Universitat de Vic
Escola Politècnica Superior

Department of Psychological Medicine and Neurology,
University of Cardiff
University Hospital of Wales

TREBALL FINAL DE GRAU

IDENTIFYING THE PARENT-OF-ORIGIN OF *DE NOVO* SNVs IN SCHIZOPHRENIA

Laura Escudero Monreal

Grau en Biotechnologia

Tutor UVIC: Professor Josep Bau Macià
Tutors UHW: Dr George Kirov, Dr Lyudmila Georgieva

Vic, juny de 2013

Universitat de Vic
Escola Politècnica Superior

Resum de Treball Final de Grau

Grau en Biotecnologia

IDENTIFYING THE PARENT-OF-ORIGIN OF *DE NOVO* SNVs IN SCHIZOPHRENIA

Laura Escudero Monreal

Tutors: Professor Josep Bau Macià, Dr George Kirov, Dr Lyudmila Georgieva

Summary

Several studies over the last few years have shown that newly arising (de novo) mutations contribute to the genetics of schizophrenia (SZ), autism (ASD) and other developmental disorders. The strongest evidence comes from studies of de novo Copy Number Variation (CNV), where the rate of new mutations is shown to be increased in cases when compared to controls [23, 24]. Research on de novo point mutations and small insertion-deletions (indels) has been more limited, but with the development of next-generation sequencing (NGS) technology, such studies are beginning to provide preliminary evidence that de novo single-nucleotide mutations (SNVs) might also increase risk of SZ and ASD [25, 26]

Advanced paternal age is a major source of new mutations in human beings [27] and could thus be associated with increased risk for developing SZ, ASD or other developmental disorders. Indeed, advanced paternal age is found to be a risk factor for developing SZ and ASD in the offspring [28, 29] and new mutations related to advanced paternal age have been implicated as a cause of sporadic cases in several autosomal dominant diseases, some neurodevelopmental diseases, including SZ and ASD, and social functioning. New single-base substitutions occur at higher rates at males compared to females and this difference increases with paternal age. This is due to the fact that sperm cells go through a much higher number of cell divisions (~840 by the age of 50), which increases the risk for DNA copy errors in the male germ line [30] . By contrast, the female eggs (oocytes) undergo only 24 cell divisions and all but the last occur during foetal life.

The **aim** of my project is to determine the parent-of-origin of de novo SNVs, using large samples of parent-offspring trios affected with schizophrenia (SZ). From whole exome sequencing of 618 Bulgarian proband-offspring trios affected, nearly 1000 de novo (SNVs or small indels) have been identified and from these, the parent-of-origin of at least 60% of the mutations (N=600) can be established. This project is contained in a main one that consists on the determination of the parental origin of different types of *de novo* mutations (SNVs, small indels and large CNVs).

Table of contents

CHAPTER THREE: MATERIALS, METHODS AND RESULTS

CHAPTER FOUR: OUTCOMES AND BENEFITS

CHAPTER ONE: INTRODUCTION

Neurological and mental disorders occur often, with approximately 450 million people suffering from them worldwide. According to the World Health Organization, mental and neurological diseases are responsible for approximately 1% of deaths and account for approximately 11% of the disease burden worldwide, a statistic that is expected to rise to 14.7% by the year 2020. Neurological and mental disease is a broad category covering a large number of disorders. The origins of these disorders, such as epilepsy, Schizophrenia, Alzheimer disease, Parkinson disease, cerebrovascular disease, depression, and brain cancer are difficult to determine.

Like most other common diseases, neurological disorders are hypothesized to be highly complex, with interactions among genes and risk factors playing a major role in the process. Many rare mendelian genetic disorders, such as cystic fibrosis, are influenced by the effects of a single gene. Statistical methodologies were developed to detect these single-disease genes and were very successful. The problem is that these Methods were not developed in the context of detecting interactions among genes associated with common diseases.

In recent years, it has become obvious that for common disorders, there may be more complex interactions among genes with and without strong independent main effects. These effects will be more difficult to detect using traditional methodologies.

1.1 SCHIZOPHRENIA

1.1.1 HISTORY

Schizophrenia is a common and severe psychiatric disorder with a worldwide prevalence of 1%. Descriptions of schizophrenia can be found in works of literature from earliest written history. Schizophrenia-like symptoms are described in individuals labelled as seers and prophets, as well as devils and witches. An understanding of schizophrenia as a human brain disease did not develop until the 19th century, but it wasn't until the middle of the 20th century that antipsychotic drug treatments became widely available [1].

It was first described in 1896 by Emil Kraepelin who used the term dementia praecox to distinguish this illness from manic depressive illness and dementia of the elderly [2]. Kraepelin further divided this patient group into three categories: hebephrenic, catatonic and paranoid. However, the term dementia praecox was thought by some to imply a pessimistic outcome in all cases and so, in 1908 Eugen Bleuler introduced the term Schizophrenia which literary means splitting the mind [3]. This reclassification implied a disruption of usually integrated thought processes and feelings, and took into account the frequent partial or full social recovery seen in people diagnosed with schizophrenia.

Now, at the beginning of the 21st century, there have been two generations of antipsychotic medications, several known risk genes, an evolving anatomy, but still no basic disease formulation.

1.1.2 SYMPTOMS

The symptoms of schizophrenia are conventionally divided into two main categories, positive and negative. The positive symptoms can include hallucinations, both auditory and visual, delusions, thought disorder and bizarre behaviour. The negative symptoms, which reflect a loss of normal function, constitute emotional flattening, apathy, poor motivation and drive, impaired abstract reasoning and poor personal hygiene. This results in a deterioration of professional function and social withdrawal. Depressive symptoms are also common in schizophrenia patients although these are secondary [1].

POSITIVE SYMPTOMS (psychotic symptoms)

Hallucinations

- **Auditory >** the most common type. More than 70 percent of patients (suggested by the IPSS) higher in industrialised societies, and up to 98percent in other studies. The voices have a negative content, patients will hear expletives, threats, demeaning personal comments, and accusations of vile thoughts or behaviours. Emotional experiences, most often sadness, often proceed hallucinations, but pacients also common experience the somatic symptoms of anxiety, and less frequently fear or anger, before the recurrence of hallucinations. Social stressors and physical illness and chronic pain, all can increase the frequency of hallucinations.
- **Visual>** are less common than auditory, but they are not rare. They have a prevalence from to 55 percent to one third of patients at some time of their illness. They have been suggested to predict a more severe illness. All kind of visual hallucinations have been described, from the most common ones, like images of animate objects, people, parts of people, religious images, fantastic creatures, inanimate objects like flashes of light, shadows, to illusions (distortions of objects seen in the environment) and less often, distortion of the world itself .

  The content of both auditory and visual hallucinations is often dependent on the culture of the person experiencing them.
- **Olfactory, gustatory and tactile>** They are present in a range of 15 to 25 percent of patients. These hallucinations can take on a broad variety of forms; feelings on being touched, burned and cut are the most common; also sensations of electric shocks, tearing or stabbing; and the physical sense that somehow other people or magical beings enter into and exit their bodies. Olfactory and gustatory hallucinations are reported for a minority of patients. As with the other ones, the experience tends to be unpleasant, with smells of rotting meat, garbage and feces, and the taste of blood or metal frequently described.

Delusions

They are defined as a false belief based on incorrect inference about external reality that is firmly sustained despite what almost everyone believes and despite what constitutes incontrovertible and obvious proof of evidence to the contrary and not one ordinarily accepted by other members of the person's culture. The delusions that are held by persons with schizophrenia are quite flexible, and can vary from fleeting concerns about personal meanings taken from the way a television news reader intonated a particular word, to a preoccupation with ideas about how giant corporations and governments are organised to persecute a particular person with schizophrenia. Instead of arising

from a new kind of perception, delusions often arise from social stressors, and threats to self-esteem seem particularly prone to provoke persecutory or grandiose delusions in response.

NEGATIVE SYMPTOMS

Negative symptoms represent a loss or diminution of normal functions and stand in contrast to positive symptoms where there are perception, cognitions and behaviours added to normal mental functions. It could be claimed that negative symptoms are the most important symptoms in schizophrenia because its severity predicts long-term disability better than the severity of psychotic or disorganisation symptoms.

Scale for the Assessment of Negative Symptoms (SANS)	
Affective Flattening or Blunting	Unchanging Facial Expression
	Decreased Spontaneous Movements
	Paucity of Expressive Gestures
	Poor Eye Contact
	Affective Nonresponsivity
	Lack of Vocal Inflections
	Global Rating of Affective Flattening
	Inappropriate Affect
Alogia	Poverty of Speech
	Poverty of Content of Speech
	Blocking
	Increased Latency of Response
	Global Rating of Alogia
Avolition - Apathy	Grooming and Hygiene
	Impressistence at Work or School
	Physical Anergia
	Global Rating of Avolition - Apathy
Anhedonia - Asociality	Recreational Interests and Activities
	Sexual Interest and Activity
	Ability to Feel Intimacy and Closeness
	Relationships with Friends and Peers
	Global Rating of Anhedonia-Asociality
Attention	Social Inattentiveness
	Inattentiveness During Mental Status Testing
	Global Rating of Attention
	Scale for the Assessment of Positive Symptoms (SAPS)

Scale for the Assessment of Positive Symptoms (SAPS)	
Hallucinations	Auditory Hallucinations
	Voices Commenting
	Voices Conversing
	Somatic or Tactile Hallucinations
	Olfactory Hallucinations
	Visual Hallucinations
	Global Rating of Severity ofm Hallucinations
Delusions	Persecutory Delusions
	Delusions of Jealousy
	Delusions of Sin or Guilt
	Grandiose Delusions
	Religious Delusions
	Somatic Delusions
	Ideas and Delusions of Reference
	Delusions of Being Controlled
	Delusions of Mind Reading
	Thought Broadcasting
	Thought Insertion
	Thought Withdrawal
	Global Rating of Severity of Delusions
Bizarre Behaviour	Clothing and Appearance
	Social and Sexual Behavior
	Aggressive and Agitated Behavior
	Repetitive or Stereotyped Behavior
	Global Rating of Severity of Bizarre Behavior
Positive Formal Thought Disorder	Derailment (Loose Associations)
	Tangentiality
	Incoherence
	Illogicality
	Circumstantiality
	Pressure of Speech
	Distractible Speech
	Clanging
	Global Rating of Positive Formal Thought Disorder

DISORGANIZATION

The disorganization syndrome includes the formal thought disorder, bizarre and catatonic behaviours, inappropriate affect and attention impairments. It appears to be the most heritable of

Motor symptoms

Disturbance of motor activity seems to be most related to the disorganisation symptoms of schizophrenia, although some studies suggest that these symptoms are an independent dimension of psychopathology. Motor behaviours can include subtle repetitive hand movements or broad, complex, and purposeless movements that include limbs and trunk. Symptoms of catatonia are also included, although is important to note that catatonia may be as common in cases with brain injury and in psychotic mood disorders as it is in schizophrenia. It is suggested to be motor behaviour generated with a marked decrease in reactivity to the environment, and in its appearance it does suggest that central motor programmes are engaged without direction from frontal areas that direct higher level planning.

Thought disorder

It refers to the disorganisation of the form of thought, and not content. Speech is frequently stilted or vague, and sentences may be incomplete, it may be tangential, so that the associative chain moves obliquely off topic. With the progression of time, speech can further deteriorate, and inappropriate content can intrude.

DEPRESSION AND ANXIETY

Depression
Most people with schizophrenia will experience a significant depression and anxiety during the course of their illness, and it has been claimed that depression is an integral part of the disorder. Clinicians and family may not notice depression or anxiety in their patients and relatives with schizophrenia, or may be too distracted by patient's positive and negative symptoms.
Anxiety
Although is very prevalent, there is little known about anxiety in schizophrenia. Experience suggests that it can precipitate violence and suicidal ideation and that can lead to increases in psychosis and disorganization and depression.

SUICIDE

It accounts for a good part of the excess of mortality that is usually found in schizophrenia. From 20 to 40 percent will make a suicide attempt sometime during their illness. Recent analysis suggests that 5 percent of the people with schizophrenia will commit suicide, with increased risk early in the course of illness [1].

VIOLENCE

For those with schizophrenia and their families and advocates, the real problem is that they are more likely to be victims of violence than the perpetrators. Although this is true, it also remains true that the evidence available indicates that a small percentage of people with schizophrenia will commit a disproportionate amount of violence and even murder.

INSIGHT

Lack of insight was the most common symptom of schizophrenia. It requires some ability to reflect on oneself, in parallel with a certainty about one's existence and confidence in the limits of one's being. Poor insight is associated with decreased compliance, worse overall function, increased levels of psychopathology, recurrent illness, and poor outcomes.

1.1.3 DIAGNOSTIC

There are no available biological markers or pathognomonic symptoms that can be used to diagnose an individual with schizophrenia and so a clinical diagnosis relies on behavioural observations and self reported abnormal experiences. Thus, there is a need for standardised procedures that a diagnostician can follow in order to make an accurate diagnosis.

There have been changes in the diagnostic criteria of schizophrenia in order to uniform it, which will contribute to better communication among clinicians and researchers from different countries.
Actually, the diagnostic criteria is based on the fifth version of the Diagnostic and Statistical Manual of Mental disorders (DSM-5) but before he DSM system was introduced, in 1980 the diagnosis of schizophrenia was highly subjective and early cross-national studies highlighted the international differences in the breadth and style of diagnosis [2] These differences led to many patients with manic-depressive illness being diagnosed with schizophrenia.

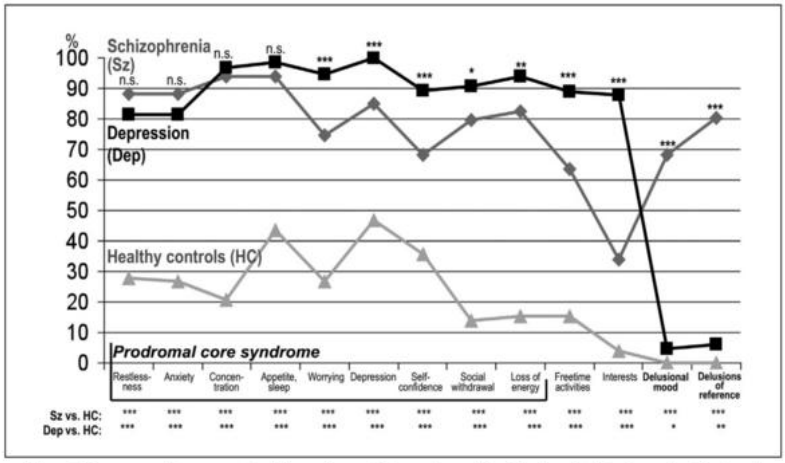


Figure 1. Frequency of symptoms (period prevalences %) in patients with schizophrenia (Sz), depression (Dep) and healthy controls (HC) Symptoms with ranks 1 to 10 and prevalences > 5% in any of the three groups.
McNemar test: n.s. = not significant; * $p < 0.05$; ** $p < 0.01$, *** $p < 0.001$.

Until now, the DSM-IV-TR specified the criteria for diagnosis and therefore introduced a standard with which to train and guide diagnosticians worldwide. [4] The components of an individual's illness can vary between cases. To allow for this phenotypic heterogeneity the DSM-IV-TR system categorised the symptoms into groups: Paranoid, Disorganised, Catatonic, Undifferentiated, Simple and Residual.
This reflects the complex nature of the schizophrenic phenotype in that schizophrenia seems to affect a wide range of brain systems and produces diverse signs and symptoms but is clearly recognisable as a syndrome.

The DSM is periodically reviewed and revised since it was first published in 1952. The previous version of *DSM* was completed nearly two decades ago; since that time, there has been a wealth of new research and knowledge about mental disorders. This May 2013 appeared the fifth edition of the DMS, with the following changes from DMS-IV. [16]

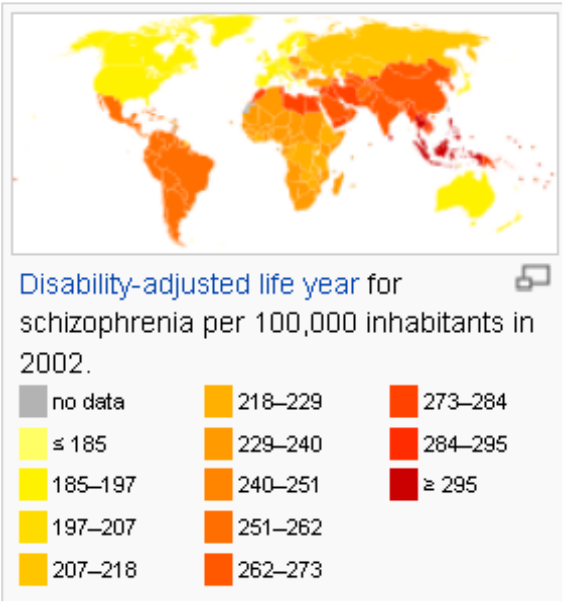
- All subtypes of Schizophrenia were deleted (paranoid, disorganized, catatonic, undifferentiated, and residual).
- A major mood episode is required for schizoaffective disorder (for a majority of the disorder's duration after criterion A is met).
- Criteria for delusional disorder changed, and, in DSM-5, delusional disorder is no longer separate from shared delusional disorder.
- In DSM-5, catatonia in all contexts requires 3 of a total of 12 symptoms. Catatonia may be a specifier for depressive, bipolar, and psychotic disorders; part of another medical condition; or another specified diagnosis.

1.2 CLINICAL EPIDEMIOLOGY OF SCHIZOPHRENIA

1.2.1 INCIDENCE AND PREVALENCE

Schizophrenia is a severe form of mental illness affecting about 7 per thousand of the adult population, mostly in the age group 15-35 years. Though the incidence is low (3-10,000), the prevalence is high due to chronicity. Schizophrenia Ranks among the top 10 causes of disability in developed countries worldwide. [5]

The global distribution of schizophrenia according to WHO (World Health Organisation) 2002



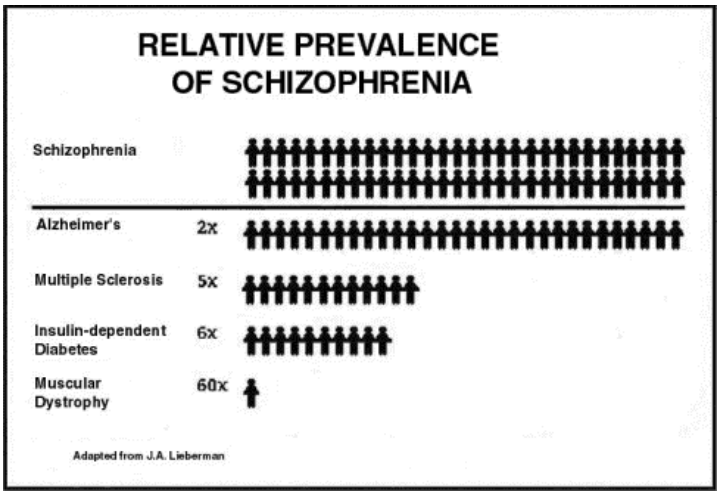
To date, no population or culture has been identified in which schizophrenia does not occur. The rates of its occurrence are broadly comparable, but this does not imply that the incidence of the disorder is uniform across all populations.

The **incidence** rate is the estimated annual number of first onset cases in a defined population per 1000 persons at risk, and the **prevalence** of the disease is the estimated number of cases per 1000 persons at risk in a population at a given time or over a defined period.

Objective biomarkers of the disease process are still lacking, so onset is usually defined as the point in time when clinical manifestation become recognizable and can be diagnosed according to specified criteria. A systematic review of incidence data from some 160 studies in 33 countries, published between 1965 and 2001 estimated the median value of 0.15 and mean value of 0.24 per 1000, with a fivefold range of the rates, and a tendency for recent studies to report lower rates.
Since schizophrenia cases in prolonged remission are likely to be missed in point prevalence surveys, it will always be useful to estimate the lifetime prevalence by supplementing the assessment of the present mental state data with data about past episodes of the disorder. A systematic review of 188 studies in 46 countries, published between 1965 and 2002 estimated the median value for point prevalence at 4.6 per 1000 persons, and for lifetime prevalence at 7.2 per 1000.[1]

[Note: The term 'prevalence' of Schizophrenia usually refers to the estimated population of people who are living with Schizophrenia at any given time. The term 'incidence' of Schizophrenia refers to the annual diagnosis rate, or the number of new cases of Schizophrenia diagnosed each year.]

The prevalence of schizophrenia compared to other well-known diseases.



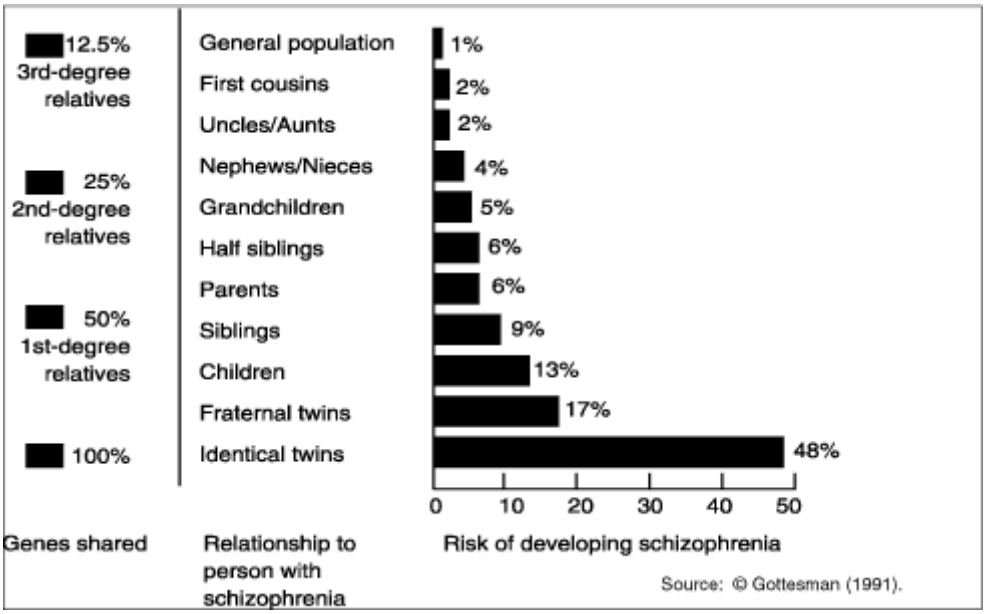
- Schizophrenia affects about 24 million people worldwide.
- Schizophrenia is a treatable disorder, treatment being more effective in its initial stages.
- More than 50% of persons with schizophrenia are not receiving appropriate care.
- 90% of people with untreated schizophrenia are in developing countries.
- Care of persons with schizophrenia can be provided at community level, with active family and community involvement.

There are effective interventions (pharmacological and psychosocial) available and the cost of treatment of a person suffering from chronic schizophrenia is about US$2 per month; the earlier the treatment is initiated, the more effective it will be. However, the majority of the persons with chronic schizophrenia do not receive treatment, which contributes to the chronicity. [6]

1.2.2 FACTORS INFLUENCING THE VARIATIONS IN OUTCOME

The outcome of schizophrenia is a result of a **genetic predisposition** combined with an **environmental exposures** and / or stresses during pregnancy or childhood that contribute to, or trigger, the disorder. Already researchers have identified several of the key genes - that when damaged - seem to create a predisposition, or increased risk, for schizophrenia. The genes, in combination with suspected environmental factors are believed to be the factors that result in schizophrenia. These genes that seem to cause increased risk of schizophrenia include the DISC1, Dysbindin, Neuregulin and G72 genes, but it has been estimated that up a dozen or more genes could be involved in schizophrenia risk.

One of the most positive areas of schizophrenia research today is in the area of identification of early risk factors for development of schizophrenia, and prevention of schizophrenia in those people who are predisposed to the disease.



The risk of getting schizophrenia. After a person has been diagnosed with schizophrenia in a family, the chance for a sibling to also be diagnosed with schizophrenia is 7 to 9 percent. If a parent has schizophrenia, the chance for a child to have the disorder is 10 to 15 percent. Risks increase with multiple affected family members. [7]

1.2.3 THE BURDEN OF COMORBIDITY

There is a significant comorbidity in schizophrenia, comprised mainly of common diseases that affect schizophrenic patients more frequently than attributable to chance, as well as certain rare conditions or abnormalities which tend to co-occur with the disorder. Persons with schizophrenia, especially those who are homeless or injection drug users, are at increased risk for potentially life-threatening communicable diseases, such as HIV/AIDS, hepatitis C, and tuberculosis. Among the chronic no communicable diseases, they have higher rates than expected rates of epilepsy, diabetes, arteriosclerosis, and ischemic heart disease. Obesity and the concomitant metabolic syndrome involving insulin resistance are becoming increasingly common problems in schizophrenic patients. Moreover, certain genetic or idiopathic disorders have been reported to co-occur with schizophrenia.

Substance abuse is at the present by far the most common associated health problem among patients with schizophrenia, and may involve any drug abuse or a polydrug combination. The addictive use of cannabis, stimulants, and nicotine is disproportionately high among schizophrenic patients and may be related to the underlying neurobiology of disorder. Early use of cannabis increased the risk of psychosis in a dose-related manner, especially in persons at high genetic risk of schizophrenia. [1]

1.2.4 RISK FACTORS OF THE ONSET OF SCHIZOPHRENIA

There are several risk factors operating during early development involved in the onset of schizophrenia, among them, the risk factor related with the paternal age is relevant for this project.

Paternal age

Large population based cohort studies from Israel, the United States, Denmark and Sweden, provided strong evidence consistently linking advanced paternal age to the risk of schizophrenia in the offspring. There is a higher risk in schizophrenia (around three to four times) in the offspring of fathers who are older than 50 at the time of conception, compared to the offspring of fathers in their early 20s.

In evidence on gene environment interactions, there is a stronger association between paternal age and schizophrenia in people without a family history. **This observed pattern of effect lends support for two genetic theories. The first and most likely is that advancing paternal age results in accumulation of de novo mutations in the germ cells of older fathers,** or second, advancing paternal age interferes with the DNA-methylation process of gene expression.

Season of birth

Winter birth, in people who later develop schizophrenia is a robust epidemiological finding, at least in the northern hemisphere. It is likely to be a proxy indicator for some seasonally fluctuating environmental factor. The most popular hypothesis relate to seasonal variation in exposure to intrauterine viral infections around the time of birth, or variation in light, temperature/weather, or external toxins.

Pregnancy and birth complications

The literature suggests that pregnancy and birth complications can have a small effect on the risk of later development of schizophrenia, as it has been demonstrated in a large number of published studies. Investigators found three main categories of obstetric complications to have significant estimates:

- Abnormal fetal growth and development: low birth weight, congenital malformations, and small head circumference.
- Complications of pregnancy: bleeding, pre-eclampsia, diabetes, and rhesus incompatibility.
- Complications of delivery: Asphyxia, uterine-atony, and emergency caesarean section. Taken together, they seem to implicate an increased risk of hypoxia.

Other putative prenatal risk factors

A study from New York City found a 10 to 20 percent increased risk for schizophrenia in people who had serologically confirmed prenatal rubella exposure, while further studies have implicated prenatal exposure to toxoplasmosis, poliovirus, and other common respiratory infections. It has been suggested that the effect may be in part due to cytokines and chemokines, which mediated host response to infection. There is some evidence that maternal-fetal genotype incompatibility effects increases the risk of developing schizophrenia.

There might be higher rates of schizophrenia in the offspring of mothers who experienced significant levels of stress during pregnancy, such as the death of a spouse or living through a natural disaster or military invasions. Nutritional deficiency in pregnancy may also increase the risk. [1]

1.3 GENETIC EPIDEMIOLOGY OF SCHIZOPHRENIA

1.3.1 FAMILY STUDIES

In order to determine if susceptibility to schizophrenia involves a genetic component, it is essential to first demonstrate whether the illness is clustered within affected families. This can be quantified by estimating the risk of developing the disease in different classes of relative and determining whether this is greater than the population average.

Overall both the old and new family studies established the high familial loading of schizophrenia, with siblings consistently showing about eight- to tenfold increased risk of being ill, compared with the rest of the general population. This is much greater than the risk reported for any environmental factor studied.
If a condition is found to be more common in relatives of probands, this could be due to shared genes or to shared environmental factors. Therefore, in order to confirm that a disorder is genetic, other types of studies need to be conducted, such as twin and/or adoption studies, which control for the environmental factors.

1.3.2 TWIN STUDIES

Following the demonstration of familial clustering, twin studies have been used to determine whether this is due to genetic factors or is the result of shared environmental factors. The rationale behind twin studies is that if a disease is caused predominantly by genetic factors, then it would be expected that concordance rates between MZ twins (who share 100% of their DNA) would be greater than DZ twins (who on average share 50% of their DNA). Conversely, if the disease is caused by environmental factors, then the rate of concordance between MZ and DZ twins would be the same assuming that both types of twin share environmental influences to the same extent .
The concordance rate refers to the proportion of co-twins who are also affected, or the proportion of twin pairs where both twins are affected.

After comparing different studies, it is clear that schizophrenia occurs more frequently in the relatives of patients, and that MZ concordance is greater than DZ concordance. These findings could still be explained to some extent by shared environmental factors. One way to control for them is provided by adoption studies.

1.3.3 ADOPTION STUDIES

Although family and twin studies have shown that schizophrenia aggregates within families there is always the possibility that this clustering may be the result of a shared familial environment. In order to clarify this problem adoption studies have been used. The general principle behind adoption studies is that if there is a genetic component to the disorder studied, the similarity between

adopted children and their biological parents should be higher than the similarity between adopted children and their adoptive parents.
In the different studies performed, a much higher rate of illness was found in children whose biological parents had schizophrenia.

All adoption studies come to remarkably conclusions: children of schizophrenic parents have a risk of developing schizophrenia, schizoaffective disorder, or other narrow spectrum schizophrenic disorders of at least 10 percent, even when they are adopted away very soon after birth. The risk is very similar to the risk among children with schizophrenic parents reported in family studies, **suggesting that most of the transmission is genetic, rather than an effect of upbringing**. This high risk was not due to the stress of adoption either, as several studies had control samples of adoptees of normal parents, where the risk of narrow spectrum schizophrenia was 0 to 2 percent (similar to the population risk). [1]
Adoption studies have therefore convincingly supported the results of family and twins studies in demonstrating the significant role of genetic factors in schizophrenia.

1.3.4 WHAT IS INHERITED?

The evidence from family, twin and adoption studies clearly implicates an important genetic contribution to schizophrenia. However the facts that no study has found MZ twins are 100% concordant **suggest environmental factors play a role and that the mode of inheritance, like that of other common disorders, is complex and non-Mendelian**.

Genetic epidemiologists suggest that the genes that predispose to schizophrenia do not respect the diagnostic categories used by psychiatrists. There is good evidence for a spectrum of milder phenotypes associated with risk to schizophrenia, the so called **extended phenotype**, and also that there is an overlapping risks with other disorders, in particular bipolar disorder.
The fact that the risk of schizophrenia among relatives is much smaller than what would be expected from a genetic disorder has troubled many schizophrenia researchers. On the other hand, it has been observed that although some do not meet the diagnostic criteria for the disorder, some of their relatives exhibit a certain degree of psychopathology, which could be explained as a variable penetrance of the disease genes, called at the beginning of the 20th century as “latent schizophrenia”.
It appears that what is transmitted in families is a liability to develop not only schizophrenia, but also schizotypal and paranoid personality disorders and other psychotic illnesses.

1.4 MAPING COMPLEX TRAITS

1.4.1 LINKAGE STUDIES

There are two main strategies to find specific genes that cause schizophrenia: Linkage and association studies. A third strategy emerged more recently: searching for copy number variation (CNV).
Linkage studies make no assumption about specific genes involved in the ethiology of the disorder, while until recently association studies had focussed on candidate genes. Recently, linkage studies have been followed by association studies of the genes contained within the chromosomal regions implicated and have produced some of the more replicated findings. The availability of array-based single-nucleotide polymorphism (SNP) genotyping, has allowed association studies to start targeting the whole genome.

Linkage studies in schizophrenia have been reviewed many times over the years and subjected to several meta-analyses. There have been at least 27 whole genome studies that analysed between 1 to 294 pedigrees containing between 32 and 669 individuals affected with a narrow definition of schizophrenia. These studies have been based on very different sample sizes, with correspondingly different statistical power, and they used different methods and different quality controls. Several meta-analyses have reached somewhat different conclusions. Large numbers of genes were implicated and a few genes were consistently identified in more than a small subset of studies.

The first conclusions that can be drawn is that no single gene for schizophrenia exists, confirming that **is not a single gene disorder**. Several genomic loci have received support from several studies. There are two explanations why the studies so far have produced such different linkage findings.

- Different genes operate in different populations.
- Schizophrenia is caused by the effect of many genes of small effect, so that the studies had no power to detect the loci. It has been estimated that 4900 pedigrees would be required to have 80 percent power to detect a locus accounting for 5 percent of variance in liability to schizophrenia at α=0.001. [1]

1.4.2 MODE OF INHERITANCE

The inability of linkage studies to unambiguously identify linkage signals and the absence of a clear mode of transmission in the vast majority of families affected with schizophrenia indicates that this is not a simple Mendelian disorder, but a disorder of **complex inheritance**.
There are two main hypotheses about the genetic background of common diseases (including schizophrenia):

- The common disease/common variant (CDCV) hypothesis proposes that common diseases are caused by common variants. This model suggests a joint action of several common

genetic variants, each of which has a small effect on illness susceptibility, together with the environmental factors.

- The rare variants hypothesis postulates that multiple rare variants in different gens, which have low population frequencies, operate in different individuals

However, it is most likely that both mechanisms operate in common diseases, including schizophrenia, as **both high and low-frequency alleles have been found to contribute to several common diseases.** There are rare chromosomal aberrations that appear to have a high penetrance and cause a small number of cases of schizophrenia.

1.4.3 ASSOCIATION STUDIES

Several teams followed their positive linkage findings with fine mapping association studies of genes in linked chromosomal regions and identified the most plausible candidates to date.

- **Dystrobrevin-Binding Protein 1 (DTNBP1)**

Dysbindin binds both α and β-dystrobrevin, which are components of the dystrophin glycoprotein complex, located in both the sarcolemma of muscle and the brain. Konrad Talbot end colleagues found that the presynaptic dystrobrevin-independent fraction of dysbindin is reduced in schizophrenic brain within certain glutamatergic neurones of the hippocampus, and this is associated with increased expression of vesicular glutamate transporter type 1. Moreover, a reduction in glutamate release has been demonstrated in cultured neurons with reduced DTNBP1 expression. Variations in DTNBP1 might confer risk by altering presynaptic glutamate function.

- **Neuregulin 1 (NRG1)**

It encodes multiple proteins with a diverse range of functions in the brain, including cell-cell signalling, ErbB receptor interactions, axon guidance, synaptogenesis, glial differentiation, myelination, and neurotransmission. Any of these could potentially influence the susceptibility of schizophrenia.

Other candidate genes are: DAOA, G72/G30, COMT, RGS4, CAPON, PRODH and AKT1, but their status remains uncertain. [1]

1.5 GENETIC ARCHITECTURE OF SCHIZOPHRENIA

Although a genetic component of schizophrenia has been acknowledged for a long time, the underlying architecture of the genetic risk remains a contentious issue. Early linkage and candidate association studies led to largely inconclusive results. More recently, the availability of powerful technologies, samples of sufficient sizes, and genome-wide panels of genetic markers facilitated systematic and agnostic scans throughout the genome for either common or rare disease risk variants of small or large effect size, respectively. Although the former had limited success, the role of rare genetic events, such as copy-number variants (CNVs) or rare point mutations, has become increasingly important in gene discovery for schizophrenia. Importantly, recent research building upon earlier findings of de novo recurrent CNVs at the 22q11.2 locus, has highlighted a de novo mutational paradigm as a major component of the genetic architecture of schizophrenia. Recent progress is bringing us closer to earlier intervention and new therapeutic targets.[8]

1.5.1 GENOME WIDE ASSOCIATION STUDIES

A genome-wide association study is an approach that involves rapidly scanning markers across the complete sets of DNA, or genomes, of many people to find genetic variations associated with a particular disease. Once new genetic associations are identified, researchers can use the information to develop better strategies to detect, treat and prevent the disease. Such studies are particularly useful in finding genetic variations that contribute to common, complex diseases, such as asthma, cancer, diabetes, heart disease and mental illnesses.

With the completion of the Human Genome Project in 2003 and the International Hap Map Project in 2005, researchers now have a set of research tools that make it possible to find the genetic contributions to common diseases. The tools include computerized databases that contain the reference human genome sequence, a map of human genetic variation and a set of new technologies that can quickly and accurately analyze whole-genome samples for genetic variations that contribute to the onset of a disease.[21]

Technological advances in SNP genotyping made such studies feasible. Two companies, Affymetrix ad Illumina, designed microarrays that are capable of simultaneously genotyping hundreds of thousands of SNPs on a single array.

To carry out a genome-wide association study, researchers use two groups of participants: people with the disease being studied and similar people without the disease. Researchers obtain DNA from each participant which is then purified, placed on tiny chips and scanned on automated laboratory machines. The machines quickly survey each participant's genome for strategically selected markers of genetic variation, which are called single nucleotide polymorphisms, or SNPs.

If certain genetic variations are found to be significantly more frequent in people with the disease compared to people without disease, the variations are said to be "associated" with the disease. The associated genetic variations can serve as powerful pointers to the region of the human genome where the disease-causing problem resides. However, the associated variants themselves may not directly cause the disease. They may just be "tagging along" with the actual causal variants. For this reason, researchers often need to take additional steps, such as sequencing DNA base pairs in that particular region of the genome, to identify the exact genetic change involved in the disease. [21]

Availability of **next-generation whole-genome or whole-exome sequencing now permits the study of *de novo* mutations** (point substitutions or single nucleotide variants (**SNVs**) and small insertions or deletions (indels)) in a systematic genome-wide manner [17,18]
Pilot studies focusing on specific synaptic genes identified a small number of putative *de novo* mutations in individuals with schizophrenia[19]. However, the full contribution of rare *de novo* SNVs and indels to schizophrenia remains unknown.

1.5.2 CNV : COPY NUMBER VARIATION

In order to identify the genetic variants that increase susceptibility to schizophrenia, it appears clear that SNV will not account for all the genetic susceptibility, and that there is a need to explore other sources of genetic variability. Therefore, the structural variation in human chromosomes is the most promising, as large chromosomal aberrations can cause several neurodevelopmental disorders, disrupting gene function, inactivating, or even duplicating genes. Technological advances have recently enabled the use of high resolution techniques, which have been instrumental in identifying up to 20 percent of sufferers with mental retardation and autism, where most of the chromosomal changes that appear causative have arisen de novo, such as spontaneous mutations in the parental germ cells.
Schizophrenia has certain features that suggest a partial overlapping etiology with mental retardation and autism, including a tendency to show delayed development and lower intelligence quotient, language and communication problems, and a higher rate of minor physical anomalies. There have been many numerous reports associating schizophrenia and large chromosomal abnormalities, two of them providing convincing evidence for the location of a susceptibility gene.

- **Chromosome 22q11.2 Deletion Syndrome**

The first and best replicated finding of a chromosomal aberration that increases the risk of schizophrenia is a homozygous deletion of chromosome 22q11.2, also known as DiGeorge/velocardiofacial syndrome (VCFS). The prevalence of this syndrome has been estimated at between 1 in 3.900 to 1 in 9.700 children. [1] Carriers of the deletion present with enormous phenotypic heterogeneity, with the most common features being cardiac anomalies, hypocalcemia, cleft lip/palate, renal abnormalities, skeletal abnormalities, developmental delay, especially speech delay, and behavioural and psychiatric disorders.
The deletion is caused by the presence of four blocks of low copy number repeats (LCRs) in this region, named LCR A-D, which are larger, more complex and have a higher homology than any other

LCRs in the human genome. The deletion affects typically a 3-Mb region, although a tenth of cases are caused by a smaller deletion, and at least 35 genes are present within the commonly deleted region.
Most commonly is presented as de novo mutation, but between 8 and 28 percent of cases is inherited from parents who have a mild phenotype.
Several studies have shown that adults with 22q11.2 deletions have a high risk of schizophrenia, as the deletion accounts for up to 3 to 6 percent of the cases with schizophrenia.
Reports on linkage to 22q11 suggest that variants in genes mapping to this region might contribute to cases of schizophrenia that do not have 22q11 deletions. One good candidate gene in the region appears to be catecholamine-O-methyl transferase (COMT), and evidence in favour of a role in schizophrenia susceptibility has been reported for several other genes including TBX1, GNB1L, PRODH and ZDHHC, although there is no compelling evidence yet. It remains possible that the high risk for schizophrenia resulting from these deletions, reflect haploinsufficiency of more than a single gene.

- **DISC1**

The other major finding on a chromosomal abnormality comes from a balanced chromosomal translocation (1,11) (q42, q14.3) that showed very strong evidence for linkage to a fairly broad phenotype consisting of schizophrenia, bipolar disorder and recurrent depression. The translocation was found to disrupt two genes on chromosome 1: DISC1 and DISC2. DISC2 contains no open reading frame and may regulate DISC1 expression by antisense RNA. It has been suggested that the disruption of DISC1 might contribute to schizophrenia by affecting neuronal functions dependent on intact cytoskeletal regulation, such as neuronal migration, neurite architecture, and intracellular transport. [1]
At the moment, there are suggestions that DISC1 variants might confer susceptibility to a range of phenotypes, including schizoaffective disorder, bipolar disorder, and recurrent major depressive disorder as well as schizophrenia.

Genome structural variation has been known for a long time, but its extent was not appreciated until recently. The earlier technologies available allowed us to improve in these studies. Is known that every person in the general population carries a surprisingly large number of such CNVs, many of which involved single-copy genes. Despite this, very few disorders have been associated with CNVs, but the list is expected to rise. Several studies have been published, and analysis of data reported indicates that some CNVs shown to contribute towards the pathogenesis of autism are also involved in schizophrenia. These findings demonstrate the important role of structural chromosomal variation in the pathogenesis of the disorder, and bring hope that other smaller or rarer CNVs will be identified soon, when more samples are investigated, and even higher-resolution platforms become available.

1.6 DE NOVO MUTATIONS

De novo mutations are mutations observed in a child but not in his or her parents, and they are assumed to have occurred in one of the parental germ lines.

All genetic variation arises via new mutations; therefore, determining the rate and biases for different classes of mutation is essential for understanding the genetics of human disease and evolution. Recent studies have shown that **76% of new mutations originate in the paternal lineage and provide unequivocal evidence for an increase in mutation with paternal age**. Although most analyses have focused on single nucleotide variants (SNVs), studies have begun to provide insight into the mutation rate for other classes of variation, including copy number variants (CNVs), microsatellites, and mobile element insertions (MEIs).[9]

The replication of the genome before cell division is a remarkably precise process. Nevertheless, there are some errors during DNA replication that lead to new mutations. If these errors occur in the germ cell lineage (i.e., the sperm and egg), then these mutations can be transmitted to offspring. Some of these new genetic variants will be deleterious to the organism, and a select few will be advantageous and serve as substrates for selection. There-fore, knowledge about the rate at which new mutations appear and the properties of new mutations is critical in the study of human genetics from evolution to disease.

Over the past few years, it has become feasible to generate large amounts of sequence data (including the genomes of parents and their offspring), and it is now possible to calculate empirically a genome-wide mutation rate. In addition, much interest has focused on under-standing the role of de novo mutations in human disease.

Recent genome-wide studies (on all individuals from a nuclear family) of the SNV mutation rate in humans have started to converge. These studies based on whole-genome sequencing and direct estimates of de novo mutations give an average SNV mutation rate of $1.16*10^{-8}$ mutations per base pair per generation [95% confidence interval (CI) of the mean: 1.11–1.22] in 96 total families. [10-14]

1.6.1 DE NOVO MUTATIONS IN SCHIZOPHRENIA

Schizophrenia has a strong genetic component. However, despite its high heritability, a large fraction of individuals with schizophrenia do not have a family history of the disease[22]. Although largely ignored in earlier efforts to model disease risk, *de novo* germ line mutations may account for a substantial fraction of sporadic schizophrenia cases.

Several studies in the last 5 years have shown that newly arising (de novo) mutations contribute to the genetics of schizophrenia (SZ). This will replenish genetic variants removed by natural selection and could, in part, explain why SZ prevalence has remained stable in the general population despite low fecundity. The strongest evidence to date for the association between SZ and de novo mutation comes from studies of de novo copy number variation (CNV), where the rate of de novo CNV mutation is shown to be increased in cases when compared with controls, and genes disrupted by these mutations are enriched for those encoding proteins involved in synaptic function and development. [15] Recent studies involving next-generation sequencing technology have provided preliminary evidence that de novo single-nucleotide mutations might also increase risk of SZ.

CHAPTER TWO: IDENTIFYING THE PARENT OF ORIGIN OF *DE NOVO* SNVS IN SCHIZOPHRENIA

2.1 BACKGROUND

Several studies over the last few years have shown that newly arising (de novo) mutations contribute to the genetics of schizophrenia (SZ), autism (ASD) and other developmental disorders. The strongest evidence comes from studies of de novo Copy Number Variation (CNV), where the rate of new mutations is shown to be increased in cases when compared to controls [23, 24]
Research on de novo point mutations and small insertion-deletions (indels) has been more limited, but with the development of next-generation sequencing (NGS) technology, such studies are beginning to provide preliminary evidence that de novo single-nucleotide mutations (SNVs) might also increase risk of SZ and ASD [25, 26]
While the extent of involvement of de novo mutations is currently unknown, the evidence for their involvement offers a new approach to detect potentially pathogenic variants and may help to explain partially the causes and inheritance pattern of SZ and also how SZ prevalence remains stable in the general population despite low fecundity.

Advanced paternal age is a major source of new mutations in human beings [27] and could thus be associated with increased risk for developing SZ, ASD or other developmental disorders. Indeed, advanced paternal age is found to be a risk factor for developing SZ and ASD in the offspring [28, 29] and new mutations related to advanced paternal age have been implicated as a cause of sporadic cases in several autosomal dominant diseases, some neurodevelopmental diseases, including SZ and ASD, and social functioning.

New single-base substitutions occur at higher rates at males compared to females and this difference increases with paternal age. This is due to the fact that sperm cells go through a much higher number of cell divisions (~840 by the age of 50), which increases the risk for DNA copy errors in the male germ line [30] . By contrast, the female eggs (oocytes) undergo only 24 cell divisions and all but the last occur during foetal life.

Much less is known about the parent-of-origin and the effect of increased parental age on the genesis of CNVs. A recent study on mental retardation found de novo CNVs to be mostly of paternal origin and to be associated with increased paternal age[31]. This was particularly evident for non-recurrent CNVs that arise by replication based mechanisms such as non-homologous end joining (NHEJ) or microhomology-mediated break-induced repair (MMBIR). By contrast, recurrent de novo CNVs are often flanked by segmental duplications that mediate the generation of these rearrangements through non-allelic homologous recombination (NAHR) mechanism which may be similar during both paternal and maternal meiosis. No parent-of-origin or parental-age effect has been demonstrated for this class of CNVs, which occur relatively frequently and recur because of the predisposing chromosomal architecture [31, 32].

The widespread use of high resolution genomic microarrays and whole genome NGS has now made possible the analyses of rare de novo mutations (from SNVs and small indels to large-scale CNVs). Currently only few studies have been able to determine the parent-of-origin of de novo mutations. Overall, **de novo mutations are more frequent in SZ and ASD patients than in normal individuals and the mutational burden increases with paternal age** [33].

De novo point mutations are predominantly paternal in origin and positively correlated with paternal age at the time of conception of the child [26, 29, 34]. These results are very encouraging, however, the number of families analysed so far is too small to draw definitive conclusions about the correlation with maternal age, and whether the paternal effect is the same for single substitutions, small indels and large CNVs.

2.1.1 THE AIM OF THE PROJECT

The **aim** of the project is to determine the parent-of-origin of de novo SNVs, using large samples of parent-offspring trios affected with schizophrenia (SZ). This project is contained in the main one that consists on the determination of the parental origin of different types of *de novo* mutations (SNVs, small indels and large CNVs) using large samples of parent-offspring trios affected with schizophrenia or bipolar affective disorder.

2.2 SAMPLE DESCRIPTION AND DATASETS USED FOR *DE NOVO* SNVS DETECTION

The department where I am doing my final project is leading a whole exome sequencing project of 618 Bulgarian proband-offspring trios affected with SZ. Nearly 1000 de novo mutations (SNVs or small indels) have been identified and 430 of those are predicted to change the amino-acid sequence. About 90% of high-quality calls for non-synonymous and synonymous mutations have been confirmed with Sanger sequencing of the trios.

The first steps of the project before I came involved: from DNA collection, to validation of the called *de novo* SNV with Sanger sequencing. And with that they obtained around 650 coding and 450 non-coding de novo mutations for analysis. From these, the parent-of-origin of at least 60% of the mutations (N=600) can be established.

2.3 DETERMINING THE PARENT OF ORIGIN OF *DE NOVO* SNVS

The aim of the project is to find the parental origin of *de novo* mutations (not inherited).They could be originated in the oogenesis (mother) or what is thought to be more likely, in the spermatogenesis (father). In order to prove it, an informative SNP near each *de novo* SNV will be selected, and the child and parents genotype looked (Table 1).

	de novo SNV	SNP near the SNV
Child	**C**/T	A/G
Father	T/T	A/G
Mother	T/T	G/G

Table 1. The genotype of the parents and the chid is showed. The child presents the allele **C** as the *de novo* mutation. The parents present an informative SNP (Heterozygous for the child and the father, and homozygous for the mother).

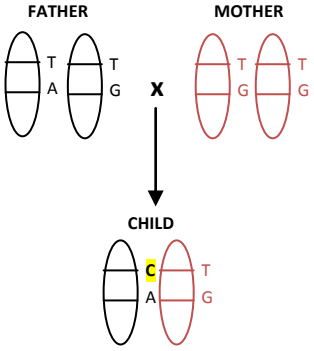


Figure 1. The *de novo* SNV (C) is transmitted from the father. We can see how is in the same chromosome as the allele A (father) of the informative SNP.

So, the first part of the project consisted on explore the NGS runs and the Affymetrix 6.0 data for the presence of other inherited informative SNPs in close proximity (within ~5000bp) of the de novo mutation, that lie (or don't lie) on the same sequence read (or pair-end read) with the mutation (Table 2). If the de novo mutation is on the same read as the inherited mutation, then it has occurred in the parent who transmitted it (and vice versa) (Figure 1).
The method works only if the proxy SNP is informative, which means that the child is heterozygous, and the parents are either homozygous for the alternative alleles, or only one of them is heterozygous for the transmitted SNP.

	Chr	POSITION	REF	ALT	GENE	SAMPLE ID	CHILD-META	PAT-META	MAT-META
de novo SNV	1	151028149	T	A	CDC42SE1	1402-1	[AD=6,7]	[AD=21,0]	[AD=17,0]
proxy SNP	1	151026815	C	T	CDC42SE1	1402-1	0/1	1/1	0/1

Table 2. Information of the position in the chromosome, the reference and the alternative alleles, the gene, the sample id, and the genotype of the child, and the parents for the *de novo* SNV with its proxy SNP is showed.

Once the list of all the de novo SNV, each with its informative SNP, was created, groups were made depending on the distance between the de novo and the proxy (within 1kb, 2kb, 3 kb,...), as the PCR conditions were going to be different depending on the size. And then, the primers for the first, second, third and the fourth set were designed. For each couple (proxy SNP + de novo SNV) a Forward and a Reverse primer for the 1st PCR, and then allele-specific primers for the 2nd PCR (using the Forward or Reverse from the 1st PCR) was needed.

Once the primers arrived, the laboratory work consisted on the method based on allele-specific Polymerase Chain Reaction (PCR) of the inherited SNPs that lie within ~5000bp away from a de novo mutation, to establish again whether the inherited SNP was on the same DNA strand as the mutation.

Briefly, the method included two-stage PCR. In the first stage a PCR product that encompasses both the *de novo* mutation and the inherited SNP was obtained (PCR1). This product was used for allele-specific PCR (PCR2) using two allele-specific primers, each complementary to the alternative alleles (e.g. C or T) of the inherited SNP or de novo mutation, and one universal primer placed some 50-100bp distal to the de novo mutation (or the inherited SNP, Figure 2).

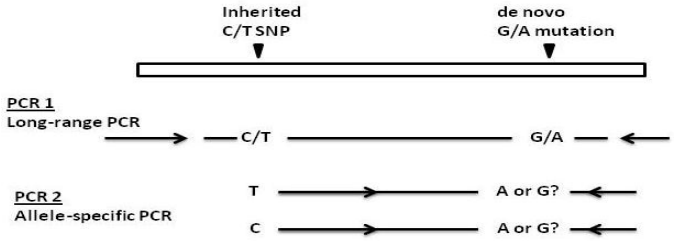
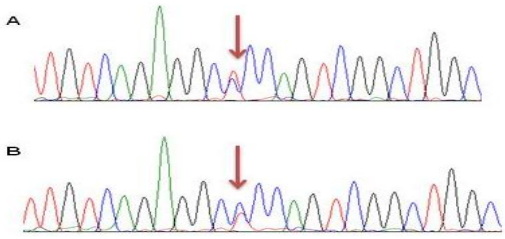


Figure2: Two-stage PCR amplification. Stage 1, a product that encompass both the inherited and the de novo SNP is obtained. Stage 2, this product is used for allele-specific amplification followed by sequencing.

The PCR2 was more efficient on the DNA strand that carried the allele complementary to the primer sequence (the C or T allele), and therefore the sequencing trace was higher for the allele at the de novo site, which was on the same strand. This was visible on the sequence trace (Figure 3).

Figure 3. Sequencing of allele-specific PCR product for allele T (A) and allele C (B). The position of the SNP is shown with red arrow. Allele T of the inherited SNP is on the same strand as allele A from the de novo mutation (T as sequenced in reverse direction), because in the allele-specific PCR product for allele T of the inherited SNP, allele T (red peak) from the de novo mutation is preferentially amplified Figure 2A). The opposite applies for the alternative allele (Figure 2B)



Following this process the parent of origin of *de novo* SNV is determined.

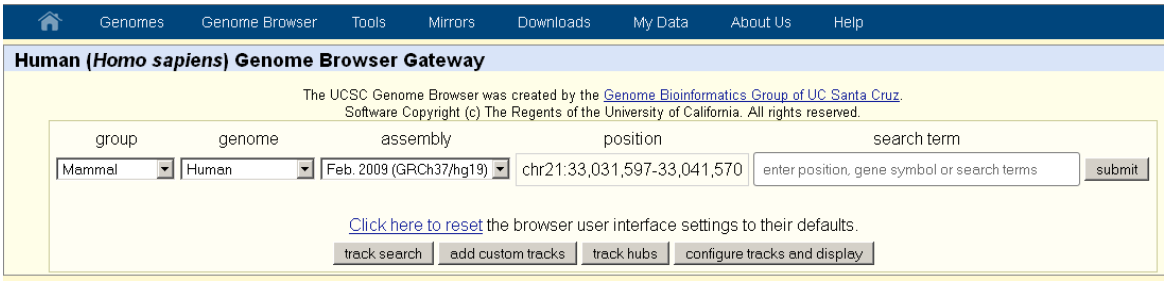
CHAPTER THREE: MATERIALS, METHODS AND RESULTS

3.1 PRIMERS

3.1.1 PRIMERS DESIGN

UCSC GENOME BIOINFORMATICS

In order to design the primers, first of all, the DNA was obtained, including the *de novo* SNV and the proxy SNP, with 500 bp upstream and downstream, in order to have enough place to find proper primers. For that, the UCSC Genome Bioinformatics website (http://genome.ucsc.edu/) was used.



Once the Fasta sequence was obtained, the previous step was repeated adding 5bp upstream and downstream in order to find the exact position of the *de novo* SNV and the proxy SNP.

Note: These are the colours I used along the project to identify them.

PRIMER 3

Once the sequence in FASTA is labelled with the specific colours, the part of the sequence that I wanted to be part of the product of the 1st PCR was included inside brackets "[...]", so the programme PRIMER 3 (http://frodo.wi.mit.edu/) could find the primers including that region.

In order to chose the best pair of primers from the list proposed by PRIMER3 they had to follow these rules:

- The primer that will be used for sequencing (the one opposite to the allele-specific one) should be located between 100 and 150 bp from the de novo SNV or the proxy SNP.
- Temperature: Forward and Reverse primers should have the same Temperature.
- Length: around 20 bp
- GC content: 50% or more hybridizes better with the sequence.
- Avoid self complementary
- Any: 3 or 4 is OK, when higher the number, there are more probabilities that it hybridizes between them or in another part.

- Max self complementary: less than 5, and we will also check it in another program Dobril Primer.
- Avoid poly-A, as the sequencing will not work properly

If the any of the primers showed was good enough, I had to create one that followed the rules.

DOBRIL PRIMER

In order **to check for the self complementary of the primer**, the DOBRIL PRIMER application (http://m034.pc.uwcm.ac.uk/FP_Primer.html) was used, where the polymorphism we wanted to target had to be between brackets. Once the primer was submitted, a list of the possible primers similar to it, and a dG(kCal/mol) value was given. We decided not to accept any primer with a **dG** above l**3**l.

The Dobril Primer also showed the **opt str.**

All possible primers with the so given parameters are:

THE FORWARD PRIMERS ARE:

primer sequence	number of bp	% GC	Tm [50mM K+]	dG (kCal/mol)		opt str	sub_opt str	Max dimer	Rank
accaggattaccagcctatagag	23	47.83	61.21	***-1.30***		opt	sub_opt	6	0.71
caccaggattaccagcctatagag	24	50.00	63.62	***-1.30***		opt	sub_opt	6	0.74
ccaggattaccagcctatagag	22	50.00	60.49	***-1.30***		opt	sub_opt	6	0.68

any comments or if you have problems using the programme do not hesitate to e-mail me at IvanovDK2@cf.ac.uk

And the most important, was to make sure the 1st PCR primers didn't **hybridize in other parts of the genome**. To check that, the tool BLAST or IN SILICO PCR, from the **UCSC Genome Bioinformatics** website was used.

2nd PCR PRIMERS

In order to create the primers for the 2nd PCR, there were two options:

- The Forward primer contained the first SNV (2)+ Reverse primer from PCR1
- The Reverse primer contained the last SNV (2) + Forward primer from PCR1

  *2: means that there are 2 primers specific for the SNV (one containing the reference and the other the alternative)

In order to **increase specificity**:

- Another mismatch was introduced in the allele-specific primers.
  Changing T<>G and C<>A.

		PCR	mismatch added a-c g-t	PRODUCT SIZE		len	tm	gc	any
GIT1	1	GIT1_PCR1_F	AGCCTCACCCCAGCAGTG	2061		18	62.49	66.67	3
		GIT1_PCR1_R	TGCCTGTTAACCCGGAATAC			20	59.82	50	6
		GIT1_PCR2_F_C	CATAGTCACTGTCGGCTCCc**C**	2022	allele specific primer over DENOVO	21	60.73	57.14	5
		GIT1_PCR2_F_T	CATAGTCACTGTCGGCTCCc**T**						
			use GIT1_PCR1_R						
BIRC6	2	BIRC6_PCR1_F	AAAGGGACTGGCTTTGGAAC	1583		20	60.48	50	5
		BIRC6_PCR1_R	GGATTATTTACCTGAATCATTTCG			24	58.06	33.33	6
		BIRC6_PCR2_R_C	ACATGGCTGGGATGAGGCc**G**	1548	allele specific primer over PROXY	20	65.79	60	4
		BIRC6_PCR2_R_T	ACATGGCTGGGATGAGGCc**A**						
			use BIRC6_PCR1_F						

We only wanted to amplify the concrete allele with the mutation, so introducing the mismatches, we made sure that it would hybridize correctly.

Sometimes, some of them weren't specific enough, and in order to increase specificity, we had to dilute more the 1st PCR product, reduce the number of cycles, etc.

SET 1 (16 *de novo* + proxy SNP)

#	SNP	chr	position	REF	ALT	GENES	func	SAMPLEID	BOX	WELL
1	denovo	1	227885581	G	T	.	intergenic-region	2064-1	4	C10
1	proxySNP	1	227885726	G	A	ZNF847P		2064-1		
2	denovo	3	148858887	G	A	HPS3	missense	1704-1	23	H2
2	proxySNP	3	148859191	C	T	HPS3		1704-1		
3	denovo	5	82789312	A	G	VCAN	onic,intronic,intronic,intr	1158-1	10	D4
3	proxySNP	5	82789647	A	G	VCAN	synonymousSNV	1158-1		
4	denovo	19	49300035	C	T	BCAT2	intronic,intronic	2155-1	9	H7
4	proxySNP	19	49300605	T	C	BCAT2		2155-1		
5	denovo	4	187003495	G	A	TLR3	missense	2232-1	9	D6
5	proxySNP	4	187004074	C	T	TLR3	nonsynonymousSNV	2232-1		
6	denovo	15	42148742	T	C	SPTBN5	missense	1516-1	15	B4
6	proxySNP	15	42149472	T	C	SPTBN5	nonsynonymousSNV	1516-1		
7	denovo	4	69202841	C	CAT	YTHDC1	frameshift,frameshift	2044-1	7	B12
7	proxySNP	4	69203646	A	G	YTHDC1		2044-1		
8	denovo	9	131847401	G	A	DOLPP1	intronic,intronic	1064-1	10	A1
8	proxySNP	9	131848382	T	A	DOLPP1		1064-1		
9	proxySNP	3	45637253	C	T	LIMD1	synonymousSNV	2145-1	6	C11
9	denovo	3	45637495	C	G	LIMD1	missense	2145-1		
10	proxySNP	13	76055602	G	A	TBC1D4	nonsynonymousSNV	2138-1	BB	E2
10	denovo	13	76055883	A	G	TBC1D4	silent	2138-1		
11	proxySNP	3	97852229	T	A	OR5H1	nonsynonymousSNV	2106-1	4	B12
11	denovo	3	97852516	A	G	.	intergenic-region	2106-1		
12	proxySNP	16	84213114	C	T	TAF1C	synonymousSNV	1042-1	3	F2
12	denovo	16	84213434	A	T	TAF1C	missense	1042-1		
13	proxySNP	3	184295008	C	A	EPHB3		2065-4	BB	D1
13	denovo	3	184295448	C	T	EPHB3	silent,splice	2065-4		
14	proxySNP	14	105238820	G	C	AKT1		2234-1	11	A1
14	denovo	14	105239278	C	T	AKT1	issense,missense,missens	2234-1		
15	proxySNP	4	1806044	C	T	FGFR3		2157-1	13	H2
15	denovo	4	1806503	C	T	FGFR3	intronic,intronic,intronic	2157-1		
16	proxySNP	3	125868513	A	G	ALDH1L1		1513-1	13	F3
16	denovo	3	125869273	C	T	ALDH1L1	silent,silent,silent	1513-1		

SET 2 (24 *de novo* + proxy SNP)

#	SNP	chr	position	REF	ALT	GENES	func	SAMPLEID	BOX	WELL
1	denovo	17	27902699	C	T	GIT1	missense,missense	2302-1	16	C5
	proxy	17	27904617	C	A	GIT1				
2	denovo	2	32773107	C	G	BIRC6	intronic	2016-1	2	F2
	proxy	2	32774505	C	T	BIRC6	synonymous SNV			
3	denovo	16	135513	C	T	MPG	ense,missense,miss	2265-1	15	F8
	proxy	16	136888	T	C	NPRL3				
4	denovo	19	36632048	C	T	CAPNS1	silent,silent	2156-1	11	C3
	proxy	19	36633277	C	G	CAPNS1				
5	denovo	11	124856585	G	T	CCDC15	intronic	2401-1	30	C5
	proxy	11	124857708	G	A	CCDC15	nonsynonymous SNV			
6	denovo	2	233406102	G	A	CHRNG	silent	2270-1	16	B4
	proxy	2	233407120	C	T	CHRNG				
7	denovo	3	36778993	G	A	DCLK3	silent	2138-4	BB	H2
	proxy	3	36779992	G	A	DCLK3	synonymous SNV			
8	denovo	8	144940886	C	T	EPPK1	missense	3019-1	2	G10
	proxy	8	144940117	C	A	EPPK1				
9	denovo	3	142682039	C	T	PAQR9	nonsense	2276-1	18	E7
	proxy	3	142681249	G	A	PAQR9	synonymous SNV			
10	denovo	10	101565129	T	C	ABCC2	intronic	1354-1	22	G1
	proxy	10	101564012	C	G	ABCC2	synonymous SNV			
11	denovo	8	9588574	G	A	TNKS	intronic	1017-1	4	C2
	proxy	8	9590755	T	A	TNKS				
12	denovo	7	6452471	G	A	DAGLB	nonsense,nonsense	1067-1	10	B2
	proxy	7	6449967	T	G	DAGLB	synonymous SNV			
13	denovo	3	48701533	C	T	.	intergenic-region	1114-1	3	F5
	proxy	3	48699519	C	T	CELSR3	synonymous SNV			
14	denovo	17	72349742	C	A	KIF19	intronic	2149-1	6	D12
	proxy	17	72346868	C	T	KIF19	nonsynonymous SNV			
15	denovo	5	141022521	T	C	FCHSD1	intronic	1139-1	19	A1
	proxy	5	141019881	G	C	FCHSD1				
16	denovo	1	171511275	G	C	PRRC2C	missense	1182-1	23	A1
	proxy	1	171510055	A	G	PRRC2C	nonsynonymous SNV			
17	denovo	3	49693941	G	T	BSN	missense	1311-1	23	B12
	proxy	3	49692454	G	A	BSN	nonsynonymous SNV			
18	denovo	21	28214245	G	A	ADAMTS1	missense	1358-1	21	E2
	proxy	21	28212760	G	A	ADAMTS1	synonymous SNV			
19	denovo	1	151028149	T	A	CDC42SE1	or-in45ag,intronic,s	1402-1	23	E12
	proxy	1	151026815	C	T	CDC42SE1				
20	denovo	17	80010157	C	CG	GPS1	intronic,intronic	2125-1	6	H7
	proxy	17	80008392	G	T	RFNG	synonymous SNV			
21	denovo	1	78430612	C	T	FUBP1	silent	2187-1	22	B3
	proxy	1	78429408	G	C	FUBP1				
22	denovo	4	437663	C	T	ABCA11P,ZNF721	missense	2208-1	10	E12
	proxy	4	436067	C	A	ZNF721	nonsynonymous SNV			
23	denovo	9	140509155	C	CAGGAGG	ARRDC1	codon-insertion	2222-1	8	B4
	proxy	9	140508031	A	G	ARRDC1				
24	denovo	3	184041660	C	T	EIF4G1	;silent,silent,silent,	2308-1	28	E2
	proxy	3	184039666	A	G	EIF4G1	nonsynonymous SNV			

3.1.2 WORKING DILUTIONS

Once the primers arrived, with its technical datasheet, the process could be started.

	1	2	3	4	5	6	7	8
A	FR1_ZNF847P_PCR1_F	FR1_ZNF847P_PCR1_R	FR1_ZNF847P_PCR2_F_mG	FR1_ZNF847P_PCR2_F_mT	FR9_LIMD1_PCR1_F	FR9_LIMD1_PCR1_R	FR9_LIMD1_PCR2_R_mG	FR9_LIMD1_PCR2_R_mC
B	FR2_HPS3_PCR1_F	FR2_HPS3_PCR1_R	FR2_HPS3_PCR2_R_mC	FR2_HPS3_PCR2_R_mT	FR10_TBC1D4_PCR1_F	FR10_TBC1D4_PCR1_R	FR10_TBC1D4_PCR1_R_mA	FR10_TBC1D4_PCR1_R_mG
C	FR3_VCAN_PCR1_F	FR3_VCAN_PCR1_R	FR3_VCAN_PCR2_R_mA	FR3_VCAN_PCR2_R_mG	FR11_OR5H1_PCR1_F	FR11_OR5H1_PCR1_R	FR11_OR5H1_PCR1_R_mA	FR11_OR5H1_PCR1_R_mG
D	FR4_BCAT2_PCR1_F	FR4_BCAT2_PCR1_R	FR4_BCAT2_PCR2_R_mT	FR4_BCAT2_PCR2_R_mC	FR12_TAF1C_PCR1_F	FR12_TAF1C_PCR1_R	FR12_TAF1C_PCR2_F_mC	FR12_TAF1C_PCR2_F_mT
E	FR5_TLR3_PCR1_F	FR5_TLR3_PCR1_R	FR5_TLR3_PCR2_F_mG	FR5_TLR3_PCR2_F_mA	FR13_EPHB3_PCR1_F	FR13_EPHB3_PCR1_R	FR13_EPHB3_PCR2_F_mC	FR13_EPHB3_PCR2_F_mA
F	FR6_SPTBN5_PCR1_F	FR6_SPTBN5_PCR1_R	FR6_SPTBN5_PCR2_F_mT	FR6_SPTBN5_PCR2_F_mC	FR14_AKT1_PCR1_F	FR14_AKT1_PCR1_R	FR14_AKT1_PCR2_F_mG	FR14_AKT1_PCR2_F_mC
G	FR7_YTHDC1_PCR1_F	FR7_YTHDC1_PCR1_R	FR7_YTHDC1_PCR2_F_mC	FR7_YTHDC1_PCR2_F_mCAT	FR15_FGFR3_PCR1_F	FR15_FGFR3_PCR1_R	FR15_FGFR3_PCR2_R_mC	FR15_FGFR3_PCR2_R_mT
H	FR8_DOLPP1_PCR1_F	FR8_DOLPP1_PCR1_R	FR8_DOLPP1_PCR2_F_mG	FR8_DOLPP1_PCR2_F_mA	FR16_ALDH1L1_PCR1_F	FR16_ALDH1L1_PCR1_R	FR16_ALDH1L1_PCR2_R_mC	FR16_ALDH1L1_PCR2_R_mT
	F	R	AS1	AS2	F	R	AS1	AS2

Table 3: SIGMA datasheet with the plate layout for the first set of primers (1-16).
F: Forward R: Reverse AS: Allele-Specific

In a new plate, I prepared the primer dilutions. The cross (X) in the following table refers to the primer I used for the 2nd PCR and the Sequencing.

	1	2	3	4	5	6	7	8
A		X			X			
B	X				X			
C	X				X			
D	X					X		
E		X				X		
F		X				X		
G		X			X			
H		X			X			
	F	**R**	**AS1**	**AS2**	**F**	**R**	**AS1**	**AS2**

In the following table the working dilutions are showed. For PCR1 : Forward + Reverse, PCR2: AS1 + Forward or Reverse; AS2 + Forward or Reverse, SEQUENCING: Forward or Reverse. Diluted with ultrapure water.

	1	2	3	4	5	6	7	8	9	10
A										
B										
C	3 µL P1	3µL P3	3µL P4				3 µL P5	3µL P7	3µL P8	
		3µL PX	3µL PX	3µL PX				3µL PX	3µL PX	3µL PX
D	3µL P2	(1/2)	(1/2)	(1/2)			3µL P6	(5/6)	(5/6)	(5/6)
E	95µL H20	95µL H20	95µL H20	97µL H20			95µL H20	95µL H20	95µL H20	97µL H20
F										
G										
H										
	PCR1	**PCR2**		**SEQ**			**PCR1**	**PCR2**		**SEQ**

For the second set of primers, the same procedure was followed although instead of for 16, for 24 *de novos* with its proxy.

Working dilutions for **SET 2**:

	1	2	3	4	5	6	7	8	9	10	11	12
A		x				x				x		
B	x					x			x			
C		x				x				x		
D	x					x			x			
E	x				x					x		
F	x					x				x		
G		x			x					x		
H		x				x				x		
	F	**R**	**AS1**	**AS2**	**F**	**R**	**AS1**	**AS2**	**F**	**R**	**AS1**	**AS2**

Plate layout

	1	2	3	4	5	6	7	8	9	10	11	12
A												
B												
C	3µL P1	3µL P3	3µLP4	3µL PX (1/2)	3µL P5	3µL P7	3µL P8	3µL PX (5/6)	3 µL P9	3µL P11	3µL P12	3µL PX (9/10)
D	3µL P2	3µL PX (1/2)	3µLPX (1/2)		3µL P6	3µL PX (5/6)	3µL PX (5/6)	97µL H20	3µL P10	3µL PX (9/10)	3µL PX (9/10)	97µL H20
E	95µL H20	95µL H20	95µL H20	97µL H20	95µL H20	95µL H20	95µL H20		95µL H20	95µL H20	95µL H20	
F												
G												
H												
	PCR1	**PCR2**		**SEQ**	**PCR1**	**PCR2**		**SEQ**	**PCR1**	**PCR2**		**SEQ**

3.2 POLYMERASE CHAIN REACTION (PCR)

The Polymerase Chain Reaction allowed us to amplify selectively our specific target DNA sequence within the total genomic DNA of each patient.

For the PCR, polymerase chain reaction buffer, forward and reverse oligonucleotide primers, deoxynucleotide (dNTP) mix, sterile molecular grade water, and DNA polymerase were needed. PCR reactions were carried in a final volume of 16 µL. Enzyme used was Taq polymerase, and PCR was performed on 4ng of DNA.

	1x	**x**
Buffer	1.2	
200 mcM dNTPs x4	1	
TaqQ	0.06	
H2O	3.74	
DNA = 4µL		
Primers mix = 2 µL		
PCR mix volume = 10µL		

The dNTPs were previously prepared to a final concentration of 5mM all.

dNTPs (100mM/each)		dNTPs **(1.25mM/each)** 5mM/all	
dATPs		12.5µL A	
dGTPs	For PCR →	12.5µL G	+ 950 µL H2O
dCTPs		12.5µL C	
dTTPs		12.5µL T	

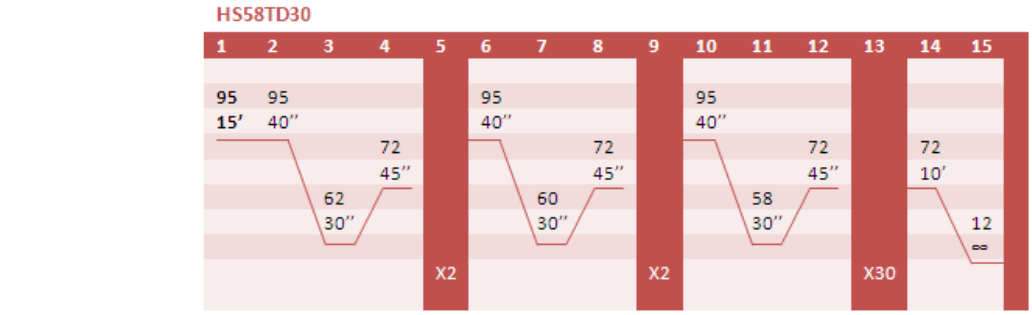
3.2.1 OPTIMISATION OF THE 1ST PCR

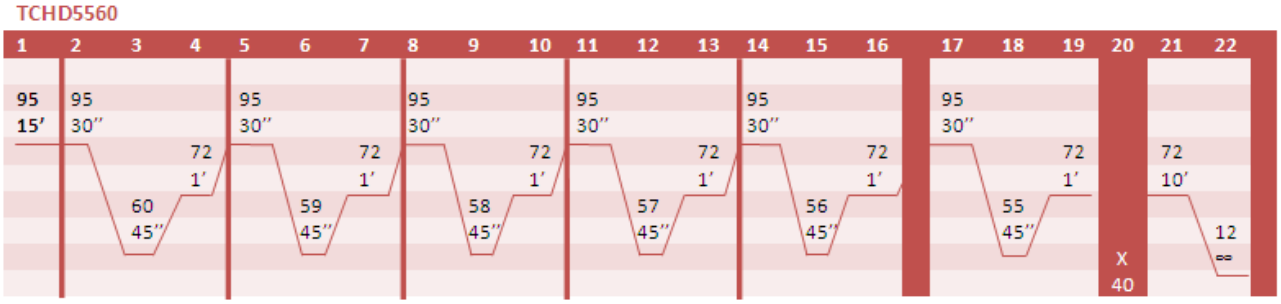
Different PCR programmes were run in order to optimise the Polymerase Chain Reaction and obtain a good product. All of them followed the following conditions, and in order to find the proper specific conditions for each set of primers, little variations were introduced between them. The extension time also needed to be modified depending on the length of the PCR products (1 min each 1000bp).

PCR		
Denature	95 °C	
Annealing	50-67 °C	
Extension	72 °C	1min/1kb

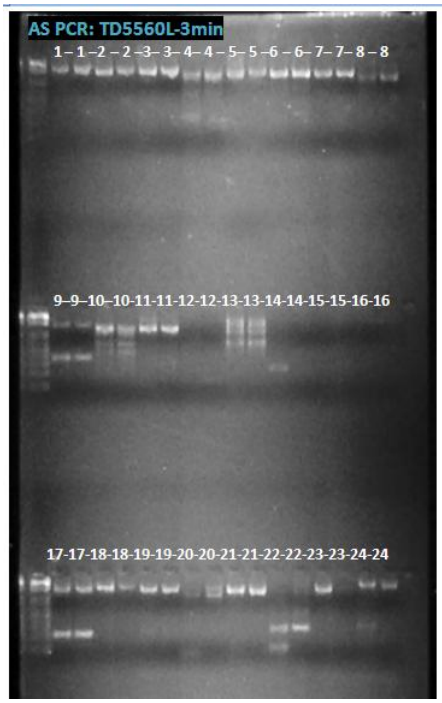
To summarise, the programs that worked better are the following ones:

For the **1ST SET**, both **HS58TD30** (Touch down from 62 degrees to 58 for the annealing temperature, and 30 cycles extension then) and **TCHD5560** (Touch down from 60 to 55, and 40 cycles extension) worked properly for the 1st PCR, although the second one worked better (elongation time of 1 min).





For the **2ND SET** (24 more *de novos* with its proxy), the extension time was increased to 3min, as the biggest fragments were of 3kb. In this case, the **HS58TD30L** worked better, as with the other one, unspecific products appeared (probably because the touchdown of the first one jumped from 2 degrees, which increased specificity).



#	GENE	Fragment size (bp)
1	GIT1	2061
2	BIRC6	1583
3	MPG	1540
4	CAPNS1	1489
5	CCDC15	1341
6	CHRNG	1213
7	DCLK3	1178
8	EPPK1	1014
9	PAQR9	1198
10	ABCC2	1293
11	TNKS	2438
12	DAGLB	2900
13	CELSR3	2426
14	KIF19	3087
15	FCHSD1	2917
16	PRRC2C	1739
17	BSN	1971
18	ADAMTS1	
19	CDC42SE1	1815
20	GPS1	2463
21	FUBP1	1455
22	ZNF721	2597
23	ARRDC1	1302
24	EIF4G1	2119

But anyway, 8 of 24 genes failed (labelled in yellow), and the rest of them, at the end of the process the sequencing results weren't really clear.

In order to find the specific conditions in which the primers would work properly, all the primers were run with other DNA samples at 60 degrees as the annealing temperature.
They worked better, only failed for a few genes.

For the ones that failed a Gradient PCR, (where in each column the annealing temperature was different) was run, and the proper temperature for the genes that didn't work at 60 degrees was found.

For the next sets, the best would probably be to run all of them at 60 degrees in the first time. And probably with other DNA samples first, so the patient's DNA is not wasted.

In the following table, the best temperature for each primer is showed.

Gene #	Annealing T°C	Gene #	Annealing T°C
1	60	**13**	-
2	60	**14**	-
3	60	**15**	60
4	55-57	**16**	60
5	60	**17**	60
6	60	**18**	60
7	60	**19**	60
8	68	**20**	-
9	55	**21**	60
10	60	**22**	64
11	60	**23**	60
12	60	**24**	60

Finally, we run for the 1st PCR of the 2nd set, with the patient's samples, with these conditions and programs:

For most of them at 60 degrees → **HS603**

- 2 TD from 64 to 62, 3 cycles each, and then at 60 35cycles.

For 4 and 9 at 55 degrees→ **HS553**

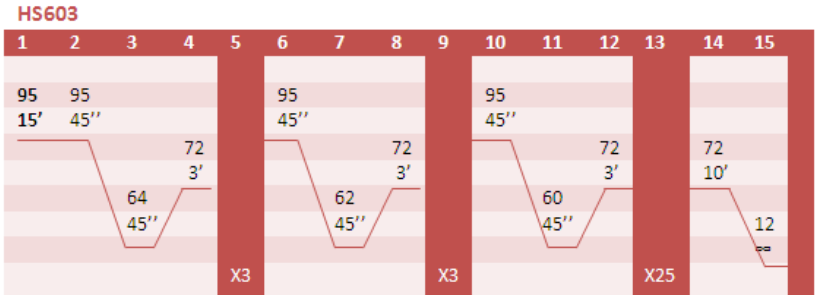
- Without TD, at 55 degrees 30 cycles.

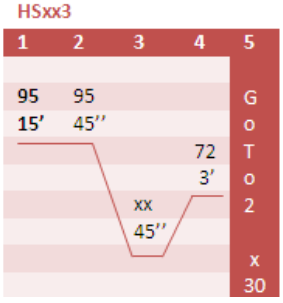
For 8 at 68 degrees → **HS683**

- Without TD, at 68 degrees 30 cycles.

For 22 at 64 degrees → **HS643**

- Without TD, at 64 degrees 30 cycles.





3.3 ALLELE SPECIFIC PCR

For the allele specific PCR, oligonucleotide primers were designed so as to discriminate between target DNA sequences that differ by a single oligonucleotide in the region of interest. The primers were designed to differ in the at the nucleotide that occurs at the extreme 3' terminus, because the DNA synthesis step in a PCR reaction is crucially dependent on correct base pairing at 3'end. Under stringent conditions, a mismatched primer doesn't initiate replication, whereas a matched primer does. The appearance of an amplification product therefore indicates the genotype.

The reagents and DNA used were the same and in the same concentrations as for the 1st PCR and in the same concentrations. The primer mix used was the specific for the 2nd PCR.

3.3.1 OPTIMISATION OF THE 2ND PCR

For the **set1**, the same programs as for the first PCR: HS58TD30, TCHD5560 were run, and also with a variation of the second one, reducing the number of cycles to 12. From the sequencing data, only a few of them worked properly and I was able to see the phase of the *de novo* mutation for them, for the others, the primers weren't specific enough, so both sequences were amplified (it didn't discriminate).

In order to increase the specificity of the primers, dilutions of the 1st PCR product were necessary (15x dilution wasn't enough, so finally, another 10x dilution more was needed).

Running the programme with only 12 cycles (TD556012) wasn't enough, (the 2nd PCR product that we can see from the gel was too poor) but the final dilution was ideal as the primers were specific (the few that work discriminate between the alleles, I could see the phase and determine the parental origin of them).
The balance between having too much product (we cannot discriminate the phase of the mutation with the proxy when sequencing) and too less (we don't know if it didn't work or if it wasn't enough cycles) needed to be found.

Reading some papers, we saw that some groups did straight away the allele specific-specific PCR (without a previous 1st PCR), So we also did it that way, because if we could avoid to do the 1st PCR, we would be able to safe in money and time. The AS-PCR with the programe TD556030 (30 cycles instead of 40) was set, but it didn't work properly, only for a few genes, which wasn't enough.

Finally, the 2nd PCR was set up, after diluting the 1st PCR product twice, and with the programme **TD556020** and **TD556030** (which are a variation of the original TCHD5560, now with 20 and 30 cycles).

After cleaning up the 2nd PCR product, preparing for the sequencing, cleaning up the product and sequencing, the results were ideal! With these conditions and the parental origin for them could be determined.

3.4 AGAROSE GEL ELECTROPHORESIS

Agarose gels were made as 2% agarose in 1x TAE buffer (40mM Tris-acetic acid, 10 mM EDTA pH 8.0). The agarose was dissolved in TAE by heating in the microwave oven on full power for 1.30 min, and cooled to 50 degrees C. Ethidium bromide solution was added (2 µL into 100 mL), and the gel solution was poured into a gel former with a comb inserted.

When set, the comb was removed, the gel transferred to an electrophoresis tank, and 1x TAE added to cover the surface off the gel. Samples were first mixed with 6x gel loading buffer (30% (v/v) glycerol, 20 mM EDTA ph 8.0, 0.25% (w/v) bromophenol blue, 0.25% (w/v) xylene cyanol in water) and located in the wells. A 1kb Plus DNA Ladder was loaded when needed beside the sample and used as a size marker.

The gel was run at 100v for approximately 30 min. Ethidium bromide stained nucleic acid was visualised using a transilluminator system.

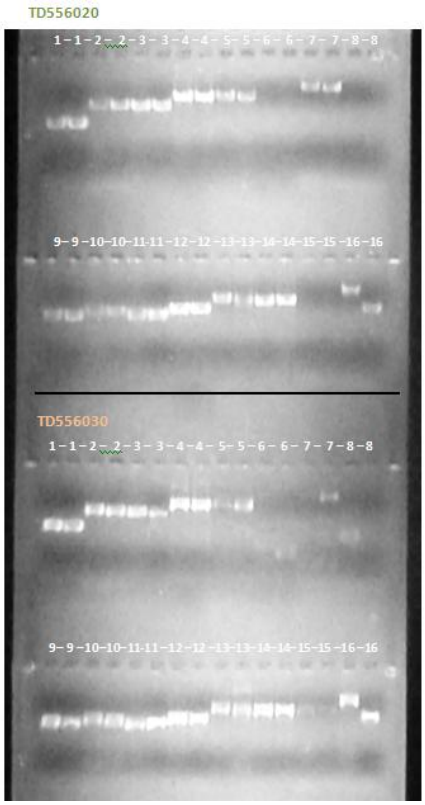


Figure 4.
Allele-Specific PCR product are checked in the agarose gel. Two different programmes were run, TD556020 and TD556030 (variations in the number of cycles of the original TCHD5560, decreasing to 20 and 30 cycles instead of 40). These PCR products correspond to **Set 1** final result, after applying the proper conditions in order to determine the parental origin of each of the 16 *de novo* of that set.

Table 3.
Set 1 plate layout sent to sequence.

	1	2	3	4	5	6	7	8
A	ZNF847P_F_mG	ZNF847P_F_mT	LIMD1_R_mG	LIMD1_R_mC	ZNF847P_F_mG	ZNF847P_F_mT	LIMD1_R_mG	LIMD1_R_mC
B	HPS3_R_mC	HPS3_R_mT	TBC1D4_R_mA	TBC1D4_R_mG	HPS3_R_mC	HPS3_R_mT	TBC1D4_R_mA	TBC1D4_R_mG
C	VCAN_R_mA	VCAN_R_mG	OR5H1_R_mA	OR5H1_R_mG	VCAN_R_mA	VCAN_R_mG	OR5H1_R_mA	OR5H1_R_mG
D	BCAT2_R_mT	BCAT2_R_mC	TAF1C_F_mC	TAF1C_F_mT	BCAT2_R_mT	BCAT2_R_mC	TAF1C_F_mC	TAF1C_F_mT
E	TLR3_F_mG	TLR3_F_mA	EPHB3_F_mC	EPHB3_F_mA	TLR3_F_mG	TLR3_F_mA	EPHB3_F_mC	EPHB3_F_mA
F	SPTBN5_F_mT	SPTBN5_F_mC	AKT1_F_mG	AKT1_F_mC	SPTBN5_F_mT	SPTBN5_F_mC	AKT1_F_mG	AKT1_F_mC
G	YTHDC1_F_mC	YTHDC1_F_mCAT	FGFR3_R_mC	FGFR3_R_mT	YTHDC1_F_mC	YTHDC1_F_mCAT	FGFR3_R_mC	FGFR3_R_mT
H	DOLPP1_F_mG	DOLPP1_F_mA	ALDH1L1_R_mC	ALDH1L1_R_mT	DOLPP1_F_mG	DOLPP1_F_mA	ALDH1L1_R_mC	ALDH1L1_R_mT
	TD556020				TD556030			

3.5 PURIFICATION OF PCR PRODUCTS

If the PCR worked, the PCR products were cleaned in order to carry on the process. The leftover primers and the P of the dNTPs needed to be removed. Exonuclease I and Shrimp Alkaline Phosphatase were used for that.

	1x
SAP	0.5 µL
Exonuclease I	0.1 µL
Water	4.4 µL
	5 µL

Programme: **SAP**	
1h	37 °C
20’	80 °C
5’	95 °C
∞	12 °C

3.6 SEQUENCING REACTION

In the Sanger method (chain terminator sequencing, dideoxychain termination method), the DNA strand to be analysed is used as a template and primed with an oligonucleotide. In the sequencing reaction, the DNA polymerase will extend from the 3'-OH end of the sequencing primer to generate complimentary strand in the presence of a mixture of four naturally occurring deoxynucleotides and one of them being a dideoxynucleotide which can not bind other deoxynucleotides because its 3' end is modified.

An improved approach called 'dye terminator sequencing' uses each of the dideoxynucleotide chain-terminators labeled with a separate fluorescent dye, which fluoresces at a different wavelength instead of labeled primers. This approach leads to a complete sequence in a single reaction, rather than the four needed with the labeled-primer approach.

The four bases are detected using different fluorescent labels. These are detected and represented as 'peaks' of different colours that can then be interpreted to determine the base sequence.

In order to obtain the product for sequencing, sequencing buffer, big dye, purified water, the forward or the reverse primer and the cleaned product of the 2nd PCR will be mixed and run in the programme called SEQUENCE in the thermocycler.

SEQUENCING MIX	**1x**
Big dye	0.25 µL
Water	0.875 µL
Seq buffer	1.875 µL
Total seq mix	**3 µL**
Primer	**2 µL**
Cleaned PCR prod	**5 µL**

Programme: **SEQUENCE** (2h)		
1	2’	96 °C
2	10’’	96 °C
3	5’’	50 °C
4	4’	60 °C
5	GO TO 2	25x
6	∞	15 °C

3.6.1 PURIFICATION USING THE ROBOT

Sequencing reactions need to be cleaned up prior to injection on the **ABI 3100 sequencer** in order to remove both the unincorporated dideoxynucleotides as well as salts from the reaction buffers. The cleaning up protocols was performed with the robot **BIOMEK NX**.

The reagents needed were the Agencourt CleanSEQ, ethanol and purified water, and the protocol used **CleanSeq.96.2.40**

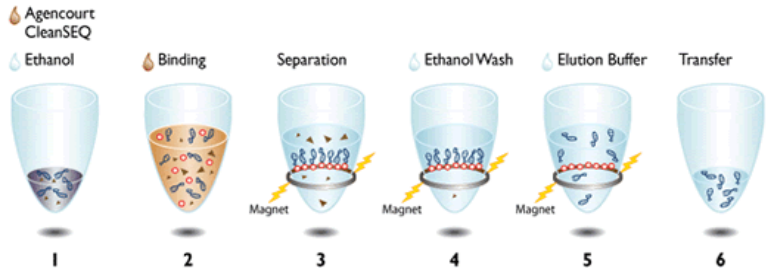
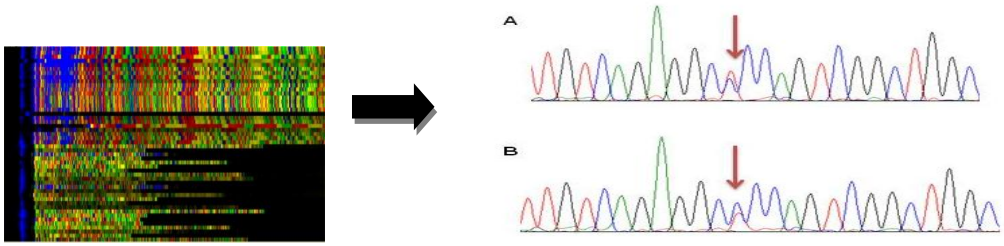


Figure 5. Process Overview. 1. Add Agencourt CleanSEQ reagent and ethanol to sequencing reaction 2. Bind sequencing products to magnetic beads 3. Separate sequencing products from contaminants with magnetic field 4. Wash with ethanol 5. Elute from magnetic particles 6. Transfer away from magnetic beads

3.6.2 ABI 3100 SEQUENCER

Each product generated will carry a fluorescent molecule at one extremity. After cleaning, the sequences are resuspended in loading buffer and run on the capillaries. A laser detects the passage of the fluorescent products and transmits this information to the computer, which will generate a text file and a chromatogram where T appears in red, G in black, C in blue, and A in green. To read the sequencing data the **Sequence Scanner (Applied Biosystems)** was used.



After sequencing the allele-specific PCR product, we were able to distinguish the phase of each *de novo* SNV with the proxy SNP. The position of the SNP is showed with red arrow. (See 2.3 Determining the parent of origin of *de novo* mutations).

CHAPTER FOUR: OUTCOMES AND BENEFITS

4.1 FINAL RESULTS

Working in the project *IDENTIFYING THE PARENT-OF-ORIGIN OF DE NOVO SNVs IN SCHIZOPHRENIA* I was able to screen for proxy SNPs for each called *de novo* SNV, from a huge dataset with the NGS information, and create a list of them. That first step of the project required a lot of time, but it had to be done in order to have an informative SNP within 5kb next to the *de novo* in order to perform the Allele-Specific PCR and sequence to read the parent of origin of each.

Once the list was created, I grouped them depending on the size (from 1kb to 5kb), and after learning how to design and select the best primers (depending on different parameters as length, temperature, self-complementarily, GC content...) using programmes such as UCSC, Primer3 and Dobril primer, I designed the primers for them.

Before starting working with the proper samples, I did a previous set up of the method with 2 other primers with non precious samples, and that helped to obtain a closer idea of which PCR programmes were working better. Anyway, along the project, I learned that different variations were always needed to be introduced, and that depending on the product size and in each primer, modifications in the programmes were needed (annealing temperature, extension time...), the same happened with the dilution of the product. For this reason, the optimisation of the method (the set up) was needed to be done for each set of primers.

For the 1st set, after changing different parameters, the best conditions were:

- 1st PCR: The PCR programme **TCHD5560**, which had 5 steps of touch down, with the annealing temperature decreasing from 60 degrees to 56, and then, 40 cycles at 55 degrees. The denaturing temperature was 95 degrees during 30 seconds, the annealing temperature with the touch down from 60 to 55 during 45' and finally, the extension step was at 72 1min (as the product size of the fragments were 1kb).
- Before the second PCR, in order to obtain a clear sequencing data, the 1st PCR products needed to be **diluted twice**, 15x and 10x.
- 2nd PCR: The PCR programme **TD556020 or TD556030** were the best, wich follow the same steps as the original TCHD5560, but instead of 40 cycles, 20 and 30.
- For the clean up, and the sequencing reaction, I always used used the programs **SAP-EXO** (after the PCRs) and with the robot the **CleanSeq** (after sequencing reaction) in the thermo cycler. Working with the **robot Biomak NX** it was also a great experience as it was the first time, and I felt comfortable using it.

For the 2st set, which fragments were up to 3kb, the optimisation was more complicated. To start with, I tried with the same programmes that worked for the 1st set, but changing the extension time.

Although only a few of them didn't work, I carried on with all of them, following the same process as for the Set 1, but the sequencing results weren't good. We wanted to obtain better results, so I tried changing the annealing temperature to 60 degrees, as most of the primers were going to work better in such conditions. And then, for the ones that didn't at 60, a gradient PCR.

After changing different parameters, the best conditions for the 1st PCR were:
For most of them at 60 degrees with the programme **HS603** (2 TD from 64 to 62, 3 cycles each, and then at 60 35cycles). And for the rest: two with the programme **HS553** (Without TD, and at 55 degrees 30 cycles), one with **HS683** (Without TD, and at 68 degrees 30 cycles), and another one with **HS643** (Without TD, and at 64 degrees 30 cycles).

With the second set I didn't obtain final results as I had to submit the project. But the sequencing data was very clear and discriminate for 10 of the 16 genes from the set 1, and I was able to work out the phase of the de novo with the proxy, and find the parental origin of the *de novo* SNV **(3 maternal : 7 paternal).**

#	GENE	Parent of Origin
1	ZNF847P	M
2	HPS3	P
3	VCAN	M
4	BCAT2	M
5	TLR3	P
6	SPTBN5	Failed PCR
7	YTHDC1	Not discriminate
8	DOLPP1	Failed PCR
9	LIMD1	P
10	TBC1D4	P
11	OR5H1	P
12	TAF1C	Failed PCR
13	EPHB3	P
14	AKT1	P
15	FGFR3	Failed PCR
16	ALDH1L1	Not discriminate

Moreover, in a previous pilot laboratory experiment performed by this group, they called the the parent-of-origin (**2 maternal : 6 paternal**) of 8 cases.

These results prove that although specific variations are needed in the set up, the method is valid to determinate the parent of origin of *de novo* mutations, and if we add both results obtained, for the moment, the parent-of-origin of *de novo* mutations is: **5 maternal: 13 paternal**.

4.2 CONCLUSIONS

Identifying the mechanisms of mutation formation in humans is of a great theoretical value in genetics and medicine. De novo mutations are important both as sources of diversity in evolution and for their immediate impact on diseases.
Establish the parental origin of up to 600 *de novo* mutations (about 60% of 1000 *de novos*) will be a long-term project. Although I was only able to determine the parent of origin of 10 *de novo* SNV (3 maternal:7 paternal), plus 8 other established in a previous study (2maternal:6paternal), once this study is finished, an accurate ratio for maternal:paternal origin of de novo SNPs found in SZ trios will be established and will also allow to perform correlations between the rate of such mutations and paternal age.

The results obtained with the main project (in which this one is involved) will have great **prognostic value** and can be implemented in the **mental health prevention programs**. In order to advice the families about the increase of risk with the **paternal age** (once the hypothesis is proved) at the time of conception of the child, and reduce the number of schizophrenia cases in the future generations.

Once this project is finished, more information about the different type of mutations related with schizophrenia will be available. The evidence for the involvement of these de novo mutations will also offer a new approach to detect potentially pathogenic variants and may help to explain partially the causes and inheritance pattern of SZ and also how SZ prevalence remains stable in the general population despite low fecundity. This will provide a better understanding of the mechanism underlying this complex disease, and will help us to be a step closer in order to find key targets for therapeutic approaches.

FUTURE DIRECTION

The relationship between genotype and phenotype in schizophrenia is likely to be mediated by complex causal pathways involving gene-gene and gene-environment interactions.

At present, no single, or major, environmental risk factor influencing the incidence of schizophrenia has been conclusively demonstrated. Further studies using large samples are required to evaluate potential risk factors, antecedents, and predictors for which the present evidence is inconclusive.

BIBLIOGRAPHY

1- B.J. Sadock, V.A. Sadock, P. Ruiz, *Textbook of psychiatry* (Lippincott Williams & Wilkins, Philadelphia, ed. 9, 2009*).*

2- Andreasen NC. Scale for the Assessment of Positive Symptoms. Department of Psychiatry University of Iowa College of Medicine *52242.* (1995)

3- Bertelsen, A. & Gottesman, I.I. Schizoaffective psychoses -- genetical clues to classification. American Journal of Medical Genetics (Neuropsychiatric Genetics), 60*, 7-11* (1995)

4- *Diagnostic and Statistical Manual of Mental Disorders Fourth Edition (DSM-IV)* (American Psychiatric Association, 1994)

5- World Health Organisation: Mental Health, 2002

6- Eric Q. Wu *et al.* The economic burden of schizophrenia in the United States in 2002. J. Clin. Psychiatry, 66-9 (2005)

7- Moldin, S.O. *et al.* Replicated psychometric correlates of schizophrenia. American Journal of Psychiatry, 148, 762-767 (1991)

8- Rodriguez-Murillo L, *et al.* The genetic architecture of schizophrenia: new mutations and emerging paradigms. Annu Rev Med 63:63-80 (2012)

9- Steinberg KM *et al.* Structural diversity and African origin of the 17q21.31 inversion polymorphism. Nat Gen. 44(8):872-80 (2012)

10- 4 Roach, J.C. *et al.* Analysis of genetic inheritance in a family quartet by whole-genome sequencing. Science 328, 636–639 5 (2010)

11- Conrad, D.F. *et al.* Variation in genome-wide mutation rates within and between human families. Nat. Genet. 43, 712–714 (2011)

12- 6 Kong, A. *et al.* Rate of de novo mutations and the importance of father's age to disease risk. Nature 488, 471–475 7 (2012)

13- Campbell, C.D. *et al*. Estimating the human mutation rate using autozygosity in a founder population. Nat. Genet. 44, 1277–1281 8 (2012)

14- Michaelson, Jacob J. *et al.* Whole-genome sequencing in autism identifies hot spots for de novo germline mutation. Cell 151, 1431–1442 (2012)

15- Rees E. *et al.* De Novo Mutation in Schizophrenia. Schizophrenia Bulletin vol. 38 no. 3 pp. 377–381 (2012)

16- *Highlights of Changes from DSM-IV-TR to DSM-5* (American Psychiatric Association, 2013)

17- O'Roak, B.J. *et al*. Exome sequencing in sporadic autism spectrum disorders identifies severe *de novo* mutations. Nat. Genet. 43, 585–589 (2011).

18- Vissers, L.E. *et al*. A *de novo* paradigm for mental retardation. Nat. Genet. 42, 1109–1112 (2010).

19- Awadalla, P. *et al*. Direct measure of the *de novo* mutation rate in autism and schizophrenia cohorts. Am. J. Hum. Genet. 87, 316–324 (2010).

20- Bin Xu *et al*. Exome sequencing supports a *de novo* mutational paradigm for schizophrenia. Nat. Genet. 43, 864–868 (2011)

21- National human genome research institute

22- Lichtenstein, P. *et al*. Common genetic determinants of schizophrenia and bipolar disorder in Swedish families: a population-based study. Lancet 373, 234–239 (2009).

23- Sebat J, Lakshmi B, Malhotra D, Troge J, Lese-Martin C, Walsh T, Yamrom B, Yoon S, Krasnitz A, Kendall J, Leotta A, Pai D, Zhang R, Lee YH, Hicks J, Spence SJ, Lee AT, Puura K, Lehtimäki T, Ledbetter D, Gregersen PK, Bregman J, Sutcliffe SJ et al. Strong association of de novo copy number mutations with autism. science 2007 Apr 20;316(5823):445-9.

24- Kirov G, Pocklington AJ, Holmans P, Ivanov D, Ikeda M, Ruderfer D, Moran J, Chambert K, Toncheva D, Georgieva L, Grozeva D, Fjodorova M, Wollerton R, Rees E, Nikolov I, van de Lagemaat LN, Bayés A,

Fernandez E, Olason PI, Böttcher Y, Komiyama NH, Collins MO, Choudhary J, Stefansson K, Stefansson H, Grant SG, Purcell S, Sklar P, O'Donovan MC, Owen MJ. De novo CNV analysis implicates specific abnormalities of postsynaptic signalling complexes in the pathogenesis of schizophrenia.Mol Psychiatry. 2012 Feb;17(2):142-53.

25- Girard SL, Gauthier J, Noreau A, Xiong L, Zhou S, Jouan L, Dionne-Laporte A, Spiegelman D, Henrion E, Diallo O, Thibodeau P, Bachand I, Bao JY, Tong AH, Lin CH, Millet B, Jaafari N, Joober R, Dion PA, Lok S, Krebs MO, Rouleau GA. Increased exonic de novo mutation rate in individuals with schizophrenia. Nat Genet. 2011 Jul 10;43(9):860-3.

26- O'Roak BJ, Vives L, Fu W, Egertson JD, Stanaway IB, Phelps IG, Carvill G, Kumar A, Lee C, Ankenman K, Munson J, Hiatt JB, Turner EH, Levy R, O'Day DR, Krumm N et al.Multiplex targeted sequencing identifies recurrently mutated genes in autism spectrum disorders. Science. 2012 Dec 21;338(6114):1619-22.

27- Crow, J. F. The origins, patterns and implications of human spontaneous mutation. Nature Rev. Genet. 1, 40–47 (2000).

28- Malaspina D, Harlap S, Fennig S, Heiman D, Nahon D, Feldman D, Susser ES. Advancing paternal age and the risk of schizophrenia. Arch Gen Psychiatry. 2001 Apr;58(4):361-7

29- Hultman CM, Sandin S, Levine SZ, Lichtenstein P, Reichenberg A. Advancing paternal age and risk of autism: new evidence from a population-based study and a meta-analysis of epidemiological studies.Mol Psychiatry. 2011 Dec;16(12):1203-12.

30- Glaser RL, Jabs EW. Dear old dad. Sci Aging Knowledge Environ. 2004 Jan 21;2004(3):re1.

31- Hehir-Kwa JY, Rodríguez-Santiago B, Vissers LE, de Leeuw N, Pfundt R, Buitelaar JK, Pérez-Jurado LA, Veltman JA. De novo copy number variants associated with intellectual disability have a paternal origin and age bias.J Med Genet. 2011 Nov;48(11):776-8.

32- Lupski JR, Stankiewicz P. Genomic disorders: molecular mechanisms for rearrangements and conveyed phenotypes. PLoS Genet. 2005 Dec;1(6):e49. Review.

33- Kong A, Frigge ML, Masson G, Besenbacher S, Sulem P, Magnusson G, Gudjonsson SA, Sigurdsson A, Jonasdottir A, Jonasdottir A, Wong WS, Sigurdsson G, Walters GB, Steinberg S, Helgason H, Thorleifsson G, Gudbjartsson DF, Helgason A, Magnusson OT, Thorsteinsdottir U, Stefansson K. Rate of de novo mutations and the importance of father's age to disease risk.Nature. 2012 Aug 23;488(7412):471-5.

34- Michaelson JJ, Shi Y, Gujral M, Zheng H, Malhotra D, Jin X, Jian M, Liu G, Greer D, Bhandari A, Wu W, Corominas R, Peoples A, Koren A et al.Whole-genome sequencing in autism identifies hot spots for de novo germline mutation. Cell. 2012 Dec 21;151(7):1431-42.