial chromosomes, single-nucleotide polymorphisms, and oligonucleotide microarrays: a platform comparison based on statistical power analysis. *DNA Res* 14(1), pp. 1-11.

Helbig, I. et al. 2009. 15q13.3 microdeletions increase risk of idiopathic generalized epilepsy. *Nat Genet* 41(2), pp. 160-162.

Henrichsen, C. N. et al. 2009a. Copy number variants, diseases and gene expression. *Hum Mol Genet* 18(R1), pp. R1-8.

Henrichsen, C. N. et al. 2009b. Segmental copy number variation shapes tissue transcriptomes. *Nat Genet* 41(4), pp. 424-429.

Hirschhorn, J. N. and Daly, M. J. 2005. Genome-wide association studies for common diseases and complex traits. *Nat Rev Genet* 6(2), pp. 95-108.

Hodgkinson, S. et al. 1987. Molecular genetic evidence for heterogeneity in manic depression. *Nature* 325(6107), pp. 805-806.

Hogart, A. et al. 2008. The comorbidity of autism with the genomic disorders of chromosome 15q11.2-q13. *Neurobiol Dis* 38(2), pp. 181-191.

Hong, C. J. et al. 2004. Association study of the human partially duplicated alpha7 nicotinic acetylcholine receptor genetic variant with bipolar disorder. *Neurosci Lett* 355(1-2), pp. 69-72.

http://www.affymetrix.com CNAT 4.0: Copy Number and Loss of Heterozygosity Estimation Algortithms for the GeneChip® Human Mapping 10/50/100/250/500K Array set, Affymetrix white paper.http://www.affymetrix.com.

Iafrate, A. J. et al. 2004. Detection of large-scale variation in the human genome. *Nat Genet* 36(9), pp. 949-951.

Ikeda, M. et al. 2010. Copy number variation in schizophrenia in the Japanese population. *Biol Psychiatry* 67(3), pp. 283-286.

Ingason, A. et al. 2010. Maternally derived microduplications at 15q11-q13 may implicate imprinted genes in psychotic illness. *Am J Psychiatry, accepted.*

Ingason, A. et al. 2009. Copy number variations of chromosome 16p13.1 region associated with schizophrenia. *Mol Psychiatry.*

International HapMap Consortium 2005. A haplotype map of the human genome. *Nature* 437(7063), pp. 1299-1320.

International Schizophrenia Consortium 2008. Rare chromosomal deletions and duplications increase risk of schizophrenia. *Nature* 455(7210), pp. 237-241.

International Schizophrenia Consortium et al. 2009. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* 460(7256), pp. 748-752.

Itsara, A. et al. 2009. Population analysis of large copy number variants and hotspots of human genetic disease. *Am J Hum Genet* 84(2), pp. 148-161.

Ivanov, D. et al. 2003. Chromosome 22q11 deletions, velo-cardio-facial syndrome and early-onset psychosis. Molecular genetic study. *Br J Psychiatry* 183, pp. 409-413.

Jacobs, P. A. et al. 1992. Estimates of the frequency of chromosome abnormalities detectable in unselected newborns using moderate levels of banding. *J Med Genet* 29(2), pp. 103-108.

Jacquemont, M. L. et al. 2006. Array-based comparative genomic hybridisation identifies high frequency of cryptic chromosomal rearrangements in patients with syndromic autism spectrum disorders. *J Med Genet* 43(11), pp. 843-849.

Jager, D. et al. 2001. Identification of a tissue-specific putative transcription factor in breast tissue by serological screening of a breast cancer library. *Cancer Res* 61(5), pp. 2055-2061.

John, J. P. et al. 2008. Expanding the schizophrenia phenotype: a composite evaluation of neurodevelopmental markers. *Compr Psychiatry* 49(1), pp. 78-86.

Johnson, S. L. 2005. Life events in bipolar disorder: towards more specific models. *Clin Psychol Rev* 25(8), pp. 1008-1027.

Jones, I. and Craddock, N. 2001. Candidate gene studies of bipolar disorder. *Ann Med* 33(4), pp. 248-256.

Jones, I. et al. 2002. *Psychiatric Genetics and Genomics*. Oxford: Oxford University Press.

Kanehisa, M. and Goto, S. 2000. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* 28(1), pp. 27-30.

Kanehisa, M. et al. 2006. From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res* 34(Database issue), pp. D354-357.

Keefe, R. S. E. and Eesley, C. 2009. *Kaplan & Sadock's Comprehensive Textbook of Psychiatry*. 9th ed. Philadelphia Wolters Kluwer, Lippincott Williams&Wilkins.

Kelsoe, J. R. 2009. *Kaplan & Sadock's Comprehensive Textbook of Psychiatry*. 9th ed. Philadelphia Wolters Kluwer, Lippincott Williams&Wilkins.

Kendler, K. S. 1986. Kraepelin and the differential diagnosis of dementia praecox and manic-depressive insanity. *Compr Psychiatry* 27(6), pp. 549-558.

Kendler, K. S. et al. 1993. A pilot Swedish twin study of affective illness, including hospital- and population-ascertained subsamples. *Arch Gen Psychiatry* 50(9), pp. 699-700.

Kieseppa, T. et al. 2004. High concordance of bipolar I disorder in a nationwide sample of twins. *Am J Psychiatry* 161(10), pp. 1814-1821.

Kim, H. G. et al. 2008. Disruption of neurexin 1 associated with autism spectrum disorder. *Am J Hum Genet* 82(1), pp. 199-207.

Kirov, G. 2010. The role of copy number variation in schizophrenia. *Expert Rev Neurother* 10(1), pp. 25-32.

Kirov, G. et al. 2009a. Support for the involvement of large cnvs in the pathogenesis of schizophrenia. *Hum Mol Genet.*

Kirov, G. et al. 2008. Comparative genome hybridization suggests a role for *NRXN1* and *APBA2* in schizophrenia. *Hum Mol Genet* 17(3), pp. 458-465.

Kirov, G. et al. 2009b. Neurexin 1 (NRXN1) deletions in schizophrenia. *Schizophr Bull* 35(5), pp. 851-854.

Kleinjan, D. A. and van Heyningen, V. 2005. Long-range control of gene expression: emerging mechanisms and disruption in disease. *Am J Hum Genet* 76(1), pp. 8-32.

Koolen, D. A. et al. 2008. Clinical and molecular delineation of the 17q21.31 microdeletion syndrome. *J Med Genet* 45(11), pp. 710-720.

Koolen, D. A. et al. 2006. A new chromosome 17q21.31 microdeletion syndrome associated with a common inversion polymorphism. *Nat Genet* 38(9), pp. 999-1001.

Kooy, R. F. 2010. Distinct disorders affecting the brain share common genetic origins. *F1000 Biol Rep* 2.

Kringlen, E. 1967. *Heredity and Environment in the Functional Psychoses*. London: William Heinemann.

Kruglyak, L. and Nickerson, D. A. 2001. Variation is the spice of life. *Nat Genet* 27(3), pp. 234-236.

Kumar, R. A. et al. 2008. Recurrent 16p11.2 microdeletions in autism. *Hum Mol Genet* 17(4), pp. 628-638.

Lachman, H. M. et al. 2007. Increase in *GSK3beta* gene copy number variation in bipolar disorder. *Am J Med Genet B Neuropsychiatr Genet* 144B(3), pp. 259-265.

Lainhart, J. E. et al. 2006. Head circumference and height in autism: a study by the Collaborative Program of Excellence in Autism. *Am J Med Genet A* 140(21), pp. 2257-2274.

Lander, E. and Kruglyak, L. 1995. Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. *Nat Genet* 11(3), pp. 241-247.

Leckman, J. F. et al. 1982. Best estimate of lifetime psychiatric diagnosis: a methodological study. *Arch Gen Psychiatry* 39(8), pp. 879-883.

Lee, J. A. et al. 2007. A DNA replication mechanism for generating nonrecurrent rearrangements associated with genomic disorders. *Cell* 131(7), pp. 1235-1247.

Lee, J. A. et al. 2006a. Role of genomic architecture in PLP1 duplication causing Pelizaeus-Merzbacher disease. *Hum Mol Genet* 15(14), pp. 2250-2265.

Lee, J. A. et al. 2006b. Spastic paraplegia type 2 associated with axonal neuropathy and apparent PLP1 position effect. *Ann Neurol* 59(2), pp. 398-403.

Lesch, K. P. et al. 2008. Molecular genetics of adult ADHD: converging evidence from genome-wide association and extended pedigree linkage studies. *J Neural Transm* 115(11), pp. 1573-1585.

Levy, S. et al. 2007. The diploid genome sequence of an individual human. *PLoS Biol* 5(10), p. e254.

Lichtenstein, P. et al. 2009. Common genetic determinants of schizophrenia and bipolar disorder in Swedish families: a population-based study. *Lancet* 373(9659), pp. 234-239.

Lieber, M. R. et al. 2003. Mechanism and regulation of human non-homologous DNA end-joining. *Nat Rev Mol Cell Biol* 4(9), pp. 712-720.

Lupski, J. R. 1998. Genomic disorders: structural features of the genome can lead to DNA rearrangements and human disease traits. *Trends Genet* 14(10), pp. 417-422.

Lupski, J. R. 2009. Genomic disorders ten years on. *Genome Med* 1(4), p. 42.

Lupski, J. R. et al. 1991. DNA duplication associated with Charcot-Marie-Tooth disease type 1A. *Cell* 66(2), pp. 219-232.

Lupski, J. R. and Stankiewicz, P. 2005. Genomic disorders: molecular mechanisms for rearrangements and conveyed phenotypes. *PLoS Genet* 1(6), p. e49.

Malhi, G. S. et al. 2009. Is lithium in a class of its own? A brief profile of its clinical use. *Aust N Z J Psychiatry* 43(12), pp. 1096-1104.

Mansour, H. A. et al. 2005. Circadian genes and bipolar disorder. *Ann Med* 37(3), pp. 196-205.

Marshall, C. R. et al. 2008. Structural variation of chromosomes in autism spectrum disorder. *Am J Hum Genet* 82(2), pp. 477-488.

McCarroll, S. A. and Altshuler, D. M. 2007. Copy-number variation and association studies of human disease. *Nat Genet* 39(7 Suppl), pp. S37-42.

McCarroll, S. A. et al. 2008. Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nat Genet* 40(10), pp. 1166-1174.

McCarthy, S. E. et al. 2009. Microduplications of 16p11.2 are associated with schizophrenia. *Nat Genet* 41(11), pp. 1223-1227.

McClellan, J. M. et al. 2007. Schizophrenia: a common disease caused by multiple rare alleles. *Br J Psychiatry* 190, pp. 194-199.

McGuffin, P. et al. 1991. A polydiagnostic application of operational criteria in studies of psychotic illness. Development and reliability of the OPCRIT system. *Arch Gen Psychiatry* 48(8), pp. 764-770.

McGuffin, P. and Katz, R. 1986. *Nature, nurture and affective disorder. In the Biology of Depression* London: Gaskell.

McGuffin, P. et al. 1994. *Seminars in Psychiatric Genetics*. London: Gaskell.

McGuffin, P. et al. 2003. The heritability of bipolar affective disorder and the genetic relationship to unipolar depression. *Arch Gen Psychiatry* 60(5), pp. 497-502.

McInnis, M. G. et al. 1993. Anticipation in bipolar affective disorder. *Am J Hum Genet* 53(2), pp. 385-390.

McIntosh, A. M. et al. 2008. The effects of a neuregulin 1 variant on white matter density and integrity. *Mol Psychiatry* 13(11), pp. 1054-1059.

McMahon, F. J. et al. 1995. Patterns of maternal transmission in bipolar affective disorder. *Am J Hum Genet* 56(6), pp. 1277-1286.

McMullan, D. J. et al. 2009. Molecular karyotyping of patients with unexplained mental retardation by SNP arrays: a multicenter study. *Hum Mutat* 30(7), pp. 1082-1092.

McQuillin, A. et al. 2010. Analysis of genetic deletions and duplications in the University College London bipolar disorder case control sample. *Eur J Hum Genet, submitted.*

Mefford, H. C. et al. 2009. A method for rapid, targeted CNV genotyping identifies rare variants associated with neurocognitive disease. *Genome Res* 19(9), pp. 1579-1585.

Mefford, H. C. et al. 2008. Recurrent rearrangements of chromosome 1q21.1 and variable pediatric phenotypes. *N Engl J Med* 359(16), pp. 1685-1699.

Merla, G. et al. 2006. Submicroscopic deletion in patients with Williams-Beuren syndrome influences expression levels of the nonhemizygous flanking genes. *Am J Hum Genet* 79(2), pp. 332-341.

Millar, J. K. et al. 2000. Disruption of two novel genes by a translocation co-segregating with schizophrenia. *Hum Mol Genet* 9(9), pp. 1415-1423.

Miller, D. T. et al. 2009. Microdeletion/duplication at 15q13.2q13.3 among individuals with features of autism and other neuropsychiatric disorders. *J Med Genet* 46(4), pp. 242-248.

Miller, O. J. and Therman, E. 2001. *Human Chromosomes*. 4th ed. Springer.

Mohyuddin, A. et al. 2004. Genetic instability in EBV-transformed lymphoblastoid cell lines. *Biochim Biophys Acta* 1670(1), pp. 81-83.

Mortensen, P. B. et al. 2003. Individual and familial risk factors for bipolar affective disorders in Denmark. *Arch Gen Psychiatry* 60(12), pp. 1209-1215.

Moskvina, V. et al. 2009. Gene-wide analyses of genome-wide association data sets: evidence for multiple common risk alleles for schizophrenia and bipolar disorder and for overlap in genetic risk. *Mol Psychiatry* 14(3), pp. 252-260.

Murphy, K. C. et al. 1999. High rates of schizophrenia in adults with velo-cardio-facial syndrome. *Arch Gen Psychiatry* 56(10), pp. 940-945.

Murphy, K. C. and Owen, M. J. 2001. Velo-cardio-facial syndrome: a model for understanding the genetics and pathogenesis of schizophrenia. *Br J Psychiatry* 179, pp. 397-402.

Murray, R. M. et al. 2004. A developmental model for similarities and dissimilarities between schizophrenia and bipolar disorder. *Schizophr Res* 71(2-3), pp. 405-416.

Murthy, S. K. et al. 2007. Detection of a novel familial deletion of four genes between BP1 and BP2 of the Prader-Willi/Angelman syndrome critical region by oligo-array CGH in a child with neurological disorder and speech impairment. *Cytogenet Genome Res* 116(1-2), pp. 135-140.

Need, A. C. et al. 2009. A genome-wide investigation of SNPs and CNVs in schizophrenia. *PLoS Genet* 5(2), p. e1000373.

Newton-Cheh, C. and Hirschhorn, J. N. 2005. Genetic association studies of complex traits: design and analysis issues. *Mutat Res* 573(1-2), pp. 54-69.

Nguyen, D. Q. et al. 2006. Bias of selection on human copy-number variants. *PLoS Genet* 2(2), p. e20.

O'Donovan, M. C. et al. 2008. Identification of loci associated with schizophrenia by genome-wide association and follow-up. *Nat Genet* 40(9), pp. 1053-1055.

Olshen, A. B. et al. 2004. Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics* 5(4), pp. 557-572.

Oskarsdottir, S. et al. 2004. Incidence and prevalence of the 22q11 deletion syndrome: a population-based study in Western Sweden. *Arch Dis Child* 89(2), pp. 148-151.

Ou, Z. et al. 2008. Microduplications of 22q11.2 are frequently inherited and are associated with variable phenotypes. *Genet Med* 10(4), pp. 267-277.

Ozeki, A. et al. 2008. A Case of Schizophrenia with Chromosomal Microdeletion of 17p11.2
Containing a Myelin-Related Gene PMP22. *The Open Psychiatry Journal* 2.

Pagnamenta, A. T. et al. 2008. A 15q13.3 microdeletion segregating with autism. *Eur J Hum Genet.*

Papolos, D. F. et al. 1996. Bipolar spectrum disorders in patients diagnosed with velo-cardio-facial syndrome: does a hemizygous deletion of chromosome 22q11 result in bipolar affective disorder? *Am J Psychiatry* 153(12), pp. 1541-1547.

Pauls, D. L. et al. 1992. Risks of affective illness among first-degree relatives of bipolar I old-order Amish probands. *Arch Gen Psychiatry* 49(9), pp. 703-708.

Perry, G. H. et al. 2007. Diet and the evolution of human amylase gene copy number variation. *Nat Genet* 39(10), pp. 1256-1260.

Pinto, D. et al. 2007. Copy-number variation in control population cohorts. *Hum Mol Genet* 16 Spec No. 2, pp. R168-173.

Portnoi, M. F. et al. 2005. 22q11.2 duplication syndrome: two new familial cases with some overlapping features with DiGeorge/velocardiofacial syndromes. *Am J Med Genet A* 137(1), pp. 47-51.

Power, C. and Elliott, J. 2006. Cohort profile: 1958 British birth cohort (National Child Development Study). *Int J Epidemiol* 35(1), pp. 34-41.

Price, T. S. et al. 2005. SW-ARRAY: a dynamic programming solution for the identification of copy-number changes in genomic DNA using array comparative genome hybridization data. *Nucleic Acids Res* 33(11), pp. 3455-3464.

Priebe, L. et al. eds. 2009. *Genome-wide screening for common and rare copy number variants in a German sample of bipolar disorder*. XVII World Congress of Psychiatric Genetics. San Diego, CA.

Pritchard, J. K. 2001. Are rare variants responsible for susceptibility to complex diseases? *Am J Hum Genet* 69(1), pp. 124-137.

Pritchard, J. K. and Cox, N. J. 2002. The allelic architecture of human disease genes: common disease-common variant...or not? *Hum Mol Genet* 11(20), pp. 2417-2423.

Purcell, S. ed. 2009. *Analysis of rare structural variants in bipolar disorder and schizophrenia*. XVII World Congress of Psychiatric Genetics. San Diego, CA.

Purcell, S. et al. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 81(3), pp. 559-575.

Raychaudhuri, S. et al. 2010. Accurately assessing the risk of schizophrenia conferred by rare copy-number variation affecting genes with brain function. *PLoS Genet* 6(9).

Redon, R. et al. 2006. Global variation in copy number in the human genome. *Nature* 444(7118), pp. 444-454.

Reich, D. E. and Lander, E. S. 2001. On the allelic spectrum of human disease. *Trends Genet* 17(9), pp. 502-510.

Rice, J. et al. 1987. The familial transmission of bipolar illness. *Arch Gen Psychiatry* 44(5), pp. 441-447.

Rihmer, Z. and Angst, J. 2009. *Kaplan & Sadock's Comprehensive Textbook of Psychiatry*. 9th ed. Philadelphia Wolters Kluwer, Lippincott Williams&Wilkins.

Risch, N. and Botstein, D. 1996. A manic depressive history. *Nat Genet* 12(4), pp. 351-353.

Risch, N. and Merikangas, K. 1996. The future of genetic studies of complex human diseases. *Science* 273(5281), pp. 1516-1517.

Risch, N. J. 2000. Searching for genetic determinants in the new millennium. *Nature* 405(6788), pp. 847-856.

Rivera-Munoz, P. et al. 2007. DNA repair and the immune system: From V(D)J recombination to aging lymphocytes. *Eur J Immunol* 37 Suppl 1, pp. S71-82.

Rujescu, D. et al. 2009. Disruption of the neurexin 1 gene is associated with schizophrenia. *Hum Mol Genet* 18(5), pp. 988-996.

Rzhetsky, A. et al. 2007. Probing genetic overlap among complex human phenotypes. *Proc Natl Acad Sci U S A* 104(28), pp. 11694-11699.

Sandkuyl, L. A. and Ott, J. 1989. Affective disorders: evaluation of a three-allele model accounting for clinical heterogeneity. *Genet Epidemiol* 6(1), pp. 265-269.

Saus, E. S., V. Escaramic, G. Crespo, JM Valero, J. Gutierrez-Zotes, A. Martorell, L. Vilella, E. Menchon, JM Estivill, X. Urretavizcaya, M. Gratacos, M. 2009. CNV overlapping GSK3β (Glycogen Synthase Kinase-3β) Gene and its Association with Mood Disorders Phenotypes in Spanish Population.

Scherer, S. W. et al. 2007. Challenges and standards in integrating surveys of structural variation. *Nat Genet* 39(7 Suppl), pp. S7-15.

Schloesser, R. J. et al. 2008. Cellular plasticity cascades in the pathophysiology and treatment of bipolar disorder. *Neuropsychopharmacology* 33(1), pp. 110-133.

Schulze, T. G. et al. 2009. Two variants in Ankyrin 3 (ANK3) are independent genetic risk factors for bipolar disorder. *Mol Psychiatry* 14(5), pp. 487-491.

Scott, L. J. et al. 2007. A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. *Science* 316(5829), pp. 1341-1345.

Sebat, J. et al. 2007. Strong association of de novo copy number mutations with autism. *Science* 316(5823), pp. 445-449.

Sebat, J. et al. 2004. Large-scale copy number polymorphism in the human genome. *Science* 305(5683), pp. 525-528.

Sebat, J. et al. 2009. Rare structural variants in schizophrenia: one disorder, multiple mutations; one mutation, multiple disorders. *Trends Genet* 25(12), pp. 528-535.

Segurado, R. et al. 2003. Genome scan meta-analysis of schizophrenia and bipolar disorder, part III: Bipolar disorder. *Am J Hum Genet* 73(1), pp. 49-62.

Serretti, A. and Mandelli, L. 2008. The genetics of bipolar disorder: genome 'hot regions,' genes, new potential candidates and future directions. *Mol Psychiatry* 13(8), pp. 742-771.

Shaffer, L. G. et al. 2006. Targeted genomic microarray analysis for identification of chromosome abnormalities in 1500 consecutive clinical cases. *J Pediatr* 149(1), pp. 98-102.

Sharp, A. J. et al. 2006a. Structural variation of the human genome. *Annu Rev Genomics Hum Genet* 7, pp. 407-442.

Sharp, A. J. et al. 2006b. Discovery of previously unidentified genomic disorders from the duplication architecture of the human genome. *Nat Genet* 38(9), pp. 1038-1042.

Sharp, A. J. et al. 2005. Segmental duplications and copy-number variation in the human genome. *Am J Hum Genet* 77(1), pp. 78-88.

Sharp, A. J. et al. 2008. A recurrent 15q13.3 microdeletion syndrome associated with mental retardation and seizures. *Nat Genet* 40(3), pp. 322-328.

Shaw-Smith, C. et al. 2006. Microdeletion encompassing MAPT at chromosome 17q21.3 is associated with developmental delay and learning disability. *Nat Genet* 38(9), pp. 1032-1037.

Shevell, M. I. et al. 2008. Array comparative genomic hybridization in global developmental delay. *Am J Med Genet B Neuropsychiatr Genet* 147B(7), pp. 1101-1108.

Shianna, K. V. and Willard, H. F. 2006. Human genomics: in search of normality. *Nature* 444(7118), pp. 428-429.

Sklar, P. et al. 2008. Whole-genome association study of bipolar disorder. *Mol Psychiatry* 13(6), pp. 558-569.

Smith, D. J. and Lusis, A. J. 2002. The allelic structure of common disease. *Hum Mol Genet* 11(20), pp. 2455-2461.

Smith, E. N. et al. 2009. Genome-wide association study of bipolar disorder in European American and African American individuals. *Mol Psychiatry* 14(8), pp. 755-763.

Smoller, J. W. and Gardner-Schuster, E. 2007. Genetics of bipolar disorder. *Curr Psychiatry Rep* 9(6), pp. 504-511.

Spence, M. A. et al. 1995. Bipolar disorder: evidence for a major locus. *Am J Med Genet* 60(5), pp. 370-376.

Spitzer, R. L. et al. 1978. Research diagnostic criteria: rationale and reliability. *Arch Gen Psychiatry* 35(6), pp. 773-782.

St Clair, D. et al. 1990. Association within a family of a balanced autosomal translocation with major mental illness. *Lancet* 336(8706), pp. 13-16.

Stankiewicz, P. and Lupski, J. R. 2002. Genome architecture, rearrangements and genomic disorders. *Trends Genet* 18(2), pp. 74-82.

Stefansson, H. et al. 2008. Large recurrent microdeletions associated with schizophrenia. *Nature* 455(7210), pp. 232-236.

Strachan, D. P. et al. 2007. Lifecourse influences on health among British adults: effects of region of residence in childhood and adulthood. *Int J Epidemiol* 36(3), pp. 522-531.

Strachan, T. and Read, A. 2004. *Human Molecular Genetics 3*. London and New York: Garland Publishing.

Stranger, B. E. et al. 2007. Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science* 315(5813), pp. 848-853.

Sudhof, T. C. 2008. Neuroligins and neurexins link synaptic function to cognitive disease. *Nature* 455(7215), pp. 903-911.

Sutrala, S. R. et al. 2007. Gene copy number variation in schizophrenia. *Schizophr Res* 96(1-3), pp. 93-99.

Szatmari, P. et al. 2007. Mapping autism risk loci using genetic linkage and chromosomal rearrangements. *Nat Genet* 39(3), pp. 319-328.

Teare, D. M. and Barrett, J. H. 2005. Genetic linkage studies. *Lancet* 366(9490), pp. 1036-1044.

The International Hapmap Project 2003. The International HapMap Project. *Nature* 426(6968), pp. 789-796.

Torgersen, S. 1986. Genetic factors in moderately severe and mild affective disorders. *Arch Gen Psychiatry* 43(3), pp. 222-226.

Tsuchiya, K. J. et al. 2003. Risk factors in relation to an emergence of bipolar disorder: a systematic review. *Bipolar Disord* 5(4), pp. 231-242.

Turecki, G. et al. 2001. Mapping susceptibility genes for bipolar disorder: a pharmacogenetic approach based on excellent response to lithium. *Mol Psychiatry* 6(5), pp. 570-578.

Turner, D. J. et al. 2008. Germline rates of de novo meiotic deletions and duplications causing several genomic disorders. *Nat Genet* 40(1), pp. 90-95.

Tuzun, E. et al. 2005. Fine-scale structural variation of the human genome. *Nat Genet* 37(7), pp. 727-732.

Ullmann, R. et al. 2007. Array CGH identifies reciprocal 16p13.1 duplications and deletions that predispose to autism and/or mental retardation. *Hum Mutat* 28(7), pp. 674-682.

Valles, V. et al. 2000. Increased morbid risk for schizophrenia in families of in-patients with bipolar illness. *Schizophr Res* 42(2), pp. 83-90.

van Bon, B. W. et al. 2009. Further delineation of the 15q13 microdeletion and duplication syndromes: a clinical spectrum varying from non-pathogenic to a severe outcome. *J Med Genet* 46(8), pp. 511-523.

Venter, J. C. 2010. Multiple personal genomes await. *Nature* 464(7289), pp. 676-677.

Vissers, L. E. et al. 2003. Array-based comparative genomic hybridization for the genomewide detection of submicroscopic chromosomal abnormalities. *Am J Hum Genet* 73(6), pp. 1261-1270.

Walker, J. et al. 2002. Schizophrenia and bipolar disorder: similarities in pathogenic mechanisms but differences in neurodevelopment. *Int Clin Psychopharmacol* 17 Suppl 3, pp. S11-19.

Walsh, T. et al. 2008. Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia. *Science* 320(5875), pp. 539-543.

Warburton, D. 1991. De novo balanced chromosome rearrangements and extra marker chromosomes identified at prenatal diagnosis: clinical significance and distribution of breakpoints. *Am J Hum Genet* 49(5), pp. 995-1013.

Weiss, L. A. et al. 2008. Association between microdeletion and microduplication at 16p11.2 and autism. *N Engl J Med* 358(7), pp. 667-675.

Weissman, M. M. et al. 1996. Cross-national epidemiology of major depression and bipolar disorder. *Jama* 276(4), pp. 293-299.

Wellcome Trust Case Control Consortium 2007. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* 447(7145), pp. 661-678.

Wellcome Trust Case Control Consortium et al. 2010. Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. *Nature* 464(7289), pp. 713-720.

Weterings, E. and van Gent, D. C. 2004. The mechanism of non-homologous end-joining: a synopsis of synapsis. *DNA Repair (Amst)* 3(11), pp. 1425-1435.

Williams, H. J. et al. 2009. Schizophrenia genetics: new insights from new approaches. *Br Med Bull* 91, pp. 61-74.

Wilson, G. M. et al. 2006. DNA copy-number analysis in bipolar disorder and schizophrenia reveals aberrations in genes involved in glutamate signaling. *Hum Mol Genet* 15(5), pp. 743-749.

Wing, J. K. et al. 1990. SCAN. Schedules for Clinical Assessment in Neuropsychiatry. *Arch Gen Psychiatry* 47(6), pp. 589-593.

Wing, J. K., Cooper, J.E., Satorius, N. 1974. *The Measurement and Classification of Psychiatric illness*. Cambridge: Cambridge University Press.

Wollstein, A. et al. 2007. Efficacy assessment of SNP sets for genome-wide disease association studies. *Nucleic Acids Res* 35(17), p. e113.

Woodworth, A. et al. 2004. Neuronal-specific synthesis and glycosylation of tenascin-R. *J Biol Chem* 279(11), pp. 10413-10421.

World Health Organisation 1992. *The ICD-10, classification of mental and behavioural disorders, clinical descriptions and diagnostic guideline*. Geneva: World Health Organisation.

World Health Organisation 2008. *The Global Burden of Disease, 2004 Update*. World Health Organisation.

Xu, B. et al. 2008. Strong association of de novo copy number mutations with sporadic schizophrenia. *Nat Genet* 40(7), pp. 880-885.

Yang, Q. et al. 2005. How many genes underlie the occurrence of common complex diseases in the population? *Int J Epidemiol* 34(5), pp. 1129-1137.

Yang, S. et al. 2009. Genomic landscape of a three-generation pedigree segregating affective disorder. *PLoS One* 4(2), p. e4474.

Yobb, T. M. et al. 2005. Microduplication and triplication of 22q11.2: a highly variable syndrome. *Am J Hum Genet* 76(5), pp. 865-876.

Young, W. 2009. Review of lithium effects on brain and blood. *Cell Transplant* 18(9), pp. 951-975.

Zeggini, E. et al. 2008. Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. *Nat Genet* 40(5), pp. 638-645.

Zhang, D. et al. 2008. Singleton deletions throughout the genome increase risk of bipolar disorder. *Mol Psychiatry*.