2012

Using Next Generation Sequencing (NGS) to identify and predict microRNAs (miRNAs) potentially affecting Schizophrenia and Bipolar Disorder

Vernell Williamson
Virginia Commonwealth University

Using Next Generation Sequencing (NGS) to identify and predict microRNAs (miRNAs) potentially affecting Schizophrenia and Bipolar Disorder

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy at Virginia Commonwealth University.

By

Vernell Seay Williamson,

Master of Arts, Wake Forest University, 1997

Master of Science, Virginia State University, 2004

Thesis advisor: Vladimir Vladimirov, MD/PhD, Psychiatry

Acknowledgement

The author wishes to thank several people. I would like to thank my family for their love, support and patience during the time it has taken for me to graduate. I would also like to thank my advisor, Dr. Vladimirov for his help and for his direction with this project. Lastly, I would to thank the various students and researchers within the Virginia Institute for Psychiatric and Behavioral Genetics, for their advice, support and counseling during my time as a student.

TABLE OF CONTENTS

Lists of tables

Lists of figures

List of Terms

Monozygotic – “identical”: type of twins that result from the splitting of one zygote into two children in utero

Dizygotic - “fraternal”: type of twins that result from the fertilization of two separate eggs by two separate sperm

Concordance – probability that a second twin will have a disease if the first is affected

Proband – the first individual examined sometimes the index

Heritability – proportion of variance explained by genetic factors

Allelic Specific Risk (λ_s) - risk to siblings having a disease compared to the rest of the population

Risk (λ) - risk of a disease to a relative compared to the rest of population

Penetrance – proportion of individuals carrying a particular variant of a gene that also expresses the phenotype

Genetic heterogeneity – a single phenotype or disorder is caused by multiple alleles at multiple loci

Prevalence – the total number of cases in a given statistical population divided by the total population

Morbidity rate - a term that can refer to the incidence rate or the prevalence rate

Incidence rate – the probability of developing a given disease within a specific period of time

Gene X Environment Interaction – “G x E”, refers to the situations where an individual's response to the environment is genetically defined

Locus – specific location of a gene

Phenotype – sum total of an organism’s observable measurable traits

Genotype – genetic makeup of an individual

DNA – deoxyribonucleic acid

Monogenetic traits – a trait determined by a single gene

Polygenic tratits - a trait for which the phenotype depends on alleles at many different genes

RNA – ribonucleic acid

LOD score - log of the odds of observing some association between two alleles.

Seed – bases 2-8 on the mature sequence of microRNA

Mature sequence – the active molecule of the microRNA, measures approximately 22 bases

Hairpin - also known as a precursor strand. Measuring ~ 60 bases, this is intermediate stage of microRNA biogenesis

Hardy Weinberg Equilibrium – principle which states that allelic and genotype frequencies within a population will stay constant unless it is perturbed through non-random mating, mutation, selection, genetic drift, gene flow, meiotic drive. Represented by $p+ q = 1$; $p^2+2pq+q^2 = 1$.

Alleles – two or more forms of a specific gene, variant or genetic locus

Allelic frequency – proportion of all copies of a variant within a population

Genotype – genetic makeup of a cell

Haplotype – a combination of SNPs that are inherited together across loci

Endophenotype – a behavioral phenotype with a strong genetic component, term borrowed from genetic epidemiology.

Global Abstract

USING NEXT GENERATION SEQUENCING (NGS) TO IDENTIFY AND PREDICT MICRORNAS (MIRNAS) POTENTIALLY IMPACTING SCHIZOPHRENIA AND BIPOLAR DISORDER

by

Vernell Seay Williamson

Advisor: Vladimir Vladimirov

The last decade has seen considerable research focusing on understanding the factors underlying schizophrenia and bipolar disorder. A major challenge encountered in studying these disorders, however, has been the contribution of genetic, or etiological, heterogeneity to the so-called “missing heritability” [1-6]. Further, recent successes of large-scale genome-wide association studies (GWAS) have nonetheless seen only limited advancements in the delineation of the specific roles of implicated genes in disease pathophysiology.

The study of microRNAs (miRNAs), given their ability to alter the transcription of hundreds of targeted genes, has the potential to expand our understanding of how

certain genes relate to schizophrenia and bipolar disorder. Indeed, the strongest finding of one recent mega-analysis by the Psychiatric GWAS consortium (PGC) was for a miRNA, though little can be said presently about its particular role in the etiologies of schizophrenia and bipolar disorder [52].

Next generation sequencing (NGS) is a versatile technology that can be used to directly sequence either DNA or RNA, thus providing valuable information on variation in the genome and in the transcriptome. A variation of NGS, MicroSeq, focuses on small RNAs and can be used to detect novel, as well as known, miRNAs [26,125, 126].

The following thesis describes the role of miRNAs in schizophrenia and bipolar disorder in various experimental settings. As an index of the interaction between multiple genes and between the genome and the environment, miRNAs are great potential biomarkers for complex disorders such as schizophrenia and bipolar disorder.

Project Overview

In the forthcoming discussion, a number of concepts will be presented regarding schizophrenia and bipolar disorder and how researchers have sought to understand the genetic architectures of these disorders. We present in this thesis in six chapters, described as follows. Chapter one is intended as an introduction to both disorders, by describing the hallmark symptoms and what is currently known in regard of their genetic structure. Additionally, chapter one presents earlier epidemiological and genetic research to highlight better the potential of miRNAs' studies to complement existing research. Chapters two, three, and four describe three separate but related studies focusing on the detection, validation, and assessment of miRNAs. These chapters are described in the manner of a research manuscript, i.e. abstract, introduction, material

and methods, results and discussion.

Because of their complex nature, schizophrenia and bipolar disorder have not yielded to cursory approaches, paradigms, or specific assays. Rather, their undoubtedly complex genetic structure necessitates an approach which integrates information from both genetic epidemiological and molecular genetic studies. In chapter five, we present a synthesis of the entire project, summarizing key findings and highlighting project limitations. Lastly, chapter six examines what future steps could be taken with this research.

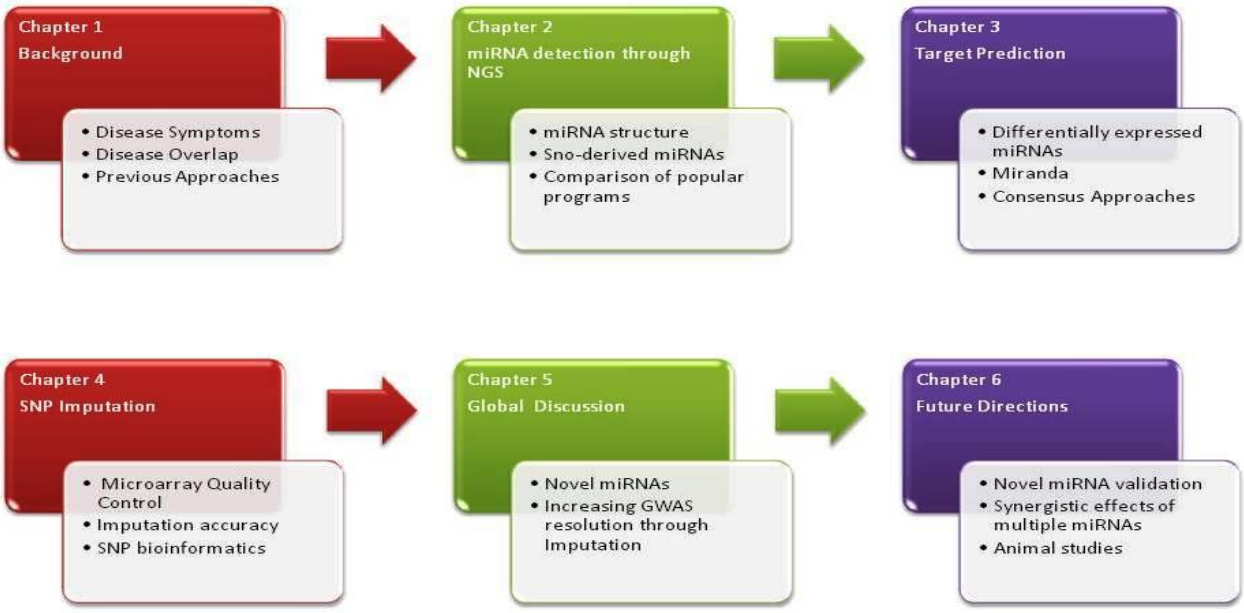


Figure 1 Project Overview

Chapter 1 Rationale and Background

Any thesis describing the use of a particular approach to the study of a given disease must provide justification for its use within the larger framework of the field. To that end, we present previous research in the field of psychiatry, starting with epidemiological, followed by genetic and expression studies, and finally with recent miRNA research. Despite their respective limitations, these studies have established that risk of developing schizophrenia and bipolar disorder has an inherited component or, stated another way, that the incidence of these disorders is genetically mediated [139]. However, mediation is different from outright control and, therefore, the field has struggled to find definite sources of causality.

Defining the Phenotype

That the genetic architecture of schizophrenia and bipolar disorder remains elusive is not surprising. Diagnosis of these disorders has been largely one of self-report and observation, relying on tools such as the diagnostic and statistic manual of mental disorders (DSM-IV), the OPCRIT+ and the ICD-10 [170, 171, 172, 173]. Indeed, a slight inconsistency across these tools has existed historically [181]. Further, considerable clinical and genetic evidence exists for overlap between schizophrenia and bipolar disorder, as well as with other disorders such as schizoaffective disorder and autism, making the problem of defining the phenotype even more acute [91,92, 95,127, 128]. Diagnostic criteria for schizophrenia can be partitioned into six separate subfields: 1) Characteristic Symptoms, 2) Social/Occupational dysfunction, 3) Duration, 4) Schizoaffective and Mood Disorder exclusion, 5) Substance/General Medical Condition

exclusion, and 6) Relationship to a Pervasive Developmental Disorder [95, 172, 211]. A threshold of minimum number of symptoms has been set in order for a diagnosis to be place, i.e. hallucinations, delusions, disorganized speech, catatonic behavior, and affective flattening need be present for a month for a diagnosis of schizophrenia to be made [95]. Bipolar Disorder can be classified as type 1 or type 2, depending on the frequency and severity of manic episodes [95,212].

The “flexibility” of the phenotypic definition, in each case, would necessarily introduce a large amount of variability, which would further complicate the search for causative variation. Regardless of the particular causes of these difficulties, it remains clear that additional avenues of research are needed if we are to truly understand the genetic architecture of schizophrenia and bipolar disorder.

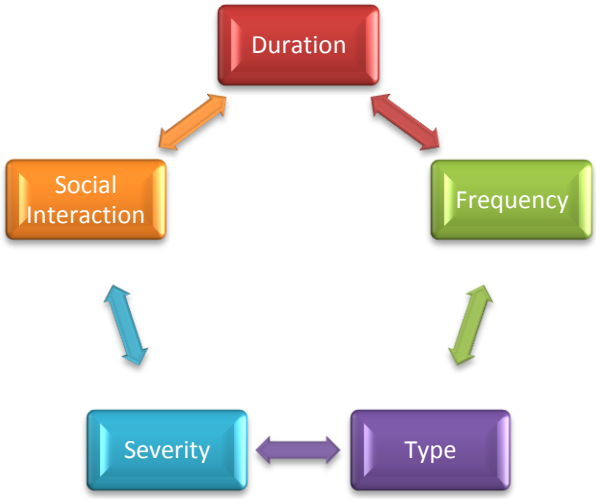


Figure 2 Categories by which schizophrenia and bipolar disorder are classified. Schizophrenia is diagnosed if two or more of the classic symptoms, e.g. delusions, hallucinations, disorganized speech/behavior, catatonic behavior, and negative symptoms are present for a significant period of time during a 1-month period. A diagnosis for bipolar disorder requires the alternation of both manic and depressive symptoms.

Schizophrenia and bipolar disorder affect approximately 1% of the population and, though not as prevalent as major depression, their toll on the quality of human life is keen and solely felt [165]. It is likely because of the complex polygenic nature of both disorders that multiple sources of information are required for a true understanding to be achieved. By presenting a brief overview of molecular psychiatry in this chapter, it is hoped that the reader will come to realize how the study of miRNAs might be used to complement existing research. That is, miRNAs, through their nature of mediating with the outside environment, might be used as a way of explaining the phenotypic variability currently observed by researchers in schizophrenia and bipolar disorder.

Moreover, a single miRNA is estimated to target as many as 200 genes. Therefore, identifying even a small group of miRNAs implicated in disease etiology might simplify the job of identifying causality considerably [53, 174]. In order to successfully integrate the study of miRNAs into molecular psychiatry, however, they must first be understood with respect to their own individual actions upon gene targets and each other. Simply put, the numbers of miRNAs must be clearly delimited and defined. To that end, this thesis presents an approach to detecting and validating novel miRNA within the larger disease framework of schizophrenia and bipolar disorder.

Genetic Epidemiology of psychiatric disease

Family, twin and adoption studies were originally conceived as a way of documenting the correlation of risk (λ) to relatives of affected persons. Family studies asked the question, “is risk in developing the disease inherited and does it aggregate in

families?” Twin studies examined the level of risk present among types of twins, i.e. monozygotic (MZ) and dizygotic (DZ). Adoption studies were able to disentangle the effects of the environment from genetics by examining the course of the disease in children adopted away from affected parents [152].

Family studies

Starting first with Rudin, family studies have shown that a child of a parent with a psychiatric disorder has a tenfold elevated risk of developing the disorder compared with general population [91,92]. Early family studies were criticized on primarily methodological grounds, but later studies suggest that risk for schizophrenia is approximately 2-9% in first degree relatives and, for bipolar disorder to range from 3-15% [139, 145]. Further, morbid risk was shown to diminish as the amount of shared genetic material diminished [140] (see figure 3). Of particular interest is the similar percentage of morbid risk for MZ twins (48%) and offspring of dual matings (46%). These two types of relationships would, conceivably, share the largest amount of genetic material.

One key criticism of family studies has been the inability to parse whether the observed effects are entirely due to genetic effects, the environment, or some combination of both [145]. Additional criticisms include small sample size and the lack of systematic ascertainment and proper controls [181]. Despite these criticisms, family studies have demonstrated consistently that risk associated with developing schizophrenia and/or bipolar disorder has an inherited element.

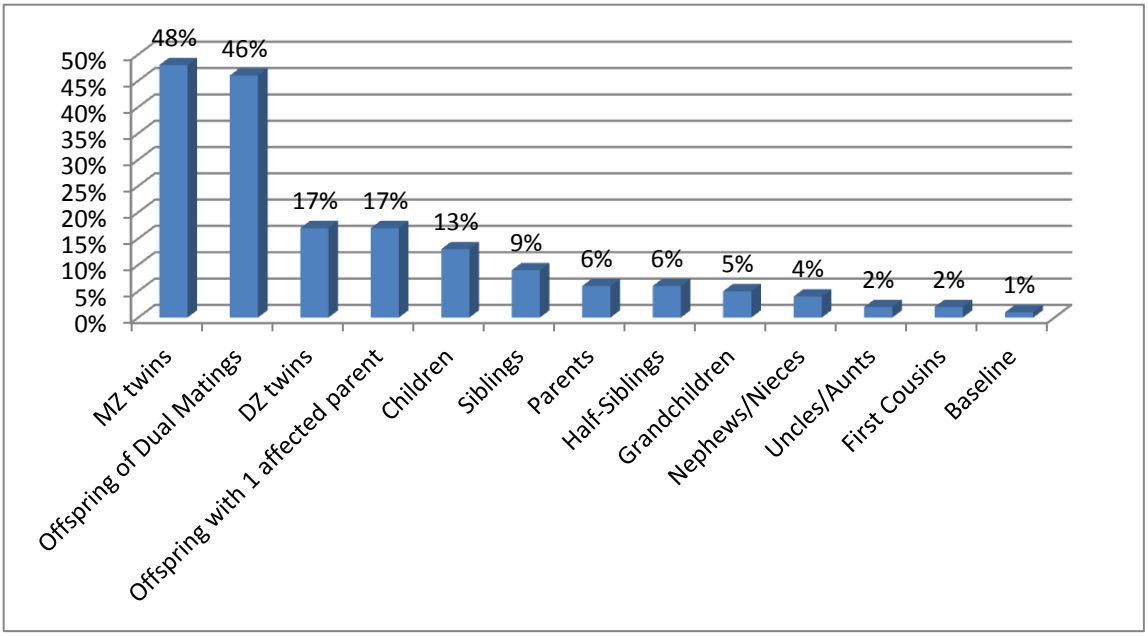


Figure 3 Lifetime morbid risk (MR) for schizophrenia in various classes of relatives. Image adapted from Gottesman.

Twin studies

Twin studies represent an approach to quantify directly the overall genetic and environmental components of risk shared by family members, measured in terms of rates of concordance between twins. Concordance is defined as the probability that a twin will have the disorder if the other is affected. For monogenetic traits, healthy monozygotic (MZ) twins demonstrate a genetic concordance of 1 and dizygotic (DZ) twins a concordance of 0.5. In complex polygenic diseases such as schizophrenia and bipolar disorder, however, the observed rates vary across studies. Such studies have also allowed for additional questions to be asked, including the role of the environment and the possibility of epigenetic mechanisms influencing discordance between monozygotic twins. Twin studies have been criticized for assuming that environments

are equal between MZ and DZ twin pairs [181].

Modern twin studies from Denmark, Switzerland, Germany and the United Kingdom have all benefited from hospital registries which have allowed for more systematic and focused study than the previous family based approaches [91,138,139]. In addition, hospital registries have enabled larger study designs, further strengthening conclusions. Concordance rates have been less than 1 and variable across individual studies, however [142, 181]. Concordance for schizophrenia has been estimated to be ~40-50% for MZ twins and 17% for DZ twins [4, 9, 140, 145, 152]. Concordance for bipolar disorder has been estimated to be ~67% for MZ twins, compared to 20% in DZ twins [4, 91, 140]. In addition, in many of these twin studies a notable degree of discordance between monozygotic twins has been observed, which some take as evidence in support of epigenetic mechanisms [3,4, 140, 152].

Adoption Studies

Adoption studies have asked whether increased risk in family members is still present even though parents and offspring do not share a common environment; there were four experimental designs employed in adoption studies [181]. In the first, children adopted away from affected parents were followed. An example of this study design is a recent Finish-based study profiling 361 families, which showed that 4.9% of children adopted away from schizophrenic mothers ultimately developed the disorder, compared to 1.1% among the offspring of control mothers [142]. The second type of design focuses on rates of disease occurrence in biological family members and controls. An example of this design type was a study of bipolar disorder which found that 7% of biological parents developed the disorder, compared to 1.8% of adoptive parents [146].

The third design type, termed “cross-fostering”, examined adopted children who were “unaffected” at the time of adoption but later came to be diagnosed with the disorder [152]. Lastly, the fourth type examined children adopted away from affected parents, along with the adoptive parents themselves, thus providing insight into the relevance of rearing environment [152]. At the core of these studies is the notion that the source of stress, i.e. the presence of a sick parent, might give rise to symptoms in children and whether the mere presence of such stress could precipitate disease onset. Despite yielding relatively low percentages, these studies suggest that risk is persistent even with a change in the rearing environment, further strengthening the argument for genetic inheritance [145,146].

The true impact of family, twin, and adoptions studies, then, has been to demonstrate that schizophrenia and bipolar disorder are substantially heritable and to quantify the relative role of genetics and environment in their etiology. In order to appreciate their full contribution to the field, one must consider these paradigms as a unit. Despite differences in diagnostic criteria and concept, these studies provide remarkably consistent findings and a rationale for pursuing genetic research towards uncovering the biological underpinnings of disease etiology.

Genetic Research

Linkage

Linkage studies represent the first attempt at discerning the particular genetic loci which comprise the genetic structure of psychiatric disease. The objective of linkage studies was to map genomic regions where sets of genes/loci were co-inherited by

affected members of family group. Linked loci segregate together during meiosis; often in linkage association, standard tandem repeats (STRs) are used as a point of focus. A logarithmic score of odds (LOD score: equation below) measures the likelihood of the observed data in a situation of no linkage (free recombination of theta value =0.5) compared to the likelihood of the data at a specific value of theta less than 0.5. A LOD of 3 or greater (a likelihood ratio equals to 1000 to 1) is considered to evidence of linkage between features.

$$LOD\ score = \log\left[\frac{likelihood\ associated\ linkage}{likelihood\ \ no\ associated\ \ linkage}\right]$$

Linkage studies have provided support for the roles of several regions in the etiologies of schizophrenia or bipolar disorder, including 6p24-22, 1q21-22, 13q32-34, 8p21-22, 6q16-25, 22q11-12, 5q21-q33, 10p15-p11, and 1q42 [86,100, 101]. Many of these reported linkage peaks encompass quite large genomic regions and have not been robust to replication efforts, i.e. genes within these linkage peaks have not yet been unambiguously identified or confirmed. Furthermore, these linkage studies were limited by small sample sizes and specification of unconfirmed genetic models and polygenetic inheritance. Of note, however, is the identification of a linkage peak on chromosome 22q13.1 and its association with what is now known about the genetic structure of Velo-cardio-facial syndrome (VCFS) [139]. This finding is illustrates how linkage studies have led to novel approaches to asking questions about a specific region and the role of structural variants in schizophrenia.

Structural variants

VCFS is caused by a copy number deletion (CND) on 22q13.1, and recent studies have documented that 28% of VCFS patients demonstrate a psychotic phenotype resulting from that deletion [139,178, 179, 181, 182]. Structural variants or genomic deletion and/or duplications underlie a number of neurodevelopmental disorders including autism and can vary in size from 1 kilobase to several megabases [182, 216].

In VCFS, the deletion size varies between1.5 to 3 megabases (Mb), affecting 35 to 60 known genes [183]. One of the genes in 22q11 is catechol-O-methyl transferase (*COMT*) which is involved in the biodegradation of catecholamine. *COMT* has a functional SNP (rs4680), which has been demonstrated significant associations in a number of candidate gene association studies (see following section). The deletion present on 22q11 is an example of a rare structural variant which until recently has been thought to play a limited role on the etiology of schizophrenia and bipolar disorder. It has been estimated that structural variants account for, at most, 10-15% of schizophrenia cases and as a whole, are not expected to explain a large amount of total population risk [181, 214]. They are seen as highly penetrant and of recent origin, often specific to individual families [214]. In a study of 418 persons, individuals with schizophrenia were found to be more than three times as likely than controls ($p=0.0008$) to have a structural variant affecting 1 or more genes [214].

Candidate Gene Association

Candidate gene association studies, in contrast to linkage analysis, search for susceptibility genes that are present in a population rather than a family and have

focused, principally, on single nucleotide polymorphisms (SNPs) as an index feature [175]. The most commonly employed experimental design in candidate gene association is the case control format because it is easier to collect large numbers of subjects, as compared to the familial cases [154, 155, 168, 176]. SNPs can be selected based on their predicted effect on a specific protein (nonsynonymous SNPs), or gene activity (splicing enhancers, stop codons) and the relationship to each other in terms of linkage disequilibrium (figure 3) [169, 176]. Synonymous or noncoding SNPs, also important, are expected to affect gene function indirectly [176]. One unique SNP function, that of affecting miRNA secondary structure, will be explored in greater detail in chapter four of this thesis. Unlike the microsatellites typical of linkage studies, SNPs are generally of low information content (i.e. nucleotide diversity) and candidate gene association studies are more successful in isolate populations [176].

The key shortcoming of candidate gene studies has been the fact that genes are investigated individually often based on *a priori* assumptions regarding its role in the biology of the disorder. Simply put, candidate gene studies have never provided a gene that has been rigorously and unambiguously replicated across multiple studies. Approximately 800 genes have been tested using this approach and none of them have been fully established [91].

A number of other genes have, however, been identified through other approaches such as positional follow-up of linkage studies and these genes have been replicated to a limited degree in association studies. These genes include 1) catechol methyl transferase (*COMT*, 22q11.21), 2) neuregulin 1 (*NRG1*, 8p12), 3) dysbindin (*DTNBP1, 6p22.3*), 4) diacylglycerol kinase (*DGKH*, 13q14.11) and 6) ankyrin-g (*ANK3*,

10q21) [100, 101, 123, 139, 155, 156, 157, 180].

In the *COMT* gene, the functional SNP rs4680 (Val108/158Met) has been proven to reduce expression of the functional isoform of *COMT* four-fold [198]. *COMT* is responsible for degradation of dopamine and is located in a region of the genome implicated in schizophrenia by linkage and CNV studies (see previous section for a discussion of VCFS). The frequency of this polymorphism varies worldwide, ranging from 1% in South American populations to 62% in Europeans [198,199]. There is evidence of over-representation of one version of this SNP (Met/Met) in poor-responders to anti-psychotics [198, 200].

The SNP rs3924999 located in the second exon of *NRG1*, changes the amino acid arginine (Arg) to glutamine (Gln) [201,202]. A number of different isoforms are well known to be produced by *NRG1*; some of these isoforms induce growth and differentiation of epithelial, neuronal and glial cells [203]. The core haplotype containing rs3924999 has been tested in a number of populations, including Chinese Han family trios. In that study, studying 246 Chinese families and using PCR-based restriction fragment length polymorphism and high-performance liquid chromatography, rs3924999 was significant in transmission disequilibrium tests ($p = 0.007752$) and the haplotype containing this SNP also demonstrated significance ($X^{2=}46.068$, $df=7$, $p <0.00001$).

Association signals for genes such as RGS4 and DISC1, originally identified through expression studies and translocation respectively, have likewise been inconclusive [181, 185, 186]. *RGS4* belongs to a gene family regulating G-protein signaling pathways. Expression of *RGS4* was shown to be decreased across the cerebral cortex of schizophrenic patients [163, 204]. Studies testing SNPs rs951436,

rs951439, and rs2661319 have generated mixed results, depending on experimental platform and population [204,205]. In a group of 218 schizophrenic Taiwanese families, no association was observed either by individual SNP or haplotype [204].

The gene *DISC1* is involved in cell proliferation, differentiation, migration, neuronal axon growth and cell-to-cell adhesion. In cell models, the truncated gene fails to interact with its binding partners, fasciculation and elongation protein zeta-1 (*FEZ1*), lissencephaly 1 protein (LIS1), and nuclear distribution element –like (*NUDEL*) [206]. The disruption of this gene was originally found to segregate with psychotic symptoms in a Scottish pedigree [205, 206]. Additional family studies detected no evidence of this disruption but instead identified SNPs associated with a reduction in hippocampal structure [205, 206, 214]. In another association study, a three-SNP haplotype (hCV219779(C)-rs821597 (G)-rs821616 (A)] was shown to be significantly associated (P = 0.002) [214]. Still, other genes such as multiple EGF-like domains (MEGF10), which do not possess a cogent biologically based reason for selection, yet have produced an positive association nonetheless [154]. Though its function was unknown at the time of its testing, *MEGF10* has since been linked to myopathy, respiratory distress and dysphagia [215].

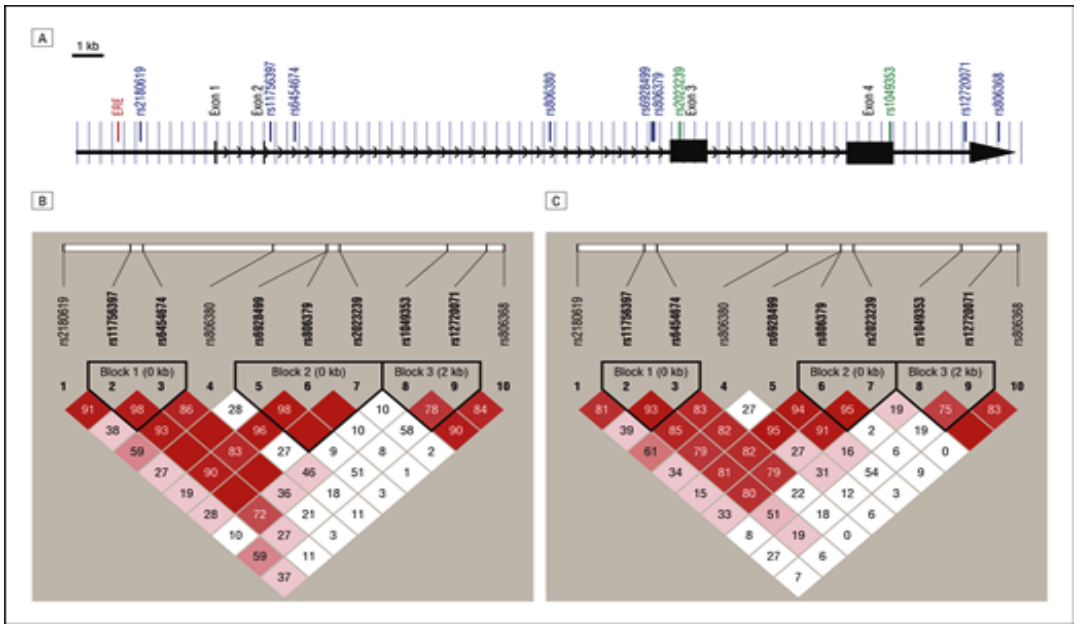


Figure 4 Linkage disequilibrium map as generated by the program Haploview. The gene pictured here is cannabinoid receptor 1 (CNR1), implicated in a number of disorders including nicotine addiction. This gene currently has 240 documented variants. Image adapted from Chen et al 2008.

Genome-wide Association (GWAS)

As a result of the completion of the human genome project, the identification of millions of catalogued polymorphisms have allowed GWAS to arise as an unbiased approach to assess genome-wide variation, while at the same time it preserves single variant resolution, that may point to a disease specific association. GWAS capitalize on the presence of linkage disequilibrium between variants to effectively minimize the number of variants needed to cover the genome and focuses on variants that occur in relatively high frequency in the population. They are, therefore, explicitly testing for common variation (minor allele frequency MAF >= 5%) present in the population. Nearly 20 separate GWAS (figure 5) have been published implicating genes such as transcription factor 4 (*TCF4*), neurogranin (*NRGN*) and the major histocompatibility

complex region (MHC) [75, 158, 159, 160]. Early GWAS used pooling and small sample sizes largely due to cost constraints, but these were quickly abandoned as the technology become more efficient to use [187,188,189]. Consortia such as the International Schizophrenia Consortium (ISC), Molecular Genetics of Schizophrenia (MGS), and the Psychiatric GWAS Consortium (PGC) have arisen to maximize sample size and available resources [75]. A number of consistent trends have begun to emerge from these GWAS. In particular the MHC region and *TCF4* have been replicated in a number of primary studies [75,158, 159, 160, 220, 221]. Stefansson et al. studied the DNA from eight separate European locations in 7662 cases and 29,053 controls and identified seven significant SNPs in the MHC region (rs6913660, rs13219354, rs6932590, rs13211507, rs3131296, rs12807809, rs9960767) that survived correction for multiple testing [221]. In the *TCF4* gene, only SNP rs9960767 was significant after follow-up ($p = 4.1 \times 10^{-9}$). The MHC region was also reported by Li *et al* in a group of Han Chinese (N = 2496 cases, N = 5184 controls), though through a smaller set of the same SNPs (rs6932590: p = 0.00096, rs3131296: $p = 1.29 \times 10^{-6}$, rs3130375: $p = 1.76 \times 10^{-5}$) [220]. One significant SNP in *TCF4* (rs2958182: $p = 3.64 \times 10^{-6}$) was identified by Li *et al*; this SNP was near to rs9960767 [220, 221].

Concerns arising from current GWAS studies include non-overlap of primary results and a limited number of findings.

Table 1. Schizophrenia genome-wide association studies and genes/loci significantly associated with risk at $P < 5 \times 10^{-8}$

			Primary sample		Replication sample				
First Author	Reference	Markers	Cases	Controls	Cases	Controls	Locus	Gene/Region	*P* value
Mah[a]	[6]	25 494	320	325	–	–	–	–	–
Lencz[b]	[18]	439 511	178	144	–	–	–	–	–
Sullivan[b]	[19]	492 900	738	733	–	–	–	–	–
O'Donovan[b]	[20]	362 532	479	2937	6666	9897	–	–	–
Shifman[a]	[21]	510 552	660	1100	–	–	–	–	–
Kirov[a]	[22]	433 680	574	605	–	–	–	–	–
Need[b]	[5]	312 565	871	863	1460	12 995	–	–	–
Shi[b,c]	[23••]	≥696 788	3967	3626	5327	16 424	6p22.1	MHC region	9.5×10^{-9}
Stefansson[b,c]	[24••]	314 868	2663	13 498	4999	15 555	6p21.3–22.1	MHC region	1.1×10^{-9} to 1.4×10^{-12}
							11q24.2	NRGN	2.4×10^{-9}
							18q21.2	TCF4	4.1×10^{-9}
Purcell[b,c]	[9••]	739 995	3322	3587	–	–	6p21.3–22.1	MHC region	9.5×10^{-9}
Chen[b]	[25]	≥446 225	1658	1655	11 380	14 708	–	–	–
Athanasiu[b]	[26]	572 888	201	305	2663	13 780	–	–	–
Williams[b,d]	[27•]	176	479	2937	18 945	38 675	2q32.1	ZNF804A	2.5×10^{-11}
Ikeda	[28•]	297 645	575	564	1990	5389	–	–	–
Steinberg[b,d]	[29•]	39	7946	19 036	10 260	23 500	2p15.1	VRK2	1.9×10^{-8}
							6p21.3–22.1	MHC region	$\leq 1.6 \times 10^{-10}$
							11q24.2	NRGN	2.8×10^{-9}
							18q21.2	TCF4	7.8×10^{-9}
							18q21.2	TCF4/CCDC68	4.2×10^{-9}
Ripke[b]	[30••]	1 252 901	9394	12 462	8442	21 397	1p23.3	MIR137	1.6×10^{-11}
							2q32.3	PCGEM1	4.7×10^{-8}
							6p21.3–22.1	TRIM26/MHC region	2.2×10^{-12}
							8p23.2	CSMD1	4.1×10^{-8}
							8q21.3	MMP16	2.8×10^{-8}
							10q24.32	CNNM2	1.8×10^{-9}
							10q24.33	NT5C2	1.1×10^{-8}
							18q21.2	CCDC68	2.6×10^{-10}
							18q21.2	TCF4	2.4×10^{-8e}
							11q24.2	STT3A	2.9×10^{-8e}
Yue	[31•]	546 561	746	1599	4027	5603	11p11.2	TSPAN18	4.8×10^{-11}
							6p21–22.1	MHC region	$<5.2 \times 10^{-10}$
Shi	[32•]	493 203	3750	6468	4383	4539	8p12	LSM1/WHSC1L1	1.27×10^{-10}
					3830	14 724	1q24.2	BRP44	9.5×10^{-9}

Figure 5 Recent GWAS generating a significant finding at a p < 5 X 10-8. A number of genes which have been replicated across the studies include TCF4 and the MHC region. Image adapted from Bergen and Petryshen (2012).

Recently in the largest GWAS to date (17836 cases and 33859 controls, respectively) the PGC group reported their most significant novel finding to be a polymorphism

(rs1625579) located in the primary transcript of a miRNA gene, hsa-miR-137, thus providing the strongest evidence for miRNA involvement in the etiology of schizophrenia and bipolar disorder. Smaller studies stemming from this GWAS have, additionally, corroborated a shared relationship between bipolar disorder and schizophrenia [52, 94, 98]. Hsa-miR-137, identified in the PGC GWAS is implicated in neural development and neurite formation [52, 161]. In stage one of this GWAS, seventeen predicted targets were enriched for association ($p < 10^{-4}$) which was nearly as twice as many as control genes. Verified gene targets for hsa-mir-137 include *TCF4*, cub and sushi domain-coding protein1 (*CSMD1*). That hsa-mir-137 and its targets were both identified on the same GWAS provides strong support for the role of miRNAs in schizophrenia.

Gene	Location	Associated SNP	P(GC-adjusted)	OR
miR137	1p21.3	rs1625579	2.65×10^{-6}	1.11
PCGEM1	2q32.3	rs17662626	1.70×10^{-3}	1.16
TRIM26	6p21.3-p22.1	rs2021722	1.55×10^{-3}	1.10
CSMD1	8p23.2	rs10503253	7.60×10^{-3}	1.08
MMP16	8q21.3	rs7004633	0.011	1.05
CNNM2	10q24.32	rs7914558	1.l07 x 10^{-3}	1.08
NT5C2	10q24.33	rs11191580	5.09×10^{-3}	1.09
STT3A	11q24.4	rs548181	0.068	1.04
CCDC68	18q21.2	rs12966547	2.29×10^{-5}	1.08
TCF4	18q21.2	rs17512836	0.085	1.08

Table 1Top Targets from PGC GWAS studies. Among them include *TCF4*, and hsa-mir-137 which has been shown to be involved in neurite development and branching.

Expression studies – Protein Coding Genes

Like GWAS, expression profiling of protein coding genes has also used high throughput technology to quantitatively assess variation on a wide scale, with the principal focus being protein-coding genes [162,163]. The intent of this research was to provide a more realistic assessment of the real time functionality of protein coding genes. Likewise, the use of postmortem brain tissue from affected subjects has been viewed as a more immediate way of addressing the problem by directly assessing expression levels in the tissue believed to be most affected, though these approaches are not without problems. Several trends have been identified in these studies including disturbances in synaptic function, energy metabolism, and oligodendrocyte function but results are inconsistent across studies [148, 149, 150,151,152, 162, 163]. Data interpretation, due to existing disparity across platforms and sample groups is difficult. Gene expression can be affected conceivably by a number of factors including upstream *cis*-acting motifs, epigenetic mechanisms, and experimental confounds. The changes in genes expression resulting from the pathology may simply not be present at the time of the death. Additionally, and perhaps more telling with respect to its usefulness in molecular psychiatry, it is difficult in gene expression studies to distinguish between gene expression changes resulting from primary pathology, or some compensatory mechanism [162, 163, 164].

Expression studies – MicroRNAs

Expression profiling of post mortem brain tissue suggest that aberrant miRNA may be linked to the etiology of schizophrenia/bipolar disorder. This type of study

represents the newest attempt in molecular psychiatry to study the molecular and genetic architecture of schizophrenia and bipolar disorder. Starting in 2007 there have been 11 separate studies that implicate miRNAs in the disease using a variety of platforms including qPCR and expression microarrays [7, 8, 27, 28, 48, 102, 103, 136,138, and more recently115]

Like expression profiling of protein coding genes, postmortem expression studies are thought of as a direct measure of activity within the brain. Specific brain regions harboring relevant affected functions, i.e. speech centers or centers of higher function have been targeted by these types of studies. These studies have been limited in scope, however, and produced limited results due to a heavy reliance on annotated miRNAs and commercial platforms. Like expression studies of protein coding genes, postmortem samples however are easily affected by confounders such as sample storage conditions, brain pH and lifetime medication used by subjects. In addition, the clinical and genetic heterogeneity underlying these early studies contributes heavily to their limitation. The studies have focused principally on the dorsolateral prefrontal cortex (Brodmann's area 9-46) and the superior temporal gyrus because of their role in working memory and social cognition. Additionally the numbers of subjects assessed in this manner have been low, varying from 13 to 105 as it has been traditionally difficult to gain samples from a large number of subjects. A total of 16 miRNAs have been identified with increased expression in these studies and 11 have miRNAs have associated with decreased expression [7, 8, 27, 28, 48, 102, 103, 136, 138]. MiRNAs which have been identified with increased expression include hsa-mir-105, -128a, -15a, -15b, -16, -17, -199*, -20a, -222, -34a, -452*, -486, -487a, -502, -652, -132, -212 and

hsa-mir-7[103]. miRNAs which have been identified with decreased expression include hsa-mir-106b, -151, -20b, -224, -30a, -30b, -30d, -30e, -383, -432, and hsa-mir-505[103].

In reviewing these studies, one can clearly see a large amount of variation in the miRNAs that are identified as dysregulated and in the directionality of their individual fold change. A meta-analysis of six studies yielded 44 miRNAs reported with any degree of consistency. Of these, only four (hsa-mir-212, -181b, -29c, and -7) were reported in more than two. The Pearson correlations (ρ) for these miRNA are depicted in the figure below. Of note, only hsa-mir-181b was reported in more than two studies with directional consistency. Though this meta-analysis was undoubtedly influenced by the way in which the expression values were identified, e.g. literature search, this nonetheless illustrates that miRNA expression studies, similarly to, genetic studies are subject to the same limitations, namely, sample size and tissue heterogeneity. At the very least, this small meta-analysis suggests that miRNA expression profiling may be as problematic as the early studies involving protein coding genes.

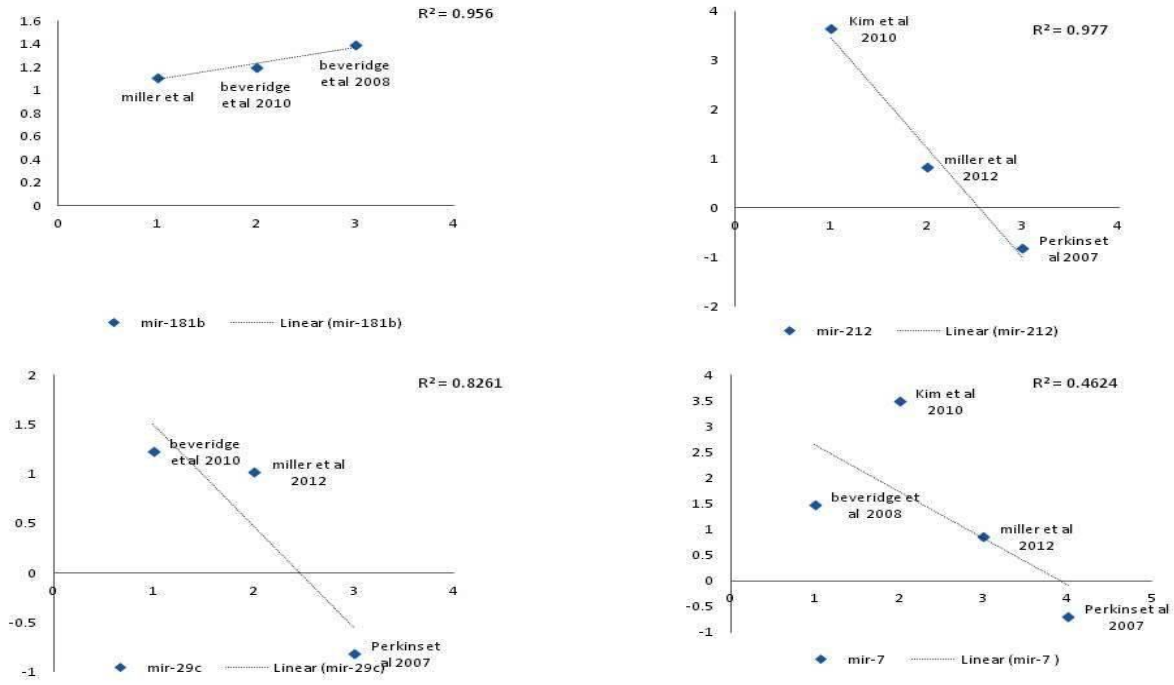


Figure 6 Comparisons of miRNAs and their reported levels was made across six expression profiling studies. A total of 44 miRNAs were reported in multiple studies (R = -.934). Only four of these hsa-mir-181b (3), hsa-mir-29c (3), hsa-mir-7 (4), and hsa-mir-212(3) were reported in more than two. The values for the respective studies are pictured in the above plots. Only the values reported for hsa-mir-181b was found to be in a consistent direction.

Expression profiling of peripheral tissues sources other than brain have been predicated on the notion that miRNA expression levels are reflective of a specific health state and as such that there should some overlap in the levels seen in the brain with that of blood. These studies undoubtedly contain greater variation as the sample sources is even farther removed and more easily affected by confounding. Blood samples are easier to obtain and could potentially be used to highlight miRNA expression differences between cases and controls as demonstrated by Lai et al (2011) [48]. In their study, seven miRNA signatures (hsa-mir-34a, -449a,

-564, -432, -548d, -572 and hsa-mir-652) were detected to correlate with negative symptoms, neurocognitive scores and event potentials. This signature was used by the researchers as a diagnostic indicator, generating an area under the curve (AUC) of 85% and receiver operating characteristics (ROC) of 95%. Blood is a heterogeneous suspension, comprised of multiple cell types and is potentially also affected by confounders such as drug exposure therefore some degree of caution should be exercised regarding these results.

From the studies reviewed above, one can see clearly that schizophrenia and bipolar disorder each possess an inherited component which, conceivably, is discernible through genetic and molecular studies. To date, researchers studying these diseases have made considerable progress in understanding the genetic architecture of both, thanks largely to the knowledge generated by the mapping of the human genome and the development of new methodology that allows a comparison of the phenotype with genetic loci. More work is needed, however, in order to explain the inter-subject variability seen in each disease and miRNAs through their control functions could service to address this problem. This field already has evidence originating from GWAS and expression studies implicating miRNAs in schizophrenia and bipolar disorder. By detecting novel miRNAs, studying their interaction with target genes and studying factors that might affect their function, we can better understand their role in schizophrenia and bipolar disorder. This thesis addresses these topics in a series of experiments using deep sequencing, qPCR, and SNP imputation.

Chapter 2. Detection of microRNAs through Next Generation Sequencing

Adapted From:

Williamson, V., Kim, A., Xie, B., McMichael, G., Gao, Y., and V. Vladimirov. (2012) Detecting miRNAs in deep-sequencing data: a software performance comparison and evaluation. Briefings in Bioinformatics, 13(1).

Williamson, V., Kim, A., McMichael, G. Bin, X., Parker, E., Gao, Y., and V. Vladimirov. (2012) Reanalysis of Six Deep Sequencing Data Sets Yields Potential Schizophrenia Related Novel MiRNA (in preparation).

Abstract

Next generation sequencing has become a preferred method for investigators interested in detecting novel miRNAs. Its depth of scope, flexibility and seemingly "agnostic approach" to data collection makes it an attractive option for those wishing to build a miRNA profile of a specific cell type or tissue. To that end, a deep sequencing experiment was performed on a commonly used cell line, neuroblastoma (ATCC: crl-2217) using the Illumina/Solexa. A total of 113 miRNAs were detected in the model cell line, 25 of which could be considered novel candidates. Based on estimates of the probable number of miRNAs found in wild-type samples, this number of miRNAs was low and prompted an assessment of the sample with additional software. A receiver operating characteristics curve (ROC) based on simulation data suggested that the software initially used in analysis, miRDeep, was in fact the most suitable for our purposes. In addition, a wide divergence in the numbers of miRNA predicted between the programs compared suggests that additional improvement is needed in the design of current software.

One of the novel miRNA predicted in the neuroblastoma data set, PRD5 was shown to be differentially expressed in schizophrenic patients from the Stanley Medical Research Institute (SMRI). Comparison of the expression values of this novel miRNA in postmortem brain tissue with the genes, zinc finger 804A (*ZNF804A)* and chromosome 10 open reading frame 26 (*C10orf26*) suggests that an interaction may be occurring. Both of these genes have been implicated in schizophrenia/bipolar disorder GWAS. In addition to the novel miRNAs that were detected in this study, five novel candidates were also identified that result from alternative pathways of biogenesis. These miRNAs,

derived from small nucleolar RNAs, were validated in a RNA tissue panel of 20 normal human tissues. The identification of these miRNAs suggests that current theories regarding the biogenic source of miRNAs should be reconsidered. In particular, it demonstrates that reads deemed unusable by virtue of their mapping location should not be discarded, but rather reassessed using different criteria.

Introduction

MicroRNA Biogenesis and Function

MicroRNAs (miRNAs) are short non-coding RNA sequences which measure approximately 18-22 bases in length. Classical miRNA biogenesis is, briefly, a three-step process that starts in the cell's nucleus and ends with the creation of the mature sequence within the cytoplasm [9, 47, 51, 53, 90]. Each step in the process results in a definable product which can be then predicted computationally and validated through Northern Blot or PCR based analyses. The primary miRNA (pri-miRNA) is a double-stranded structure that measures over 1 kilobase (kb) in length and has a guanine cap and poly-adenylated tail. In the first stage of miRNA biogenesis, the pri-miRNA is cleaved by Drosha and its partner DiGeorge critical region 8 (*DGCR8*) to form the precursor strand (pre-miRNA). The second stage in miRNA biogenesis occurs when the pre-miRNA is exported from the nucleus to the cytoplasm by exportin 5 (*EXP5*). Once in the cytoplasm, the precursor (pre-miRNA) is processed by Dicer and its partner TAR RNA-binding protein (TRBP). Dicer cleaves the arms of the pre-miRNA approximately 22 bases from the site of the Drosha cleavage, creating a double-stranded duplex containing the mature sequence and the star sequence (miRNA:miRNA*).

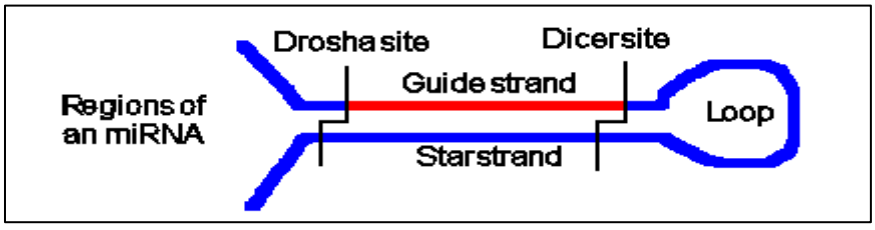


Figure 2 The guide or mature sequence is found in the precursor opposite to the star. It is the mature sequence that survives degradation and directs the gene targeting. Image adapted from http://www.gene-tools.com/TargetSites.gif.

Thermodynamic studies have suggested that the strand with the least stable base pairs (the mature sequence) at the 5' end survive degradation [53,54]. The strand opposite to the mature strand, the star sequence, is generally moved for degradation (see above figure for relative positions of the star and mature sequence within the precursor) [217]. In general, prediction algorithms focus on the second and third stage of miRNA biogenesis to assess the presence of novel miRNAs.

Currently, the main repository for the annotation of miRNAs is the website miRBase. The current release of miRBase (v18) holds over 18226 miRNAs from species as diverse as zebrafish and chicken as well as *arapidopsis*. For humans, miRBase (v18) holds 1527 precursors and 1921 mature sequences [53]. The rules for determining whether a miRNA is a true candidate are: 1) whether its predicted precursor sequence folds into a viable hairpin, 2) whether the mature sequence can be detected in a size fractionated sample and 3) whether the candidate sequence falls within one of the hairpin's respective legs or stems [93].

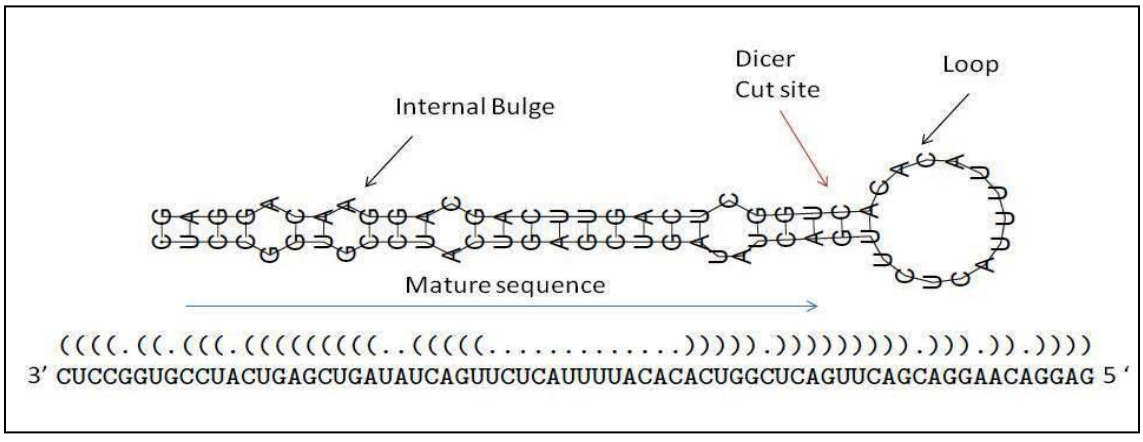


Figure 7 Classic stem-loop structure/hairpin generated by RNAfold. This program is used in many pipelines to assess the minimum free energy of candidate hairpins. Pictured is hsa-miR-24.

There is now growing evidence, stemming largely from NGS studies that miRNA-like sequences can result from non-coding RNA sources such as small nucleolar RNA (snoRNA) and transfer RNA (tRNA) [11-19,23]. In particular, snoRNAs are a family of conserved nucleolar RNA encoded in the introns of protein coding sequences that are approximately 200 bases in length. There are two officially recognized classes of snoRNA (Box C/D, Box H/ACA) which work in conjunction with other proteins in complexes called SNRPs to control 2-O-ribose methylation and pseudouridylation respectively. SnoRNAs target principally ribosomal RNA, transfer RNA and small nuclear RNA; classes are distinguished on the presence of key sequence motifs and in their interactions with molecules such as dyskerin and fibrillarin. H/ACA snoRNA are distinguished by the H motif box (consensus ANANNA; N = purine or pyrimidine) and the ACA (ACA) and C/D boxes snoRNAs are distinguished by conserved C (UGAUGA) and D (CUGA) motifs, respectively. Recently, a third (Orphan) and fourth (composite) class of snoRNAs have been suggested [73]. The function of these two classes is currently unknown but they possess the same structural motifs as the Box H/ACA and Box C/D snoRNA.

One characteristic that sets snoRNAs apart from that of miRNAs is the cellular location at which the molecule is thought to be functional. The biogenic pathways of both molecules, however, are extremely similar (figure below) with the exception of the participation of the enzyme Drosha. Drosha does not participate in the biogenesis of snoRNAs whereas it figures heavily into that of miRNAs. Several groups have found miRNAs that originate from and overlap with larger snoRNA molecules [11-19, 23]. One

group, in particular, discovered a miRNA in a HITS-CLIP sequencing experiment which effectively reduced mRNA expression of the gene cyclin-dependent kinase 19 (CDK19) by 20% [13]. HITS-CLIP sequencing, the sequencing of RNA isolated by cross-linking immunoprecipitation, is widely used form of NGS for the mapping of protein-RNA binding sites in vivo [224].

A subgroup of orphan snoRNAs are expressed exclusively in the amygdala, hippocampus, and nucleus accumbens, and in animal studies is thought to affect contextual fear conditioning and brain function [23]. Further, one cluster of this subclass, HBII-52, was also shown to regulate alternative splicing in serotonin receptor 2c and implicated in the etiology of the neurodevelopmental Prader Willi Syndrome [104].

The fact that these brain expressed snoRNAs do not function as traditional snoRNAs would, suggests that other potential regulatory functions for the molecules exist. It is conceivable, then, that orphan snoRNAs could act as an additional source for miRNAs potentially relevant to brain function and that, as such, this biogenic pathway needs to be investigated.

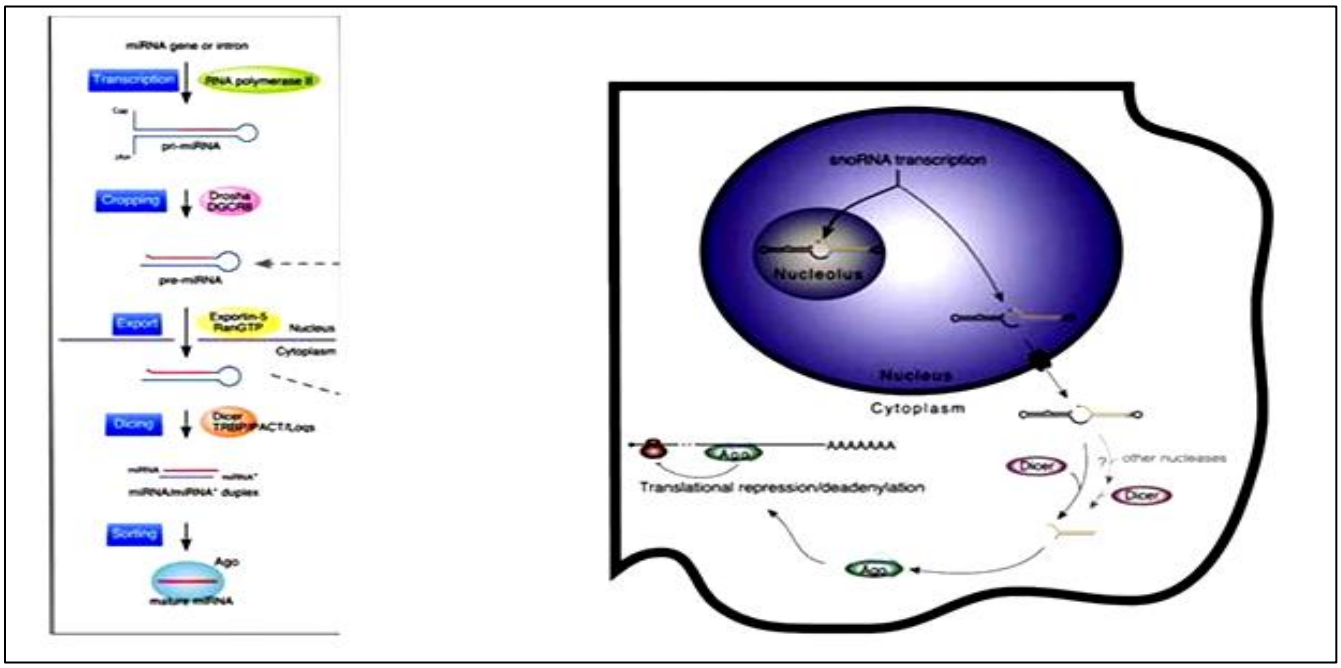


Figure 8 A comparison of the classical biogenic pathways in miRNAs and snoRNAs. Distinct similarities exist both in the enzymes used in the process and the locations within the cell where these activities occur. Pictured in the above diagram left is the miRNA biogenic pathway and right is the snoRNA biogenic pathway. Images are adapted from Miyoshi et al, 2010 and www. cipsm.com.

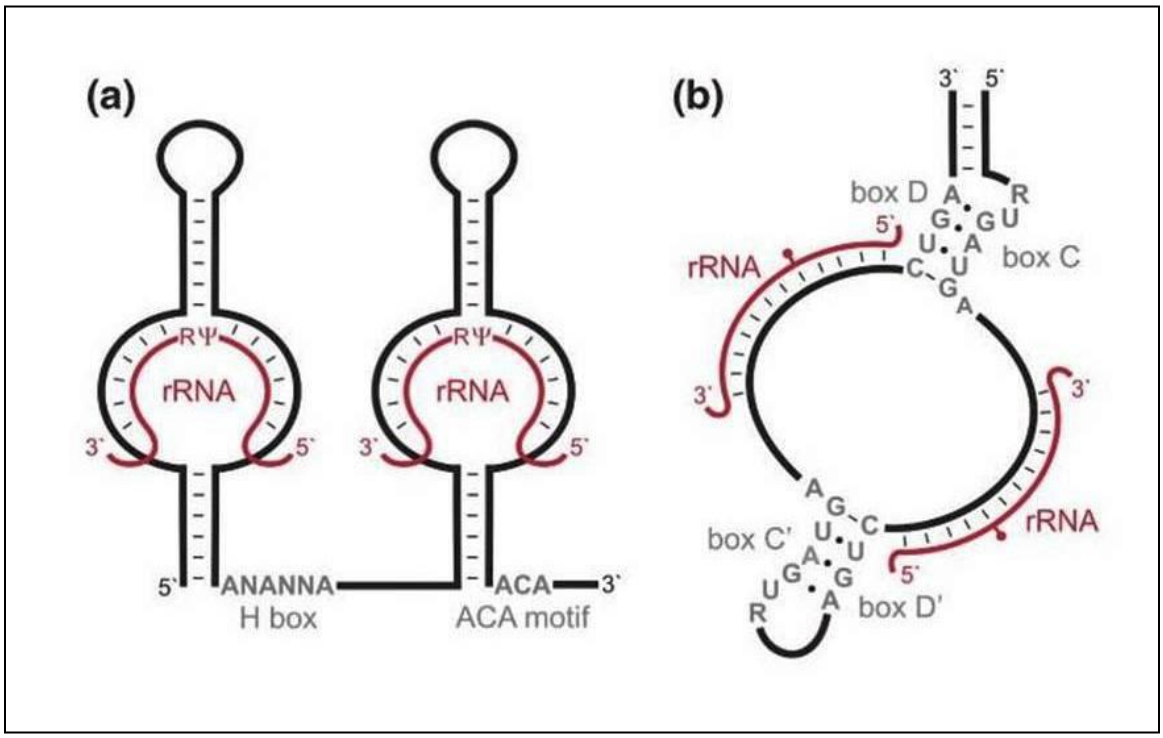


Figure 9 H/ACA (A) and C/D Box (B) structure. The secondary structure of these molecules suggests that a portion could function as a hairpin and from that yield a functional miRNA like fragment. Image taken from www-snorna.biotoul.fr

Most miRNAs regulate gene function negatively through imperfect binding with the 3' untranslated region (UTR) [51, 52, 54, 55]. Animal miRNAs pair with 3'UTR of their target genes though the "seed" region (consisting of nucleotides 2-7) at the 5' end of the mature strand. Depending on the percent match of the interaction, the miRNA can affect either the degradation of a mRNA target or its translational inhibition [90]. It has been estimated that miRNAs may influence as much as 30% of the human transcriptome [47]. Target validation for miRNAs has been slow at best and outpaced by discovery techniques. One study, using *Drosophilia* as a model organism, estimated that as high as 60% of annotated miRNAs in miRBase lack clearly defined experimentally validated targets [51, 54].

Regardless, miRNAs have been implicated in a number of human diseases including Cancer, Fragile X syndrome, and coronary heart disease [191, 207, 208]. Currently, the largest field of study has been Cancer research where great strides have been made identifying miRNA affecting oncogenes involved in throat and gastric cancer [209, 210].

Deep sequencing experimental protocol

There are a number of second generation platforms routinely used today for the sequencing of miRNA, e.g. Illumina, Roche 454 and ABI Solid. The basic protocol is the same for sample isolation but varies considerably for library preparation and actual sequencing. In sample preparation, the small RNA fraction is extracted either using phenol/chloroform or a silica-based column method and is eluted using ethanol precipitation. The sample then is fractionated and adapters are attached which are platform specific sequences. Small RNA has been shown to be more stable than mRNA and has been isolated successfully in a variety of tissue and cell types [61]. In general, deep sequencing represents improvement over earlier Sanger/capillary based techniques as there is no amplification through bacterial cloning. MiRNA comprise approximately 0.01% of the total RNA fraction and a recent comparison between methods used to isolate miRNA show that enrichment may be necessary for successful detection [61]. Typical endogenous expression of miRNA within the cell falls within 500 copies, though some cell-type specific miRNA have been demonstrated to be expressed at much higher levels [56].

There are several limitations to MicroSeq technology that must be addressed and many of these introduce biases that must be overcome before the results from

individual studies are compared. First, PCR based amplification creates sequencing errors primarily at the 3' end of the read; it has been estimated that errors due to either thermal stress on the enzyme or through editing occurs a rate of 7.2 X 10^{-6} base pairs [96]. A second issue is the limited comparability of expression values across samples due to the methodology regarding the normalization of read levels. Third, unequal representation of fragments in the library preparation may also occur due to naturally occurring imbalances in expression levels. Currently, no set statistical approach exists that allows researchers to easily compensate for biases induced during PCR amplification, library creation, or sample preparation although a few significant advances have been recently made [20, 21, 65].

Because of the acknowledged issues with this type of dataset, the relative expression levels of any miRNA identified in MicroSeq data should be approached with caution and additional validation should be performed with more established methods such as quantitative (qPCR). Several studies show the relative magnitude of MicroSeq data to correlate well ($r^2 = 0.8$) with qPCR [111,112]. Despite the issues raised above MicroSeq datasets can still be successfully used to determine presence of novel or known miRNAs within a particular sample.

Computational prediction of miRNAs and their targets

Recently, there has been an increase in the number of programs written toward the prediction of miRNAs from deep sequencing data. The manner in which this data is evaluated with respect to prediction is three fold: 1) reads are mapped to a reference genome 2) the loci to which the read is mapped is expanded computationally and folded to determine secondary structure, and 3) the secondary structure is evaluated to

determine whether it is a miRNA hairpin. Programs differ on the basis of whether they are comparative or non-comparative in nature. Comparative programs assess whether the proposed candidate is phylogenetically conserved and non-comparative programs focus more heavily on species-specific examples. The number of species used in non-comparative programs is largely up to the user and non-comparative programs are thought to more useful in the detection of newly evolved novel miRNAs.

The program, miRDeep, is a non-comparative program that predicts the presence of miRNA from deep sequencing data using conditional probabilities [83,84]. miRDeep employs a flexible format, accommodating data generated by a 454 Life Sciences/Roche or an Illumina/Solexa sequencer. Using the steps of classic biogenesis as a guide, the pipeline first compares the reads to a target genome, and then evaluates the read's suitability on a thermodynamic scale. The algorithm assumes that if a read is related to miRNA, then it must either be a portion of a star, a loop sequence or a mature sequence. The read must demonstrate characteristics similar to already annotated examples, e.g. definite evidence of a present 2nt 3' overhang. Additional characteristics include the minimum free energy of the predicted precursor strand and in miRDeep2 the level of demonstrated homology to a species closely related to the target genome. Also, miRDeep makes the assumption that because mature sequences tend to be more abundant in the cell than any other miRNA related sequence, reads which conform structurally to “mature sequences” will likewise be the most abundant in the data file.

The software generates a logarithm of the odds (LOD) score which is based in part on read frequency. If a read meets structural criteria for being considered a mature sequence and is found to be frequently represented in the data file it receives a higher

score than those that are less frequently found. Structural stability of the predicted precursor, as well as conservation of the 5’ end of the mature sequence is also factored into the LOD score through parameter fitting. Because miRDeep is a pipeline, it allows the user to choose the mapping tool and software for free energy evaluation. In this project, we used the program Oligomap to map the reads rather than Blast and RNAfold to evaluate free energy [87]. It is generally acknowledged that Blast is poorly suited to the process of mapping deep sequencing reads; Oligomap with its heuristic approach to short read mapping greatly speeds up the process.

miRanalyzer is a web-based tool based on a random forest classifier and trained on experimental data [66]. A benefit of web-based applications such as miRanalyzer is that they allow the user to analyze their results without having access to a large amount of computer resources. The first version of the software targeted seven model species (human, mouse, rat, fruit-fly, round worm, zebrafish and dog*).* Newer versions have since incorporated plant genomes and predictions based on plant models. miRanalyzer uses the program Bowtie to map the reads to the target genome [67]. Apart from specifying the number of allowable mismatches, and the acceptable p level for a credible prediction, the user is restricted, however, from employing any other changes.

The current version of the deep sequencing small RNA analysis pipeline (DSAP) differs from miRDeep or miRanalyzer in that its algorithm focuses more strongly on miRNA expression rather than prediction of known and novel forms [85]. By employing a technique where reads are clustered into unique groups and then mapped onto existing RFAM and miRNA databases (ergo only known miRNAs can be detected), the program circumvents the need for annotated genome required by the other software. In addition,

DSAP provides the user with superior processing speeds, e.g. 2 million sequences were bench marked at less than 15 minutes [85]. This program uses Supermatcher from the EMBOSS tool kit which is a combination of word match and the Smith-Waterman dynamic programming algorithm [218, 219]. The speed of DSAP is partly due to the use of Supermatcher which is designed to perform local pairwise alignments between a single sequence – typically a large one – and that of a database.

Software	Format	File format	Location
Seqbuster	Web based, executable	fasta, tab-delimited	http://davinci.crg.es/estivillLab/seqbuster/
miRExpress	Executable	sequence tag count	http://mirexpress.mbc.nctu.edu.tw/
miRNAKey	Executable	fasta, fastq	http://ibis.tau.ac.il/miRNAkey/
MirTools	Web based	sequence tag count	http://59.79.168.90/mirtools
miReNA	Executable	fasta	http://www.ihes.fr/~carbone/data8
miRTrap	Executable	fasta	http://davinci.crg.es

Table 2 Other software used in the analysis of deep sequencing data. MiRDeep, miRanalyzer represent two of the most popular open-source software used for the analysis of deep-sequencing data.

The relative strengths and weaknesses of these programs have largely been unexamined as the field has been focused on the development of viable algorithms that do not place undue stress on the computational resources of a typical laboratory interested in deep sequencing. To that end, several developers have chosen to create web-based applications that moves read mapping steps to a location offsite from the user. The above table lists several software that have been created in this vein. Two of these programs (Seqbuster, MiRTools) are web-based applications that allow the user to analyze their data off site. Our comparison of miRDeep, miRDeep2, miRanalyzer and DSAP represent one of the first attempts to compare programs in terms of their

respective suitability to the detection of novel miRNA.

Algorithmic approaches for target prediction

Successful prediction of targets for novel miRNAs involves careful consideration of homology, Watson-Crick binding, and minimum free energy (MFE). In addition, several authors have suggested that site accessibility and co-expression be considered as factors [51, 54]. Briefly, the mature sequence pairs with the gene's 3'UTR through Watson-Crick binding in three ways:1) seed only, 2) seed plus additional bases on the 5'end of the mature sequence and 3) additional bases binding on the 3' end of the mature sequence. The seed is defined as bases 2-8 on the 5' end of the mature sequence and many believe that seed binding is all that is required for a gene to be functionally affected [51,55].

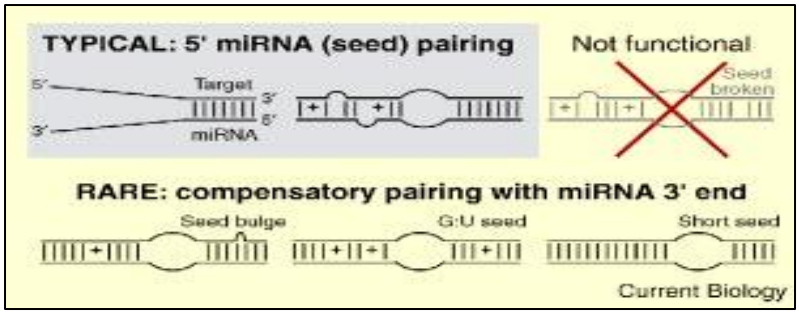


Figure 10 Ways in which the mature sequence may bind with a target gene. The seed (bases 2-8) on the 5' end of the mature sequence has been demonstrated to be instrumental in whether a miRNA targets a gene. Image adapted from Lai, Current Biology, 2005

Different programs vary with respect to the emphasis given to each respective characteristic and in the manner that the 3' UTR is defined [55]. Of the various software, miRanda, one of the oldest, is generally viewed to be the most sensitive [51]. It creates a threshold score based on homology, Watson-Crick binding, and MFE to rank its target

predictions. As the most sensitive, however, miRanda is apt to contain the highest percentage of false positives and should be screened carefully. PITA, by contrast, considers site accessibility and co-expression as key factors in target determination [47]. The $_{\Delta\Delta}$G score generated by PITA takes into account how strongly the miRNA binds to a proposed target with that target folded into its probable secondary structure. The combination of miRanda and PITA allows us to be more restrictive in the gene targets selection. While overlap is certainly desired when comparing predictions of these programs, it is small at best, estimated at ~11% [47]. A small overlap percentage, however, is desirable as it will ensure that the generated predictions are robust and less likely to be a result of type I error.

Materials and Methods

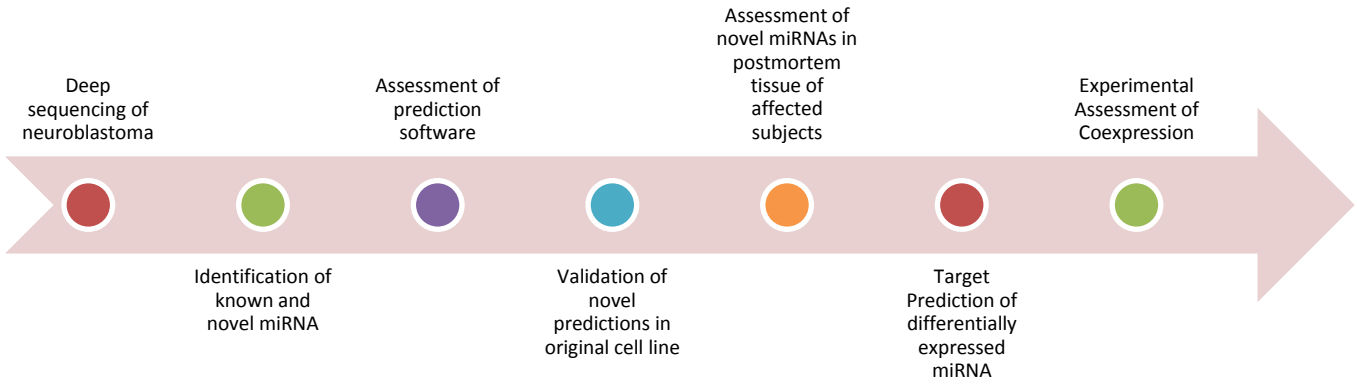


Figure 11 Steps taken in chapter two. Read processing includes filtering for contamination and other non-miRNA specimens as well as the adapter sequence utilized on the reaction.

Neuroblastoma deep sequencing

Total RNA containing the small RNA fraction was isolated from pelleted neuroblastoma cell lines (ATCC: crl-2217) using the mirVana Paris kit (Ambion) following manufacturer's specifications. After isolation and ethanol precipitation (95% Ethanol/Sodium Acetate, pH. 4.5), the RNA concentration for the neuroblastoma cell line was 4µg/µl. Library preparation, PCR reaction, and deposition were prepared according to standard protocol and slides were analyzed on an Illumina/Solexa system. The sample was sequenced at a concentration of 1 µg/µl. Post sequencing, a shell script was used to convert fastq files into two fasta files, with one containing the reads and another containing quality scores. The reads in the fasta file were then cleaned of the tag sequence and homopolymers longer than 4 nucleotides (nt) and size sorted (>18nt).

Sample contamination was determined by the percentage of reads which mapped to the *E. coli* and *M. musculus* genomes. These two organisms are common sources of contamination due to their popularity as lab models. In the neuroblastoma cell line, individual base quality diminished considerably at the 3’ end of the read sequence (e.g. enzyme depletion), similar to what has already been shown [99]. This finding was compensated prior to final processing by trimming five bases from the 3’ end. Sample contamination overall was negligible (<0.1%) eliminating only a small number of reads. Final read processing was performed by a perl script that removed redundant sequences, formatting the file into a series of reads and read counts. A total of 6,904,317 reads were analyzed from this data set to yield 1,214,402 unique reads after cleaning.

Stage	*Count*
Total Reads	*6904317*
Unique Reads following cleaning	*1214402*
Number of Times reads mapped to human genome	*16097885*
Number of rimes reads mapped 5 or fewer times	*19569*
Number of verified miRNA	*88*
Number of unknown miRNA	*25*

Table 3 Read number after each subsequent step of processing. Reads are eliminated by miRDeep that map more than 5 times to the genome because they are assumed to be degradation products of other more highly expressed RNA. Multiple hits occur in part due to the physical size of the trimmed read (18 bases). Smaller reads in general have a higher probability of mapping in multiple regions.

Validation of miRNA predictions in Neuroblastoma and a RNA tissue panel

All novel miRNA predictions generated by miRDeep, were verified in the original cell line using qPCR. Briefly, RNA was isolated from wild-type neuroblastoma cells (concentration 10ng/ul) and assessed in triplicate on a HT Fast 7900 (ABI) using Taqman Assay. Neuroblastoma cells (ATCC: crl-2217) were grown at 37˚C and 5% CO_2 in a 1:1 mixture of Eagle's Minimum Essential Medium, F12 Medium, supplemented with 10% bovine serum (FBS) and 1% non-essential amino acids. The average RNA integrity number (RIN) value for all samples tested was 8.2. The protocol for qPCR was a two step process including cDNA synthesis and Realtime qPCR and can be found in a number of papers published by our lab [8, 26]. On average the standard deviation (SD) between replicates was less than 0.03. The average Cq values were normalized using

$_{\Delta\Delta}$Ct method against an endogenous reference snoRNA RNU44.

Separately, all novel miRNA candidates derived from snoRNA were validated in a RNA panel of 20 normal Human tissues (First Choice, Life Technologies). Sample validation was performed in the same manner as described above with the exception of the source RNA originating from a commercial platform rather than a cell line. The purpose behind using a tissue panel rather than a cell line in this instance was to determine whether the novel sno-derived miRNA were differentially expressed across a wide range of normal human tissues. Differential tissue expression has been viewed as a classic indicator of miRNA status in traditionally derived miRNAs [15], therefore it was used here as a way of affirming candidate status. A graph showing the validation results of the sno-derived miRNA can be found in section entitled “Detection of novel miRNAs from alternative biogenic sources”.

Validation in Postmortem Brain Tissue from the SMRI

200 mg of postmortem brain tissue, originating from the dorsolateral prefrontal cortex (Brodman’s Area 46) were received from the Stanley Medical Research Institute (SMRI). This sample was used to: 1) determine if any identified novel miRNA was also expressed in brain tissue and 2) determine if any novel miRNA was significantly differentially expressed between cases and controls. Exclusion criteria for subjects included: 1) brain pathology, 2) central nervous system disease, 3) poor RNA quality, 4) IQ < 70, 5) age <30 years, and 6) substance abuse within one year of death. Total descriptive parameters for this study group can be found in the table below. Total RNA was isolated from approximately 100mg of this tissue using the MirVana-Paris Kit (Ambion, Texas). The RNA integrity number (RIN) was determined using nano chip

(Agilent) on the 2100 Bioanalyzer (Agilent, California) and was 7.2. The protocol for cDNA synthesis was performed according to manufacturer's recommendations and can be found in papers published by our lab [8, 26]. All novel miRNA candidates were validated in triplicate in this sample and normalized using $_{\Delta\Delta}$Ct method against the snoRNA RNU44

Profile	**Diagnosis**
Age	years at death
Sex	*Male, female*
Race	Caucasian, African American
DOD	date of death
Refrigerator Interval	from estimated time of death to refrigeration of body at ME's office (hours)
Suicide Status	death by suicide
PMI	post-mortem interval (hours)
RIN	RNA Integrity Number
Brain PH	acidity-alkalinity (log scale; 7=neutral)
Left Brain	*Side of brain*
Right Brain	*(see above)*
Brain Weight	Relative mass of brain (g)
Age Of Onset	age of first symptoms (years)
Duration Of Illness	age at death minus age of onset (years)
Time In Hospital	total for all psychiatric hospitalizations (years)
Lifetime Alcohol Use	*Years*
Lifetime Drug Use	*Years*
Smoking At TOD	if person had smoked previously but had quit in the past, this was coded "no."
Psychotic Feature	*Diagnosis*
Lifetime Antipsychotics	fluphenazine equivalents (mg)

Table 4 Descriptive parameters for the Stanley Medical Research Institute sample set

Performance Comparison of Prediction Software

Using three different programs, miRDeep (v1, 2), miRanalyzer and DSAP, we analyzed seven data sets (GSE494809, GSE494810, GSE494811, GSE494812, GSE715665, our neuroblastoma dataset, and a simulated dataset) to provide a critical evaluation of program performance [26]. Initial miRNA predictions by miRDeep were

thought to be too low and we hoped to receive confirmation by comparing them to those generated by miRanalyzer and DSAP. Additionally, we felt justified that the relative operator curves for each program generated in the simulated dataset was sufficient to determine which program might be best suited to the detection of novel miRNA candidates (figure 12). The biological data sets used in this comparison include miRNA profiles drawn on peripheral mononuclear blood cells, HL60 cells, K562 cells, breast cancer cells and neuroblastoma cells and were downloaded from Geo Omnibus [210]. In addition, a simulation data set was created using the program Flux Simulator and included 100 known miRNA that were 'spiked in' to the simulation at a prevalence of 0.1% to provide a basis against which ROC curves could be drawn. The parameters shown below in table 4 were chosen to mimic the characteristics of the neuroblastoma dataset though characteristics derived from comparisons of publicly available data demonstrated considerable variability in terms of size, GC median content and GC standard deviation.

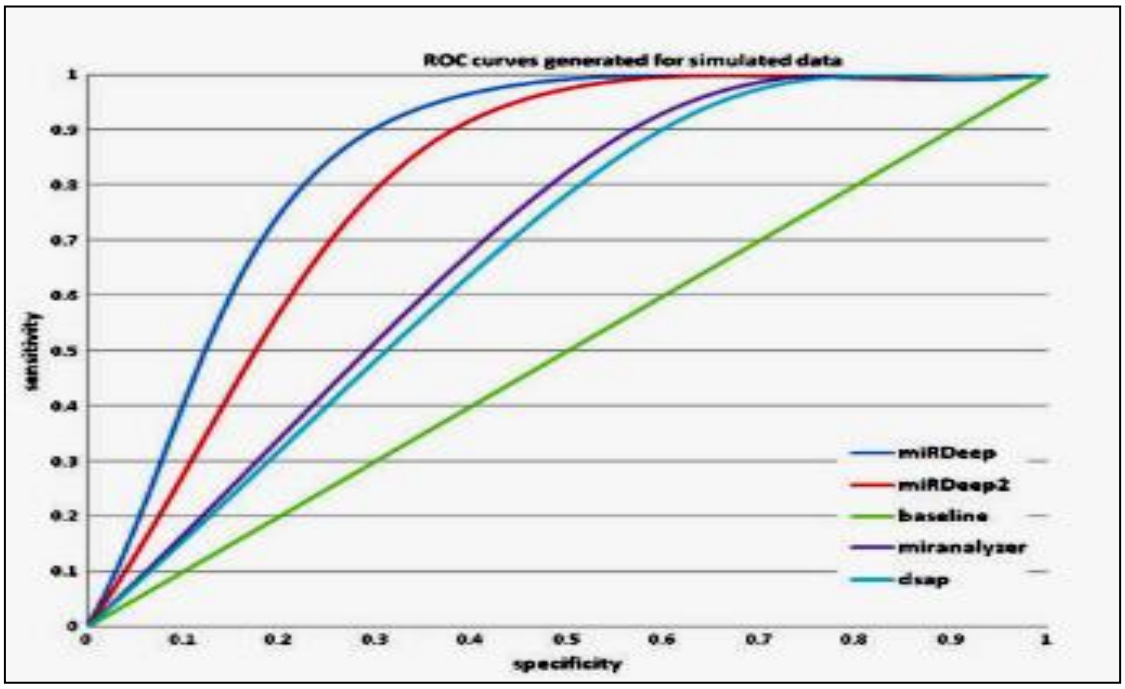


Figure 12 ROC curve were created using simulated data generated by Flux Simulator.

The ROC curves suggest that miRanalyzer and DSAP is less specific with regard to miRNA detection than miRDeep. The area under the curve (AUC) for miRDeep, miRDeep2, miRanalyzer, and DSAP was 0.94, 0.89, 0.72, and 0.70 respectively. In addition, in figure 13 a greater number of novel miRNA candidates were reported by miRanalyzer than either miRDeep or miRDeep2. miRanalyzer appears to detect a greater number of lowly expressed novel miRNAs, e.g. based on fewer unique reads and an examination of the normalized expression levels confirms this finding. Judging from the number of unique reads, these lowly expressed miRNAs detected by miRanalyzer may in fact be false positives but will be difficult to validate conclusively without costly additional study. The program comparisons suggest that, despite differing stringency levels, after adjustment for total number, they all identify a similar set of known and novel predictions. Different stringency levels are likely, however, to affect the

number of possible novel candidates for functional verification. Stringency levels may also play a key factor if one is interested in working with miRNAs that are lowly expressed or have recently arisen in an organism's course of evolution.

READ.LENGTH	35
TSS.MEAN	25
READ.NUMBER	5 000 000
NB.MOLECULES	5 000 000
GC.SD	0.1
GC.MEAN	0.5
SIZE.SAMPLING	AC
FRAG.SUBSTRATE	RNA
FRAG.METHOD	UR
FRAG.EZ.MOTIF	NlaIII
PAIRED.END	FALSE

Table 5 Parameters employed in creating simulation data

Results

Using the program miRDeep, we detected a total of 113 miRNAs (known and novel) in the neuroblastoma cells. Eighty-eight of these miRNAs were known and validated through blast (default parameters) to miRBase and 25 were considered novel. In comparing these predictions to those of miRanalyzer, we derived a list of 17 miRNA which overlapped between the two programs. 12 of these 17 miRNAs were validated successfully using qPCR with Cq values ranging from 19.10 to 36.10 (table 5). The higher value range (>30) for some of the novel miRNAs may suggest these to be false positives, however, we think they simply might be very low expressed, since some of them, originally thought by us as novel, have already been reported in miRBase (table 5). The reader should also note, looking at table 5, the apparent inability to validate precursor predictions. Despite testing precursor predictions from both miRDeep and

miRanalyzer, only 11% (2 out of 12) of the precursor predictions were successfully validated with Taqman. In contrast, seventy percent of the consensus mature predictions were successfully validated. The difference in the validation success for these two features may indicate either 1) an algorithmic flaw in the way the precursor is predicted in this software or 2) an overall flaw in the way the experimental procedures are performed initially. The precursors predicted by miRDeep and miRanalyzer when aligned were in many cases discontinuous representations of each other, and this variability undoubtedly impacted attempts at validation.

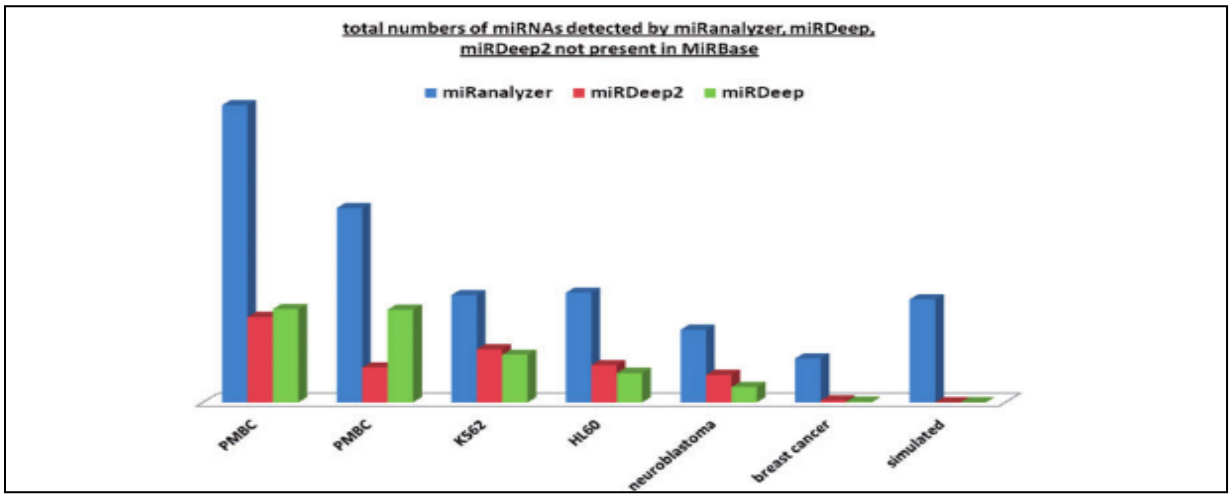


Figure 13 Novel miRNAs predicted in the neuroblastoma cell lines and six additional datasets. Predictions were generated by miRDeep, miRDeep2 and miRanalyzer

Sample	average_CQ	average_ CQ_precursor	annotated as (MirBase)
prd.mat-1	28.93	no amplification	
prd.mat-2	35.00	no amplification	Hsa-mir-3660
prd.mat-3	36.03	no amplification	Hsa-mir-4428
prd.mat-5	29.89	no amplification	
prd.mat-6	19.10	no amplification	
prd.mat-7	26.91	31.53	
prd.mat-8	34.10	no amplification	
prd.mat-11	32.25	no amplification	Hsa-mir-3131
prd.mat-13	35.45	no amplification	Hsa-mir-4421
prd.mat-14	32.52	34.11	
prd.mat-16	31.65	no amplification	Hsa-mir-2110
prd.mat-17	36.10	no amplification	Hsa-mir-4222

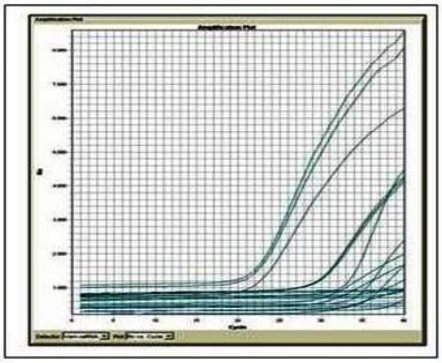


Table 6 Novel miRNA predictions that were validated in the original neuroblastoma cell line. On the right are amplification curves generated by the 7900HT.

Experimental Verification of miRNA presence in post mortem tissue

All novel miRNAs validated in the neuroblastoma cell line were also tested a second time in post mortem brain samples from SMRI to determine if any were significantly associated with schizophrenia and bipolar disorder. All novel miRNAs tested were validated successfully in postmortem brain tissue with Cq values averaging ~26. One miRNA candidate, PRD5, was shown to be significantly differentially expressed between schizophrenic subjects and healthy controls ($b_{(i)}$= -0.078; t=-2.3; p=0.025), but no differential expression was observed between bipolar subjects and controls. Due to presence of heavy outliers (figure 5) the disease effects on PRD5 expression were estimated within the robust multiple regression model (using the Huber's method) adjusting for potential confounding effects such as drug, lifetime anti-psychotics, PMI, RI and RIN. A box plot comparing the three diagnostic groups is pictured in figure 14. This initial finding prompted us to pursue PRD5 further, both bioinformatically and experimentally.

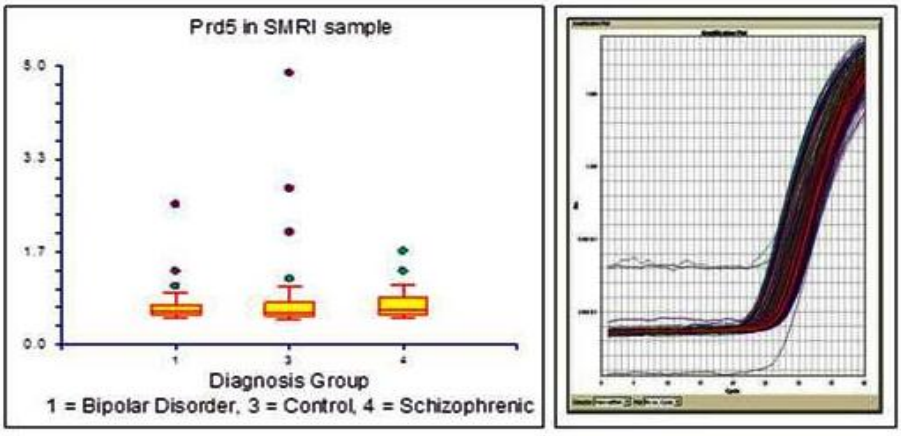


Figure 14 On the left, box plots of the three diagnostic groups. Outliers are marked as green (SD≥2) and red (SD≥3) symbols. On the right are the amplification curves generated by the 7900 HT.

Bioinformatic Analysis of PRD5 and its predicted targets

PRD5 is located in an intergenic region of chromosome 17, nearest to a cluster of homeobox (HOX) genes. HOX genes are responsible for basic body structure and regulation; of that cluster, HOXB9 has been shown to bind to B-cell translocation gene 2 (BTG2) a transcription co-regulator involved in neurite formation. Reanalysis of the datasets used in the software comparison (see previous section a discussion of the software used) determined that PRD5 was found in HL60 and K562 cells. Detection in multiple datasets along with our laboratory validation lends additional evidential support to the status of PRD5. In addition, structurally, this miRNA satisfies criteria drawn up by Sewer et al (2005), including minimum free energy, symmetric secondary structure and size of the terminal loop [88].

Target prediction

The precise prediction of miRNA gene targets is important as it provides information on the biological processes going on in the cell and allows for additional experimental study of gene candidates. The software used in target prediction have increasingly incorporated both functional (e.g. biologically based parameters) and logistic (e.g. consensus approach across multiple programs) to compensate for a high false positive rate [151]. Although, after following these approaches the false positive rates are substantially minimized [151], the remaining target predictions nevertheless still must be viewed with caution.

Thus, the PRD5 predictions were filtered with the above described functional and logistic approached in an attempt to derive viable gene candidate set for further follow

up.

Threshold screening using FDR based approaches

In one last attempt to control for type I error occurring during gene target prediction, a false discovery rate (FDR) was used. 1499 predictions generated by miRanda, TargetScan, and PITA were assessed using the Benjamini-Hochberg method. In the figure below, one should note that a high number of genes (769 < q 0.05) remained after multiple correction testing. The fact that nearly 50% of the predicted targets did not survive multiple correction testing is illustrative of the acknowledged false positive rate that is inherent in miRNA target prediction. Among the genes surviving correction for multiple testing include *ZNF804A*, *C10orf26*, and *TCF4*.

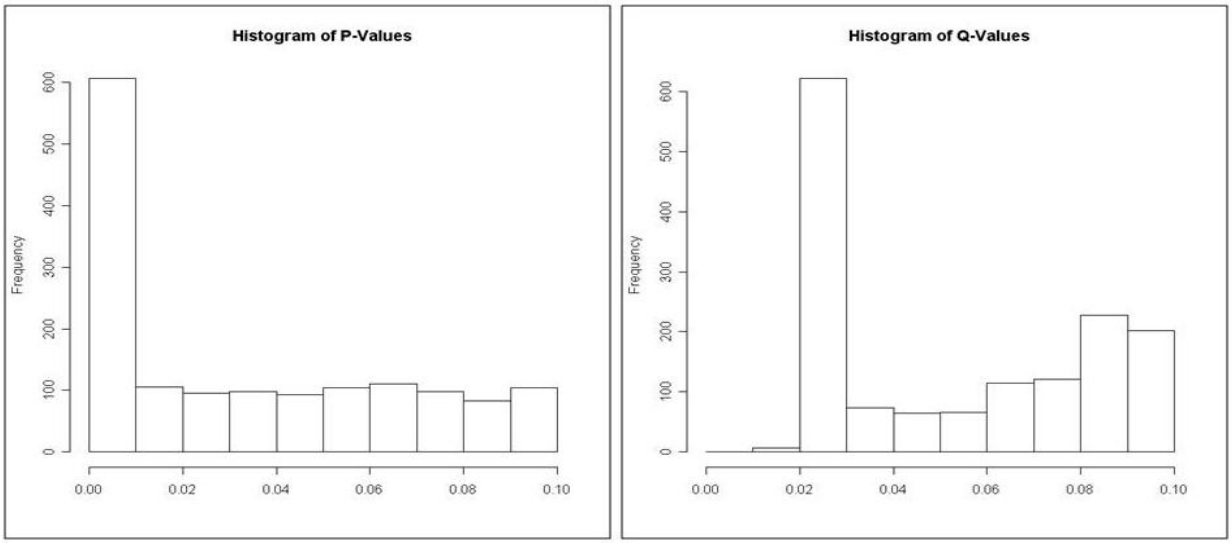


Figure 15 A false discovery rate (FDR:Benjamini Hochberg) was performed on the p values generated by the program miRanda.

Threshold screening using biologically based approaches

Next, a consensus approach was taken to filter predicted targets from miRanda, PITA (v7) and TargetScanCustom [91]. The union of these programs effectively combines an algorithmic emphasis on: 1) species-specific conservation (miRanda), 2) seed conservation (TargetScanCustom) and 3) site accessibility (PITA). In this analysis, the targets that were identified in all three programs include transcription factor 4 (*TCF4*), calcium channel, voltage-dependent , L type, alpha 1C subunits (*CACNA1c*), glutamate receptor, ionotropic, AMPA1 (*GRIA1*) and zinc finger 804A (*ZNF804*). These top targets showed consistency of the binding site across all known isoforms and high native expression. It also should be noted that *TCF4*, *ZNF804A* and *C10orf26* are three genes that survived correction for multiple testing in the previous section. All of these genes have been cited as significant in schizophrenia/bipolar disorder related GWAS studies [52].

Chi square assessment of target enrichment

To determine whether the potential targets of PRD5 generated by miRanda, TargetScan, and PITA have any bearing on schizophrenia, a chi- square test was performed comparing these targets with unrelated gene sets predicted for Multiple Sclerosis (MS), Parkinson's disease (PD), and Cancer. Genes were initially chosen from annotated databases (SZgene.org, PDgene.org, MSgene.org and cgap.nci.nih.gov) matching for the average relative size of the 3’UTR sequences of targets predicted by PRD5. These databases are comprised of gene association studies as well as meta-analyses and emphasis. Only those targets which had a seed sequence length greater than 7 bases and showed evolutionary conservation between human, mouse and rat

organism was included in the chi-square statistic. From that comparison, targets for this miRNA are moderately enriched for schizophrenia (N= 2291, df = 4, $p < 0.03$).

Experimental assessment of co-expression of miRNA targets in postmortem brain tissue

PRD 5 was validated in a two different formats, e.g. neuroblastoma and postmortem prefrontal cortex in human samples from SMRI using real time quantitative PCR (see above sections for validation experiments performed in neuroblastoma). In this section, the expression levels of PRD5 were compared to two of its predicted gene targets *C10orf26* and *ZNF804A* in the SMRI samples (figures 11,12)

The *C10orf26* and *ZNF804A* genes were assessed similarly to PRD5, using the reference genes, *IPO8*, *HMB5*, and *PPIA* as a reference. Choice of a reference gene for each set of assays was based on their level of consistency [124].

The relationship between *PRD5*, *ZNF804A*, and *C10orf26* expression levels was assessed using partial correlations, adjusting for the effect of potential confounding effects of antipsychotics on gene express levels. *C10orf26* was significantly negatively correlated ($r = -0.38$, n=102, $p = 0.004$) with PRD5 whereas *ZNF804A* was significantly positively correlated ($r=0.4$, n= 102, $p = 0.0006$). The direction of each correlation suggests that PRD5 may be targeting the 3'UTR of *C10orf26* and the 5'UTR of *ZNF804A*; the binding site alignments generated by one of the used prediction program suggest a similar conclusion.

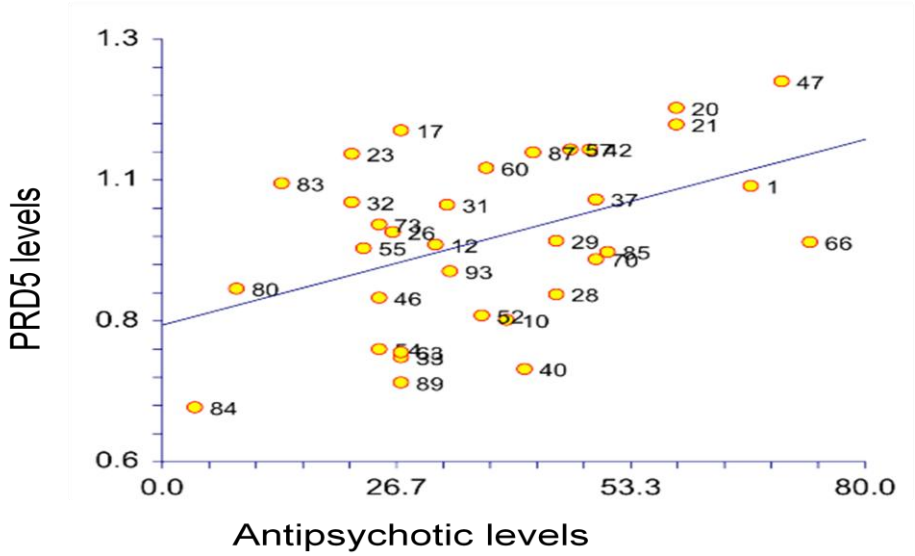


Figure 16 Effect of Lifetime Antipsychotic use on the levels of PRD5. The effect of potential confounders, e.g. gender, age, brain PH was estimated on PRD5 expression levels. Pictured on this graph are the 35 Schizophrenic and 7 Bipolar patients

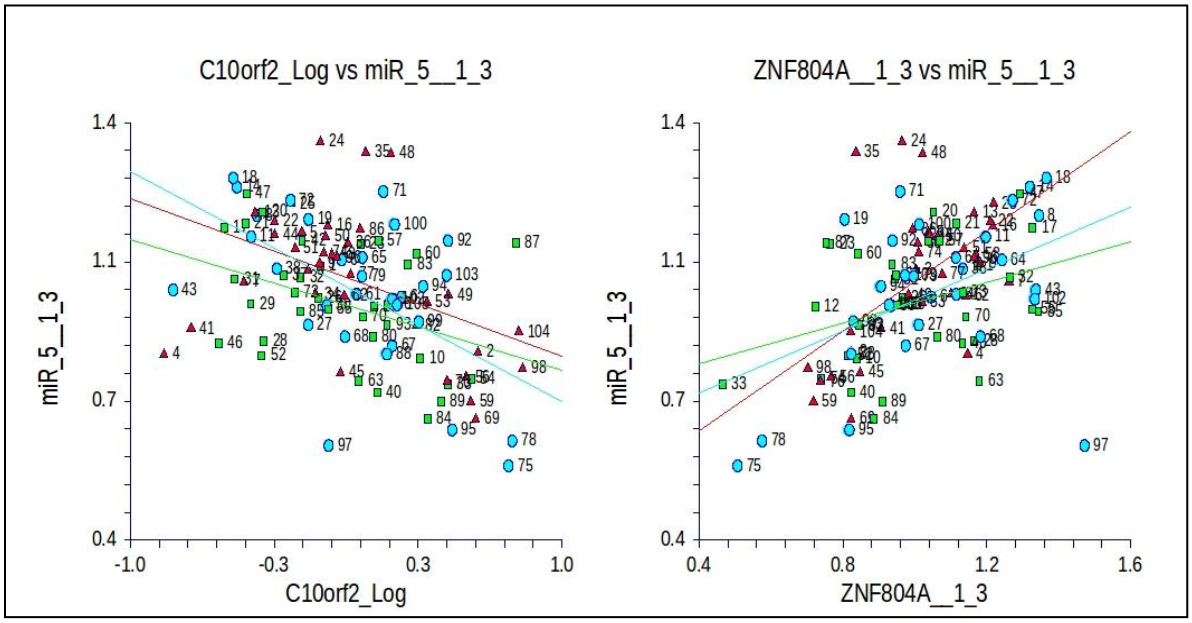


Figure 17 Correlation of expression values of novel miRNA and C10orf26 and ZNF804A. The colors green, cyan, and red indicate control, Schizophrenics, and Bipolars respectively. The correlation values for PRD5 and C10orf26 and ZNF804 were r = -.38, and r = 0.4 respectively.

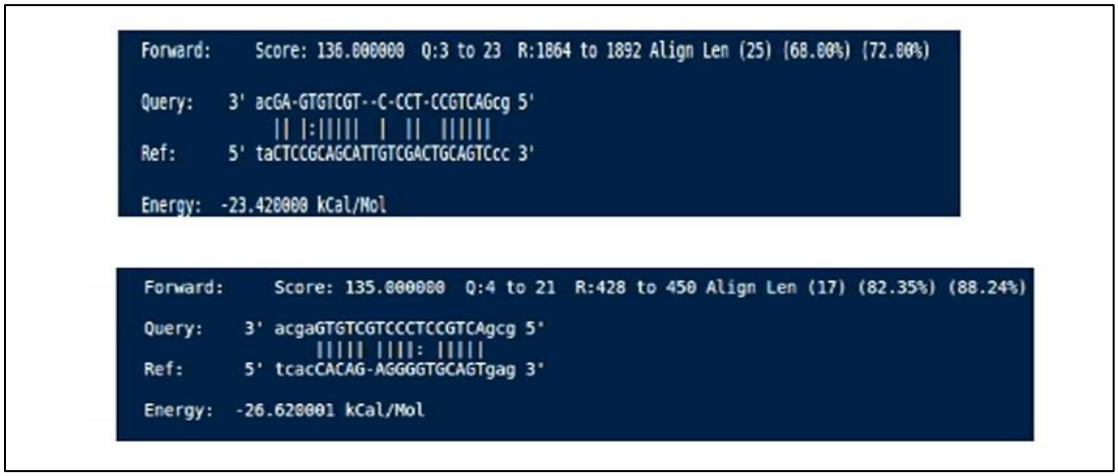


Figure 18 Binding site alignments for C10orf26 and ZNF804A generated by miRanda. Strong 3' compensatory binding along with a definite seed suggests that these genes are good probable targets for the novel miRNA PRD5. Pictured on top is the predicted binding site for the 3'UTR of C10orf26 and on the bottom is the predicted binding site for the 5' UTR of ZNF804A.

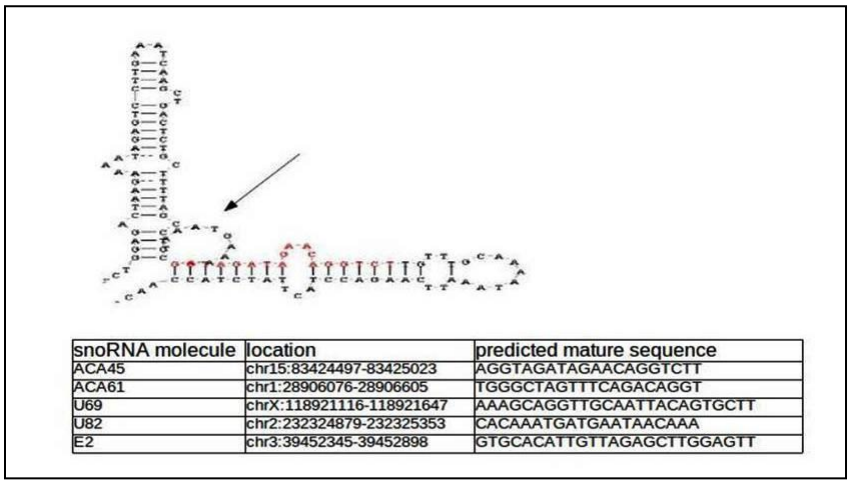
Detection of Novel miRNA from alternative biogenic sources

Also present in our in-house MicroSeq data set were five novel candidate miRNAs which were derived from snoRNAs. These additional novel miRNA were predicted from 119438 reads not originally used by miRDeep to predict the novel miRNAs described above. These reads were excluded by the software because they mapped to genomic regions traditionally thought not have potential miRNAs, i.e. those regions which contained snoRNAs. The exclusion of the reads, however reduced the amount of information gained by the experiment and so a bioinformatic pipeline was devised which reassessed the suitability of these reads and their mapped loci for miRNA prediction.

This pipeline included several steps traditionally included in miRNA prediction, i.e. assessment of precursor stability based on thermodynamics but allowed for the biogenic source to include snoRNA. First, the excluded reads were re-mapped against snoRNA-LBME-db [73]. This database houses the sequences for all currently known snoRNA sequences for humans. If the read mapped to an existing snoRNA structure, subunits of the larger molecule were created using a sliding window of 100 bases. Each subunit was then assessed using the program miPred. MiPred, a random forest program which determines the likelihood that a precursor hairpin is a true miRNA generates empirical p values for each sequence through 1000 permutations.

The initial p values from miPred ranged from 0.015 to 0.001 and percent confidence scores range from 76.7% to 52.8%. The sno-derived novel miRNA candidates mapped to E2, U69, ACA61, ACA45, and HBII-99B (table 6). One sno-derived miRNA, mapping to ACA45, detected in this pipeline was a replication of a

previous finding in a fibroblast cell lines (pictured below) [13]. In that work, the functionality of the sequence was demonstrated using a luciferase assay for cyclin-dependent kinase 19 (CDK19) [13]. It should be noted that one of the five sno-derived miRNA belongs to the orphan class of snoRNA described earlier in this chapter. Other sno-derived miRNA detected in the dataset but not tested because they have already been annotated include ones originating from 1) ACA36: hsa-mir-664, 2) ACA55: hsa-mir-4667-3p, 3) E3: hsa-mir-199a-5p, 4) HBII61: hsa-mir-1248 and 5) mgh28s-2411: hsa-mir-136-5p.



snoRNA molecule	location	predicted mature sequence
ACA45	chr15:83424497-83425023	AGGTAGATAGAACAGGTCTT
ACA61	chr1:28906076-28906605	TGGGCTAGTTTCAGACAGGT
U69	chrX:118921116-118921647	AAAGCAGGTTGCAATTACAGTGCTT
U82	chr2:232324879-232325353	CACAAATGATGAATAACAAA
E2	chr3:39452345-39452898	GTGCACATTGTTAGAGCTTGGAGTT

Figure 19 ACA45 snoRNA. Colored in red is the approximate mapped location of the dataset.

snoRNA		MiPRED score		annotation
	group	p value	% confidence	
ACA36	ACA	0.002	66.1	Hsa-mir-664
E2	ACA	0.012	69.6	
ACA55	ACA	0.015	52.8	hsa-mir-4667-3p
U69	ACA	0.008	61.8	
E3	ACA	0.001	63.7	hsa-mir199a-5p
HBI-61	ACA	0.001	76.7	Hsa-mir-1248
ACA61	ACA	0.001	59.2	
ACA45	CAJAL	0.001	60.2	
HBII-99B	CDBOX	0.001	67.3	
U82	CDBOX	0.006	58.1	
mgh28S-2411	CDBOX	0.009	56.7	hsa-mir-136-5p

Table 7 MiPred predictions for excised precursor for five potential novel sno-derived miRNA. Candidates include one described by Ender et al (ACA45) to have miRNA-like characteristics. This sno-derived miRNA was shown to associate with Argonaut proteins and to inhibit mRNA expression of cyclin-dependent kinase 19 (CDK19).

To determine whether the novel miRNA candidates were differentially expressed according to tissue type, all five were evaluated using a RNA panel of normal human tissues (First Choice, Life Technologies). Pictured in the figure below, the miRNA demonstrated expression among the twenty tissues on the panel that were different but not significantly so (two way ANOVA: df = 4, F =0.47083, p = 0.632).

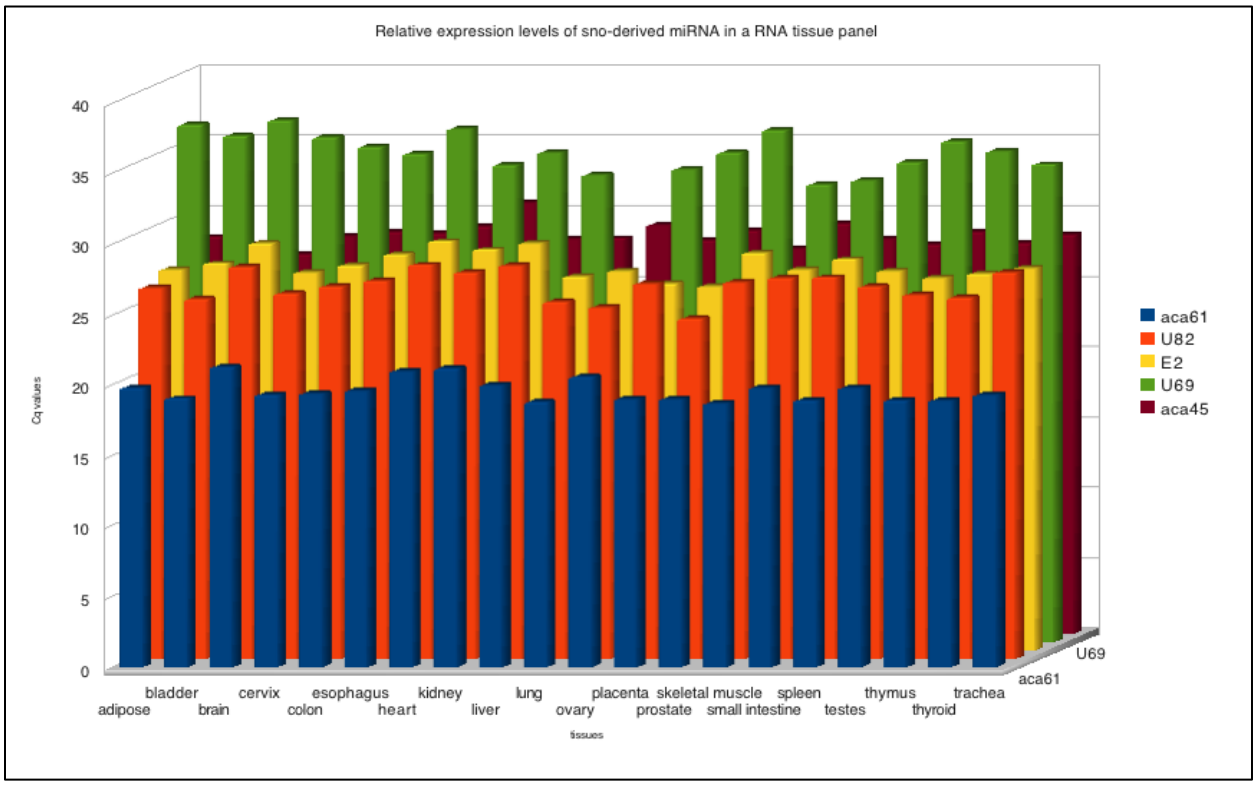


Figure 20 Relative expression values of sno-derived miRNA candidates as assayed in a RNA tissue panel of twenty normal human tissues. No significant difference was observed between the various tissues for this class of miRNA.

Chapter Discussion

Analysis of MicroSeq data is a complicated and computationally intensive process, but it is, without doubt, the most potentially productive data type generated to date with regard to the amount and types of findings generated therein. A key focus of chapter one has been the identification of novel miRNA and their evaluation in postmortem tissue samples. We successfully identified and validated in cell lines and postmortem brain tissue a number of novel miRNA demonstrating the usefulness of this technique for miRNA discovery.

Additionally, we compared a number of popular software in terms of the number

of miRNA generated as well as multiple versions of the same software in our dataset and six other publically available ones. We discovered that stark disagreement between these programs exists in terms of the precursor structure and the absolute numbers of predictions generated. ROC curves of simulated data clearly indicate that the program miRDeep is the most suitable software for the prediction of novel miRNA candidates. Because of the differences in stringency in the various programs, however, we suggest that a consensus approach across multiple programs be used when identifying novel miRNAs that will be evaluated for further study. Using a multi-program approach, we will able to successfully validate 68% (12 out of 17 overlapping) classically derived novel miRNA in the original cell line.

Additionally, another five novel miRNAs derived from snoRNA molecules were validated in a RNA tissue panel but did not show significant differences based on tissue type. It is unlikely that this group of novel miRNAs impact disease status as they did not show significant tissue based differences in expression. However, their detection suggests that alternative modes of biogenesis are possible for miRNAs and that by focusing on traditional biogenic pathways, current prediction software potentially misses a significant route of discovery. We were able to detect these miRNAs, further, in a set of reads excluded by the analytical software thereby increasing the potential effectiveness of the technology through a reassessment of current definitions for the detection of novel miRNAs. Accordingly, we might suggest a reconsideration of the underlying rationale of these software and their results.

The multiple program approach is a prudent cost effective methodology for the

identification of novel miRNAs using MicroSeq. It is difficult to directly compare the success rate, i.e. the number of miRNAs predicted versus validated, of the multi-program approach to the approaches used by other studies as most do not choose to validate their total set of predictions [210]. The current costs associated with deep sequencing and validation of novel predictions necessitates filtering approaches that generate the best possible candidates for testing. We suggest that a multi-program approach provides a good basis upon which additional selection methods can conceivably be employed.

Additional findings of this study, more pertinent to schizophrenia and bipolar disorder include the detection of the novel miRNA PRD5. While performing validation in the postmortem brain tissue of affected subjects, we identified PRD5 to be differentially expressed in schizophrenia cases only. This finding suggested that additional bioinformatic analysis and exploration of this miRNA is warranted. First, we sought to validate this miRNA in other cell lines than neuroblastoma. Using the datasets employed in the software comparison, we found presence of PRD5 in HL60 and K562 cells. HL60 are a common lab model for studying blood and K562 cells develop characteristics similar to erythrocytes. Second, a chi-square statistics comparing the predicted targets for this miRNA in a number of disease-specific gene sets showed mild enrichment for schizophrenia. Third we investigated the co-expression of this miRNA with two of its top predicted targets. From the list of predicted targets, filtered using a FDR test statistic and a multi-program approach, we chose two genes, *C10orf26* and *ZNF804A* with which to compare to PRD5. These genes survived FDR based correction, were predicted by multiple programs,

and showed binding site consistency across known isoforms. The directionality of the co-expression of *C10orf26* and *ZNF804A* with PRD5 in conjunction with the manner in which the miRNA is predicted to bind to the gene suggests that PRD5 may be involved in their regulation. Thus, all of this evidence, although tentative, strongly supports future experimental work on PRD5 in the pathophysiology of schizophrenia.

Chapter 3. Prediction of targets for differentially expressed miRNAs in the SMRI sample

Adapted From:

Kim, A.H. Reimers, M., Maher, B., Williamson, V., McMichael, O., McClay, J.L. et al. (2010). MicroRNA expression profiling in the prefrontal cortex of individuals affected with schizophrenic and bipolar disorder. Schizophrenia Research, 124 (1-3), 183-191.

Abstract

667 miRNAs were profiled using a real-time PCR based technique in the prefrontal cortex in a group of Schizophrenia (N=35) and Bipolar Disorder (N= 35) from Stanley Medical Research Institute (SMRI) [8]. Twenty-two miRNA were determined to be differentially expressed using a high-throughput commercial method (TLDA, Life Technologies); seven of which the initial twenty-two were verified a second time using single tube assay (Taqman, Life technologies). 3371 targets were predicted using miRanda for five of these differentially expressed miRNA (hsa-miR-132, hsa-miR-132*, hsa-miR-154*, hsa-miR-212, and hsa-mir-34a). In other words, each of the five miRNA genes on average predicted 634 gene targets.

Given that a single miRNA has been estimated to target approximately 200 genes, this number is likely to contain a large number of false positives and in need of filtering. We, therefore, used two approaches to filtering these predictions and to the selection of candidates for additional testing, i.e. a false discovery rate (FDR) statistic ($\pi_0 = 1$) and a biologically-based threshold that considers the effects of alternative splicing, co-expression, and mRNA site accessibility [116]. The biologically-based threshold approach attempts to account for other stochastic parameters that might impact the successful binding of a miRNA to a mRNA target sequence, in particular the consistency with which a binding site is located.

Two genes, tyrosine hydroxylase (TH) and phosphogluconate dehydrogenase (PGD), were selected that survived correction for multiple testing using the FDR statistic and the biologically based threshold approach. TH had had previously been suggested by Jacewicz et al, 2008 to be a candidate gene for schizophrenia [225]. The gene

expression level of both of these candidates was assessed in post mortem brain tissue from SMRI subjects using qPCR. Based on qPCR results, gene expression analyses show TH and PGD to be negatively correlated with hsa-miR-132 and -212 ($p = 0.0001$, 0.0017, 0.0054 and 0.017, respectively), suggesting a probable miRNA:mRNA interaction.

Introduction

MiRNAs possess several characteristics which make comparison of their relative expression levels on a large scale difficult without significant technological adjustment. First, though mature sequences are defined as 22 bases in length, sequencing of these molecules has shown experimentally their length to be more variable, ranging from ~15 to 25. This means that probes used in the detection of miRNA must be designed such that they bind with enough stability to compensate for these slight length differences. In addition, these probes must also be able to differentiate between members of a particular seed family which typically only differ by one or two bases. Second, variable GC content creates differences in melting temperatures potentially affecting the efficiency with which the probe binds to the target. Third, in the creation of cDNA, the miRNA generally lacks the poly A tail of mRNA or a consensus sequence that can be used in enrichment. Taken together, these issues have combined to make it difficult for anyone to assess miRNA variation in a high throughput manner, without making substantial adjustments in the way individual miRNA are detected and quantified.

Significant technological advances have been made, e.g. stem-loop primer design and locked nucleic acid chemistry (LNA) which facilitate high- throughput expression profiling through increased assay specificity. The stem-loop primer design was first advanced by Chen et al in 2005 [226]. Stem-loop primers bind to the 3' end of a miRNA molecule increasing specificity through base stacking and reduced spatial constraint. This primer design is extremely sensitive, capable of detecting as few as seven copies of sample per reaction with a high level of correlation between input and result ($r^2 > 0.994$) [226]. Additionally, stem-loop primers are insensitive to miRNA

precursor sequences or genomic DNA potentially diluting any observed signal.

Originally created in 1998, LNA chemistry improves melting temperature (TM) normalization and facilitates more accurate probe binding in high throughput platforms [227, 228]. In LNA chemistry the ribose moiety of a RNA nucleotide is modified with a 2' oxygen and a 4'carbon, locking the ribose sugar into a 3'-endo conformation. Because LNA monomers [2'-*0,4'-C*-methylene-β-o-ribofuranosyl monomer] mimic RNA nucleotides they can be mixed commercially in a wide variety of formats lower melting temperature substantially.

Approximately eleven separate miRNA expression profiling studies have been performed since 2007, examining the relationship between miRNA dysregulation and disease with varying results (see chapter one for an expanded discussion of these studies). Very few miRNAs have been identified that have performed with any degree of consistency in these studies - with hsa-mir-181b being the best candidate overall – but many of the targets of these implicated miRNAs include key genes already identified in genetic studies of schizophrenia and bipolar disorder. The source of this inconsistency is most likely due to the small number of subjects sample and the nature of the confounders present in post mortem brain tissue. These studies, nevertheless, suggest a rationale for exploring further the nature of the relationship between a miRNA and its target gene.

This chapter delineates the steps and issues involved in miRNA target prediction following an expression profiling study in schizophrenic and bipolar patients from the SMRI. It uses miRNA expression profiling of the schizophrenic and bipolar patients from the SMRI as an example to illustrate the cogent issues involved miRNA target prediction

and the steps required for the successful prediction of gene target prediction. In this chapter, successful prediction of miRNA targets is defined as the experimental validation of gene targets through the significant partial correlation of respective expression levels of both miRNA and predicted gene target following adjustment for experimental confounds. We explore FDR-based approaches to the filtering of gene target prediction algorithms as well as biologically based thresholds, highlighting specifically the effect of alternative splicing on the consistency of the binding site placement.

Materials and Methods

Description of samples used in profiling and gene expression validation

200 mg of postmortem brain tissue, originating from the dorsolateral prefrontal cortex (Brodman's Area 46) were received from the Stanley Medical Research Institute (SMRI). Exclusion criteria for subjects included: 1) brain pathology, 2) central nervous system disease, 3) poor RNA quality, 4) IQ < 70, 5) age <30 years, and 6) substance abuse within one year of death. Total demographics for this study group can be found in chapter 2 of this thesis. Total RNA was isolated from approximately 100mg of this tissue using the miRVana-Paris Kit (Ambion, Texas). RNA quality (RIN) was measured using nano chip on the 2100 Bioanalyzer (Agilent, California) and was 7.2. The protocol for cDNA synthesis and rtPCR was performed according to manufacturer's recommendations.

Target Prediction

Gene targets were predicted using the program miRanda (August 2010 release) from the 3'UTR sequences of all known protein coding genes (grCh37.p8). The settings for miRanda include: 1) scale = 4.0, 2) gap opening penalty = -2.0, 3) gap extend penalty = -8.0 and 4) energy threshold (kcal/mol) = -20.0. MiRanda uses affine penalties of length with respect to gap opening and extension; the algorithm also employs a scaling factor to the first eleven positions of the mature sequence to reflect 5'-3' symmetry [224]. Further, in keeping with experimental observations, four rules are applied within the algorithm with respect to the conceptualized seed portion of the mature sequence. First, no mismatches are allowed between positions 2-4 of the 5' end of the mature sequence. Second, 5 or fewer mismatches are allowed between positions 3-12 of the 5' end of the mature sequence. These first two rules protect what is generally believed to be the most crucial point of miRNA:mRNA alignment, the seed sequence. The third rule allows for 5 or fewer mismatches between positions 9 and L-5, where L is total alignment length. The last four rule states that 2 or fewer mismatches in the last five positions of the mature sequence [224]. The last two rules focus more on the 3' end of the sequence and allow for flexibility based on the any gaps which might introduced in the initial alignment.

There are currently 59871 genes annotated in the Ensemble database (GrCh37.p8), 22088 of which are classified as protein coding. The predicted targets for this project were based on protein coding genes only. A protein coding gene was defined as a locatable region of the genome which could be inherited and possessed a

combination of regulatory regions, transcribed regions, and/or functional sequence regions [229,230, 231, 232]. As there is no current consensus for the true numbers of genes present in the genome, this number represents an estimation based on a combination of *ab initio* gene prediction software and comparative approaches [229, 230, 231, 232]. The 3'UTR sequences were downloaded for each gene using Ensembl Biomart API and the median size for the sequences was 702 bases.

Screening using biologically-based approaches

All predicted targets from the previous section were screened on the basis of co-expression, number of transcripts, prediction number per genes and degree of binding site consistency across multiple transcripts. Studies have suggested that the number of times a program predicts the presence of a binding site within a particular gene correlates with the likelihood that the prediction is true [51]. In addition, a high co-expression of both target gene and miRNA means that there are sufficient numbers of both to bond within the cell. The presence of a binding site in multiple transcripts of a particular gene suggests that it might be important to the cell's function, so much so that it is maintained regardless of alternative splicing.

In this project, a computational pipeline was created which: 1) identified transcripts for each target prediction through the expressed sequence tag database (EST) Aceview, 2) aligned the transcripts using ClustalW (ebi.ac.uk/Tools/msa/clustalW2), and 3) assessed the consistency of the location of the binding site across those transcripts. In order to minimize bias, the sequence order of alignments of the individual gene transcripts was permuted 1000 times. Clustal W

performs a pairwise alignment of the first two input sequences from which a nearest neighbor guide tree is created that is used to align all other sequences. In this scheme, the sequence order impacts the initial pairwise alignment and thus the creation of the nearest neighbor guide tree. Therefore permutation of sequence order is important to reduce the possibility of this bias occurring. The settings for Clustal W included: 1) gap open penalty = 10.0, 2) gap extension penalty = 1.0 and 3) weight matrix = IUB.

Co-expression was determined by comparing values for gene targets and predicting miRNA in the Gene Expression Atlas (release 12.09) (www.ebi.ac.uk/gxa). This database is currently holds the expression data from Affymetrix expression arrays for 703295 genes across 3384 separate experiments [223].

Screening using False Discovery Rate

A number of different approaches to correction for multiple testing have been developed including Bonferroni tests, permutation testing, and empirical p value. Many of these tests can be too strict and reduce the power to observe a significant result. By focusing on the p values of a particular test and creating adjusted p values (q value), the FDR has the ability to discover significant results without sacrificing statistical power. For example, a p value of 0.05 in a specific test implies that 5% of all tests will result in false positives. If one uses a FDR based correction on the same of tests, the adjusted p values, assess a fewer number of tests, i.e. q values only assess the number of significant tests. In this example, the adjusted p values would only index the number of significant tests rather than the total experiment.

Results

Prediction of targets for hsa-mir-132 and hsa-mir-212 using biological filtering

Initially miRanda predicted a total of 4482 separates genes as potential targets; approximately 75% of these genes (3371) had multiple transcripts. When biological filtering was applied, this number was reduced further to 2156. From this number based on co-expression levels of the targets, site consistency across multiple transcripts, TH and PGD were selected further experimental tests. Both TH and PGD showed greater than 95% level binding site consistency and were highly expressed in the cerebellum , and brain.

Prediction of targets for hsa-mir-132 and hsa-mir-212 using FDR-based filtering

FDR analysis of the gene target predictions suggested an interesting pattern with respect to the miRNAs, hsa-mir-132, -212, and -154. Though both miRNAs (hsa-mir-132, -212) were predicted to have a number of significant predictions, after FDR adjustment, hsa-mir-212 indicated a higher number than hsa-mir-132. These miRNAs overlap each other in the genome and share a portion of their precursor and seed sequence hsa-mir-132(-5p: ACCGTGGCTTTCGATTGTTACT) and hsa-mir-212 (-5p: ACCTTGGCTCTAGACTGCTTACT).

The lower number of significant targets predicted for hsa-mir-132 compared to hsa-mir-212, despite their sequence overlap, suggests that hsa-mir-132 exercises a stronger control over its targets, thus causing a greater impact on the gene function and cell phenotype, respectively. Indeed, hsa-mir-132 has been shown to be implicated not only in Schizophrenia but in other psychiatric conditions such as autism, alcohol and/or

drug dependence [130].

gene	hsa-miR-132	hsa-miR-132*	hsa-miR-154*	hsa-miR-212	hsa-miR-34a
TH	4	7		10	13
C6orf60	6			6	
PPIL5	2		5	2	
ELA2A_HUMAN	4			4	
SNHG5	4			4	
Cxorf26	4		2	2	
MCOLN3	4			4	
MMP26	4			4	
ADRA1A	1	6		1	

Table 8 Extracted excerpt from target table generated for differentially expressed miRNA. Pictured are genes having multiple transcripts (> 3), multiple predictions, and a significant p value after FDR correction. The A number of filtering steps were employed against the initial predictions generated by miRanda including 1) binding site consistent across multiple transcripts and 2) high conservation of site in multiple species. The complete table with all unfiltered predictions can be found in appendix 7.

FDR analysis of the gene targets for hsa-mir-154* suggest, in addition, a large number of significant target predictions. Given that hsa-mir-154* is a star sequence and as such occurs on average less frequently this is likely to be a problematic finding. The typically lower levels of the star sequence in comparison to the mature, generally seen in the cell, make the star sequences less likely to be functional, ie. their lower levels place them at a disadvantage when competing for access to binding sites.

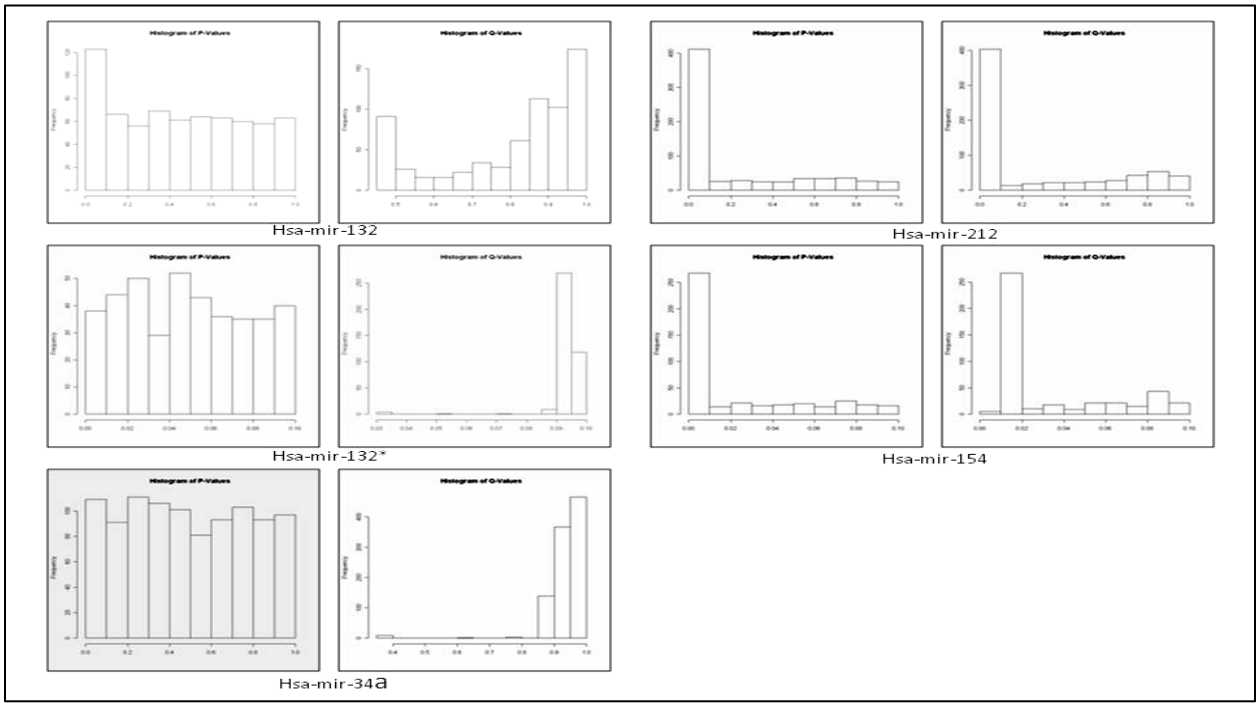


Figure 21 Density distributions calculated for predicted targets of each respective miRNA.

Experimental assessment of Co-expression

Gene target predictions for TH and PGD were verified using single tube validation (depicted in figure 17). miRNA assays were run in triplicate and normalized against the reference snoRNA RNU44 using the $2^{-\Delta\Delta}$ algorithm; gene expression assays were run in triplicate and normalized against the reference control IPO8 using the $2^{-\Delta\Delta}$ algorithm. A different reference gene was in these two studies due to the differing nature of the gene input. Spearman (p) coefficient was used to estimate miRNA correlations. Mild negative correlation between gene PGD and TH and hsa-mir-132 and hsa-mir-212 expression levels was detected (PGD: R = -0.29, p = 0.0017 hsa-mir-132, R = -0.22, p = 0.0017 hsa-mir-212; TH: R = -0.41, p = 0.0001 hsa-mir-132, R = -0.35, p = 0.00054 hsa-mir-212).

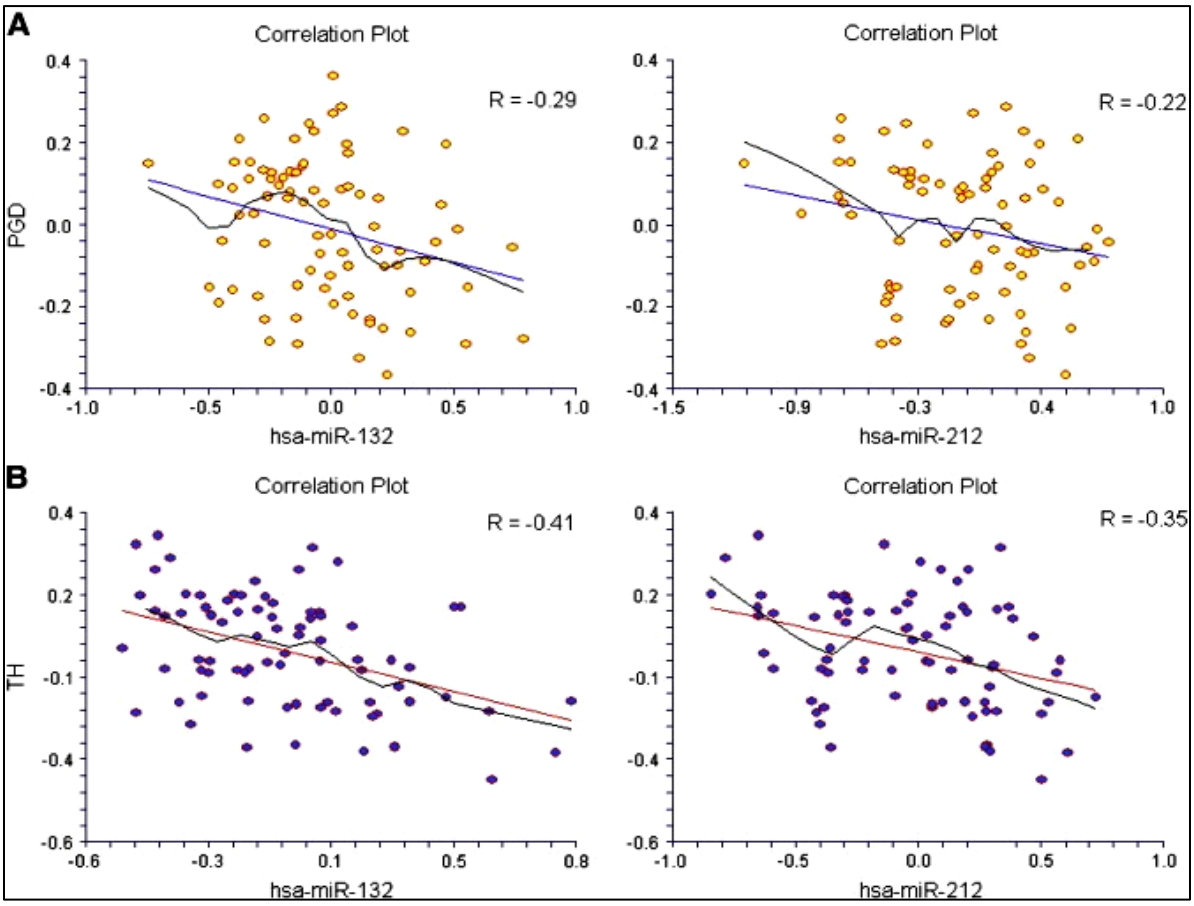


Figure 22 Spearman (p) Coefficient correlation plots for genes PGD (A) and TH (B).Values were log-transformed and raised to the power of 1/3 to approximate a normal distribution. The values were then fitted into an analysis of covariance (ANCOVA) model with pH, age, RIN, sex, and disease status as covariates. Image taken from Kim et al, 2010.

Chapter Discussion

Predicting gene targets for miRNAs is an important step in miRNA research as it positions the miRNA within the grander genomic context and allows one to infer function through associated gene networks. The process by which these targets are predicted and identified however is less than ideal and suffers from a high false positive rate. At present, the source of this error is unclear. Many authors suggest that it is due to the

small size and binding behavior of the mature sequence in animal genomes. Unlike plant miRNA, animal miRNAs bind imperfectly with their target miRNA and are affected by dynamic cellular factors such as molecule concentration, thermodynamics as well as additional features present in the genome.

The participation of these factors, thus, increases the difficulty to accurately capture and predict entirely what is occurring in the cell at any point in time. One example of those factors not adequately explained by current algorithmic approaches to miRNA prediction are miRNA sponges [233,234, 235]. MiRNA sponges" soak up" the miRNA s that would otherwise target a functional gene with a similar sequence. These sponges are often pseudo-genes which have a functional 3'UTR matching that of the target mRNA. Recently, one such "sponge" has been reported for PTEN [233,234,235].

In chapter 3 we predicted targets for five miRNA that were shown to be differentially expressed in group of Schizophrenic and Bipolar patients. 3371 targets were initially predicted for these five miRNA, with each miRNA having on average of 600. Assessment of these targets showed that 75% of them had multiple transcripts and of this 75%, 63% has multiple predictions occurring within the 3'UTR sequence. Multiple predictions for a gene could inflate a FDR test statistic. At the same time, the presence of multiple sites would suggest an increased likelihood for binding between the miRNA and the mRNA target.

Correction for multiple testing among these miRNA using a FDR based statistic effectively reduced the number of gene predictions but comparison between the individual miRNAs was problematic, especially so in the case of hsa-mir-212, and -132.

Notably, there was a difference in the number of significant predictions for hsa-mir-212 and -132, despite their overlapping seed sequences. This discrepancy may be due to some inherent difference in the relationship of these two miRNA with their respective targets or it may speak to some difficulty with comparing FDR across experiments [133]. One would expect that because these miRNAs overlap in sequence and location that they would target a similar number of genes. The two gene targets that were validated experimentally, TH and PGD survived screening by both FDR statistic and biologically based parameters. Both genes show expression patterns that negatively correlate with hsa-mir-132, -212, suggesting that the 3'UTR is targeted. Certainly, additional tests are needed to validate the interaction of hsa-mir-132, -212 with these targets, e.g western blots.

Chapter 4. Bioinformatic assessment of imputed variants with respect to miRNA efficiency

Adapted From

Yan, J., Aliev, F., Webb, B., Kendler, K., Williamson, V., et al. (2012) Using genetic information from candidate gene and genome-wide association studies in risk production for alcohol dependence. Addition Biology (submitted).

Chen, X., Chen, J., Williamson, V.S., An,SS., et al. (2009) Variants in nicotinic acetylcholine receptors alpha5 and alpha3 increase risks to nicotine dependence. Am J. Med Genet. B. Neuropsychiatr Genet. 150B(7) 926-933.

Chen, X., Williamson, V.S., An, S.S., Hettema, J.M., et al (2008) Cannabinoid receptor 1 gene association with nicotine dependence. Archives of General Psychiatry 65(7) 816-824.

Chen, X., Wang, X., Chen, Q., Williamson, V., et al. (2008) MEGF10 association with schizophrenia. Biological Psychiatry 63(5) 441-8.

Abstract

Imputation of single nucleotide polymorphisms (SNPs) has rapidly become a preferred method for increasing the resolution of SNP arrays. Imputation of 93 Affymetrix version 5 SNP microarrays was performed in order to provide increased resolution and definition of existing SNP data from the SMRI sample. The reference panel and legend files used in the imputation were based on HapMap3. Over two million additional SNPS were generated having an R^2 value greater than 0.5 and a confidence index greater than 0.99. Several steps pre-and post imputation were employed to ensure quality. Pre-imputation steps taken included: 1) examining arrays for surface flaws and experimental biases, 2) excluding SNPs based on low call rate, and 3) excluding SNPs which violated Hardy Weinberg equilibrium (HWEV). To filter SNPs for HWEV, a false discovery rate (FDR) was used. Post-imputation steps included an assessment of imputation accuracy through SNP masking and the use of confidence intervals/test statistics generated by the program IMPUTE2/SNPTEST.

Imputed SNPs were assessed bioinformatically to determine which could potentially affect the miRNA function and biogenesis. Of particular emphasis was the identification of SNPs which affect miRNA secondary structure (miRSNPs) because such a variant could theoretically impact the largest number of genes without radically changing the genome. Another area of focus was whether SNPs fell within differentially methylated regions; the rationale for this analysis stemmed from the significantly high number of miRNAs that have been found to be located on or within predicted CpG islands [184].

A large number of imputed SNPs were found to fall within CG rich isochores and

previously identified differentially methylated regions. 35% were SNPs present in the platform but that had been excluded during quality control. A much smaller number of SNPs (66) were identified as impacting miRNA regulatory function through an alteration of its secondary structure; this latter finding is in line with earlier estimates of the distribution of miRSNPs. This project demonstrates that the utility of the imputation approach for identifying relevant SNPs for study. Though the number of microarrays used in this imputation and the size of the reference panel undoubtedly affected its accuracy the fact that all of the chips were processed in the same lab using the same sample precluded any introduced bias that might have arisen through batch effects.

Introduction

Imputation has become a well established practice for increasing statistical power of GWAS studies to identify relevant variants. In imputation, the haplotype patterns of a densely genotyped reference population are used to predict the missing genotypes of a study group, using flanking genotyped SNPs from that group as a guide [29,37]. SNP imputation has been shown to boost GWAS power by as much as 10% and can be used to compare results across multiple genotyping platforms [37,40]. Several programs are now available with which to impute SNPs, e.g. IMPUTE2, PLINK, MACH, fastPHASE and the methods by which the programs make their determination vary [30-32, 36, 38]. IMPUTE2 uses a Markov chain Monte Carlo algorithm (MCMC) and has an accuracy rate of > 95% with SNPs that are in high LD and have adequate sampling density [30-32].Overall, in a comparison of the software mentioned above, IMPUTE2 was shown to provide superior results with regard to the effect of linkage disequilibrium (LD), marker density, minor allelic frequency (MAF) on imputation accuracy (IA) [33, 35]. Sources for the reference files include well characterized populations such as those represented by the HapMap 3, Seattle SNPS, and the UK10K. Recently, the 1000 Genomes Project has become a popular source for reference material surpassing the HapMap3 both in terms of numbers of individuals (~2500) and geographical regions (> 25) sampled [36].

The bioinformatic assessment of the role played by a variant in the etiology of a disorder can encompass a number of characteristics regarding function including location, effect on protein structure, species conservation, as well as its relationship to other noncoding features present in the genome. Typically, linkage disequilibrium (see chapter one for a discussion) has been used to filter the number of SNPs needed for

testing to adequately cover a particular gene.

In particular, variants which affect miRNA regulatory efficiency (miRSNPs) have been highly sought after in the last few decades because their ability to impact large numbers of genes simultaneously [44, 43, 49, 55, 56, 57, 58, 59, 60]. Falling into a number of classes, variants can either: 1) affect the secondary structure of the miRNA, 2) fall within the binding sites of a miRNA's predicted targets or 3) possess some other characteristic that necessarily impacts miRNA functionality. Assessment of the strength of a variants impact on the miRNA is determined on its location, with SNPs falling in the seed sequence or terminal loop having a higher probability of affecting functionality [174].

Materials and Methods

Sample Description

Briefly, raw cel files from 93 microarray chips (Affymetrix version 5) were downloaded from the Stanley Medical Research Institute Online Genomics database (www. stanleygenomics.org) and analyzed using methods employed by the programs Affymetrix Power Tools (APT) and Beagle call. The chips were used to index SNP and structural variation in a group of Schizophrenic, Bipolar and Control patients. The Affymetrix version 5 SNP chip is comprised of a majority of SNPs taken from the previous 500K mapping chip sets as well as an additional 1.8 million probes that can be used to measure features such as copy number variation. Description of the methods employed in sample ascertainment and processing have been described elsewhere [108]. The sample contained approximately a 2:1 male to female ratio and this was

compared to a known gender list for all samples.

For this project, the intensity values were extracted from their respective chips (“apt cel-extract”) and input into each program accordingly. Additionally, variability among individual probes was normalized using quantile normalization and a loess curve.

Quality Assessment of Microarrays prior to Imputation

As strict quality control (QC) is important to imputation, a number of steps were taken prior to imputation that would insure result superiority. All microarrays were screened using established methods Affymetrix PowerTools (APT) and an in-house R script that uses large deviations in probe intensity to detect the presence of physical flaws in microarray. Any microarrays demonstrating significant physical flaws, e.g. scratches, thumb prints, bubbles were discarded.

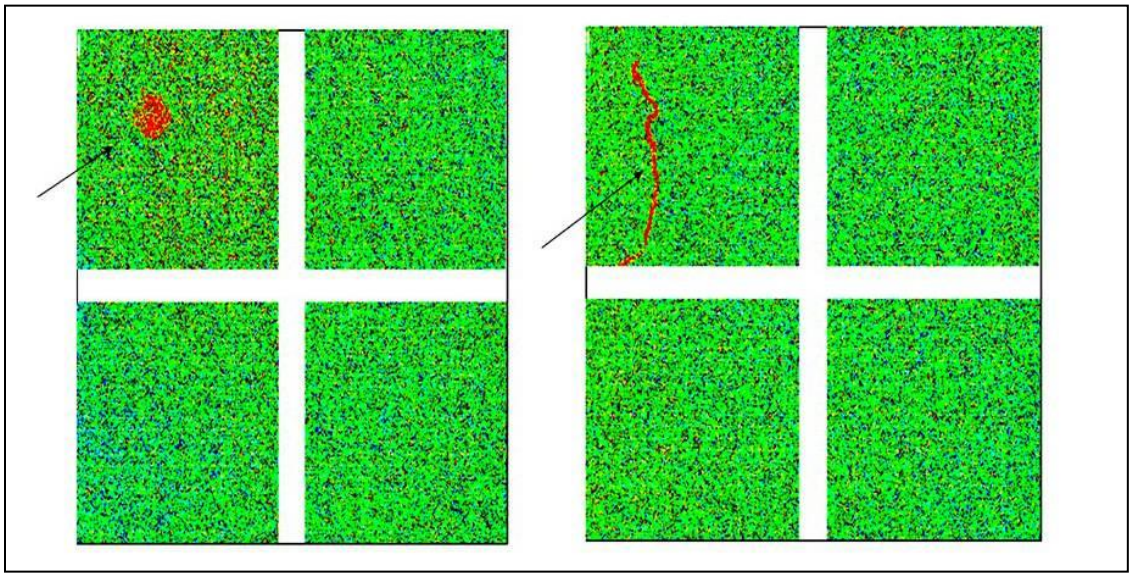


Figure 23 R scripts comparing intensity deviations were used to identify surface flaws on SNPs chips. Pictured is an example of the output generated by this script.

Determining Individual genotypes using Affymetrix PowerTools and Beagle Call

Genotypes were determined twice, first with APT and then with BeagleCall to determine agreement [38]. SNPs were removed that either:1) violated Hardy Weinberg equilibrium, 2) did not exhibit complete genotype agreement between APT and BeagleCall, or 3) that have a call rate (CR) < 95%. Call rate was defined, visualized below, as the difference between the total number of SNPs and the number of SNPs for which 95% or greater individuals were genotyped divided by the total number of SNPs assayed on the chip overall.

$$CR = (N - nocall)/N$$

Hardy Weinberg violations are often an indication of assay failure and have been shown to significantly impact the summary odds ratio of gene association studies by more than 10% [106]. A false discovery rate (FDR) was applied to the genotyped SNPs to exclude SNPs which effectively violated the Hardy Weinberg Equilibrium ($p+ q = 1,; p^2 + q^2 + 2pq = 1$). To compensate for cryptic stratification, a reference set comprised of individuals from several different populations was created that included information from HapMap 3. The table below lists number of samples and SNPs associated with each respective population in that reference panel. Cryptic stratification and cryptic relatedness is most often present when the hidden origin of members of a study group is not taken into account. These “confounders” often produce false signals of association within a GWAS study. An imputation reference panel that is made up of multiple ethnic groups can provide a wider reference frame thereby diluting these effects [29].

population	Sample number	Polymorphic SNPs
ASW	87	1543115
CEU	165	1397814
CHB	137	1341772
CHD	109	1311767
GIH	101	1408904
JPT	113	1294406
LWK	110	1526783
MXL	86	1453054
TSI	102	1419970
MKK	184	1532002
YRI	203	1493761

Table 9 HapMap3 release 3 reference panel information

Assessment of Imputation Accuracy

After imputation, SNP quality was determined both on a study-wide as well as per SNP basis. To determine study-wide imputation accuracy (IA), 5% of the original genotypes were masked for each chromosome and the chips imputed a second time focusing on regions containing the missing SNPs. Per SNP QC was determined using the `info' metric generated by IMPUTE2/SNPTest. This score (0-1) corresponds to the amount of information at an individual SNP about the population allele frequency; it can be interpreted as αN where N is the number of subjects and α is the score [29]. This score corresponds well to the dosage measure R^2 provided by MACH [37]. While, currently, there is no set threshold for screening SNPs, several authors have found an info metric of 0.3 to be acceptable [32,33]. In this project, SNPs having an 'info' metric of 0.5 and a confidence interval greater than 0.9 were retained.

Imputation of SNPs using Impute2

A Perl script was written that split each chromosome into overlapping intervals of five megabases each and then imputed SNPs using that interval. A sliding window was used, extracting data sequentially and creating regions of overlap which were later filtered. All SNPs were prephased prior to imputation using the program Shape-It [121]. GWAS files were aligned to the forward strand of the human reference sequence (Hg18). Technically it is not necessary to split chromosomes in this manner but this step was done to allow for parallel computing and to allow for quick comparison of those regions which overlapped with centromeres (centromere locations are downloadable from www.hgdownload.cse.ucsc.edu/goldenPath/hg18/database/cytoBand.txt.gz) and chromosome ends. Impute2 allows a default buffer zone of 250 kilobases which was compared to determine consistency. After imputation, the fragments were concatenated into a single file. Sex chromosomes (X and Y) were not imputed due to the problems associated with hemizygosity.

Results

Pre-imputation Quality Control

Overall, seventy–four percent of all arrays (69 out of 93) had a call rate greater than 95%. In addition, qualitative assessment of physical flaws using the previously described R script was unable to eliminate any single microarray.

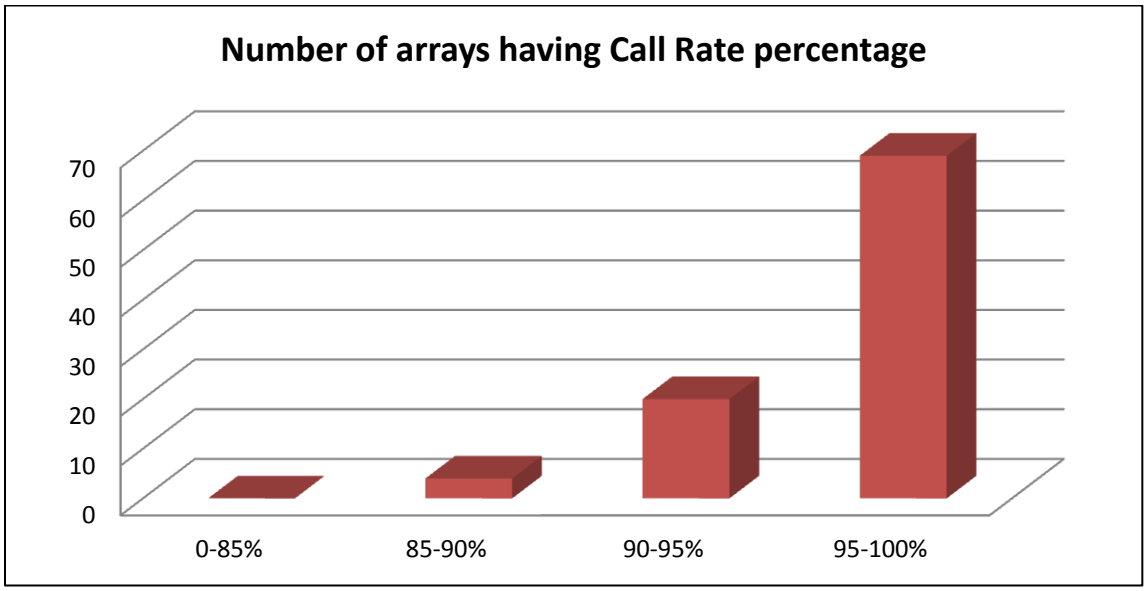


Figure 24 Call Rate was determined for each array using APT.

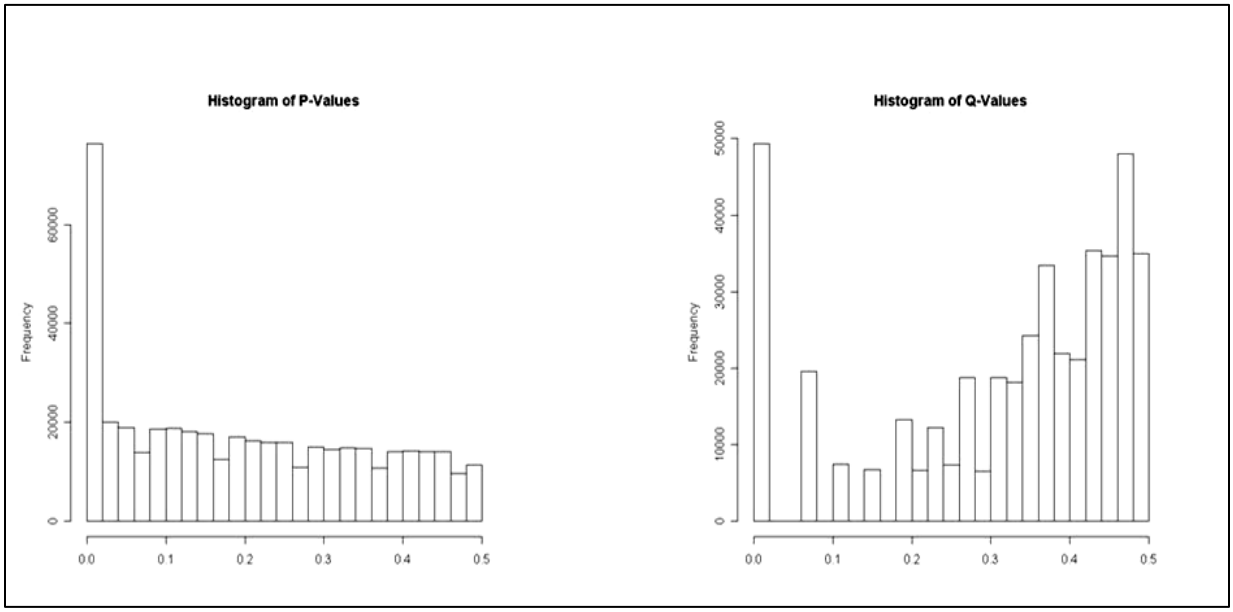


Figure 25 Calculated FDR values for HWEV. Based on an FDR (0.05), approximately 5% of SNPs should be viewed with caution as they are false positives. Pvalues pictured left and adjusted p values(Q values) pictured right

A false discovery rate statistic (FDR = 0.05) was calculated on genotypes after HWEV p

values were determined using PLINK [120]. 438,681SNPs called by Beagle and APT were assessed with the FDR test statistic; 21,934 of that number (5%) should be viewed with caution as potentially representing false positives. The inclusion of these SNPs into the imputation process may have affected the validity of the subsequent SNPs that were generated.

Post imputation Quality Control

The average imputation accuracy percentage was 97.6%. Chromosomes 1,2, and 3 demonstrated the highest accuracy level (98%, 96%, and 95%, respectively) whereas chromosome 14 demonstrated the lowest (85%). The low level of imputation accuracy for chromosome 14 may in fact be the reason why no bioinformatically significant SNPs were discovered in this region (see discussion of this in next section).

Bioinformatics of Screened SNPs

Over 2.5 million additional SNPs were generated using IMPUTE2 and screened bioinformatically. The program GTool (well.ox.ac.uk/~cfreeman/software/gwas/gtooll) was used to convert genotype probabilities into respective alleles which were then assessed. Two main areas of functional impact were considered: 1) potential methylation and 2) miRNA structure. Perl scripts were written in each case that performed analysis. To determine potentially methylated SNPs, the Perl script assessed the alleles of the imputed SNPs in terms of their nucleotide content and their respective position with regard to predicted CpG Islands, genomic isochores, and an experimental database comprised on Methylseq and BSseq data established by the Salk Institute for Biological studies [107, 117, 118]. The liftover tool (www.genome.ucsc.edu/cgi-

bin/hgliftover) was used to convert genome coordinates between the assemblies used by these three sources.

The coordinates for the CpG islands were taken from UCSC Genome Browser and are based on the parameters set by Gardiner-Garden and Fromme [118]. Briefly, in order to be considered a CpG island, a region must be longer than 200 bases, have a GC content greater than 50% and a observed/expected GC ratio greater than 0.6 [118]. The isochore coordinates used were determined by Costanini et al [107] through genomic segmentation. The methylation database (neomorph.salk.edu/human_methylome), comprised of Methylseq and BSseq data provides single base resolution maps of H1 human embryonic stem cells and IMR90 fetal lung fibroblasts as well as differentially methylated regions that can be assessed.

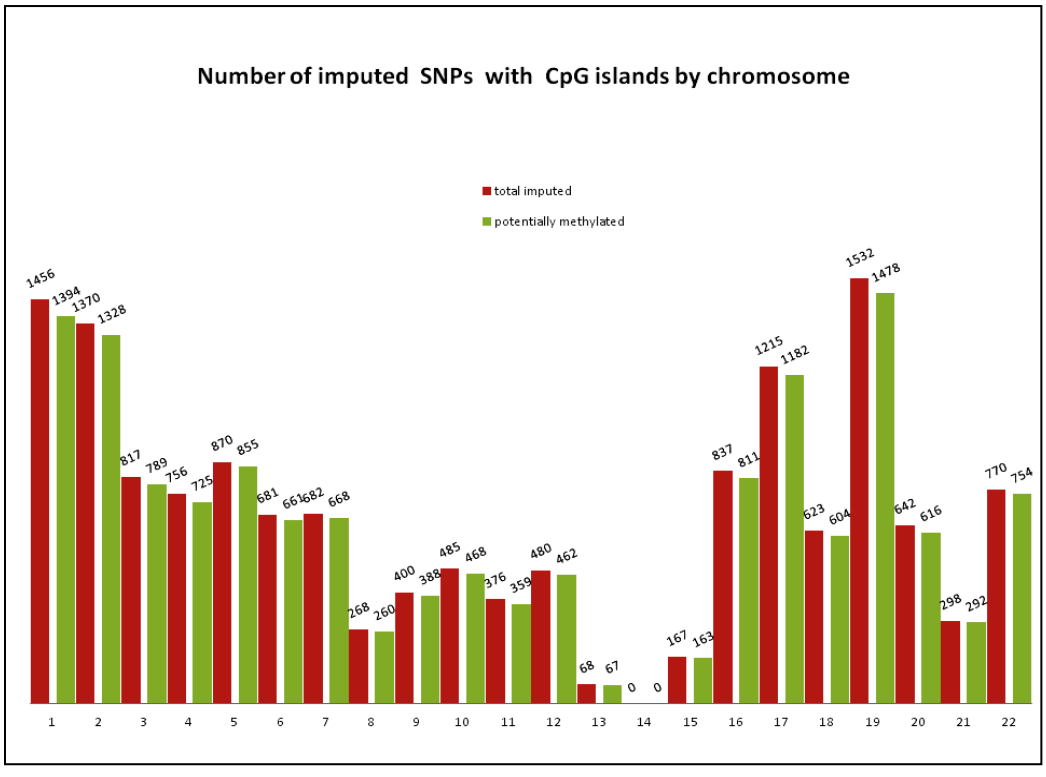


Figure 26 Number of Imputed SNPs falling within CPG islands

Interestingly, a large number of the imputed SNPs were located within CpG islands in the genome suggesting them to be of significant importance (figure26). When comparing this number to the number of genotyped SNPs present on the array with a chi-square test ($df = 1$, $N = 838461$, $p < 0.0001$) there was a significant enrichment for imputed SNPs falling in these regions. It must be said, however, that only 15% of genotyped SNPs from the original platform with a high GC content (>60%) actually survived quality control filtering (call rate, HWEV). The significant number of SNPs may in fact be the result of backwards imputation where SNPs previously excluded due to QC measures were captured by IMPUTE2.

Nevertheless, the CpG islands which contain imputed SNPs in turn fell within several prominent (H3) isochores on the respective chromosomes. In particular, almost all of the imputed methylated SNPs on chromosome 1 and 19 were found in H3 isochores having a GC % greater than 53%.

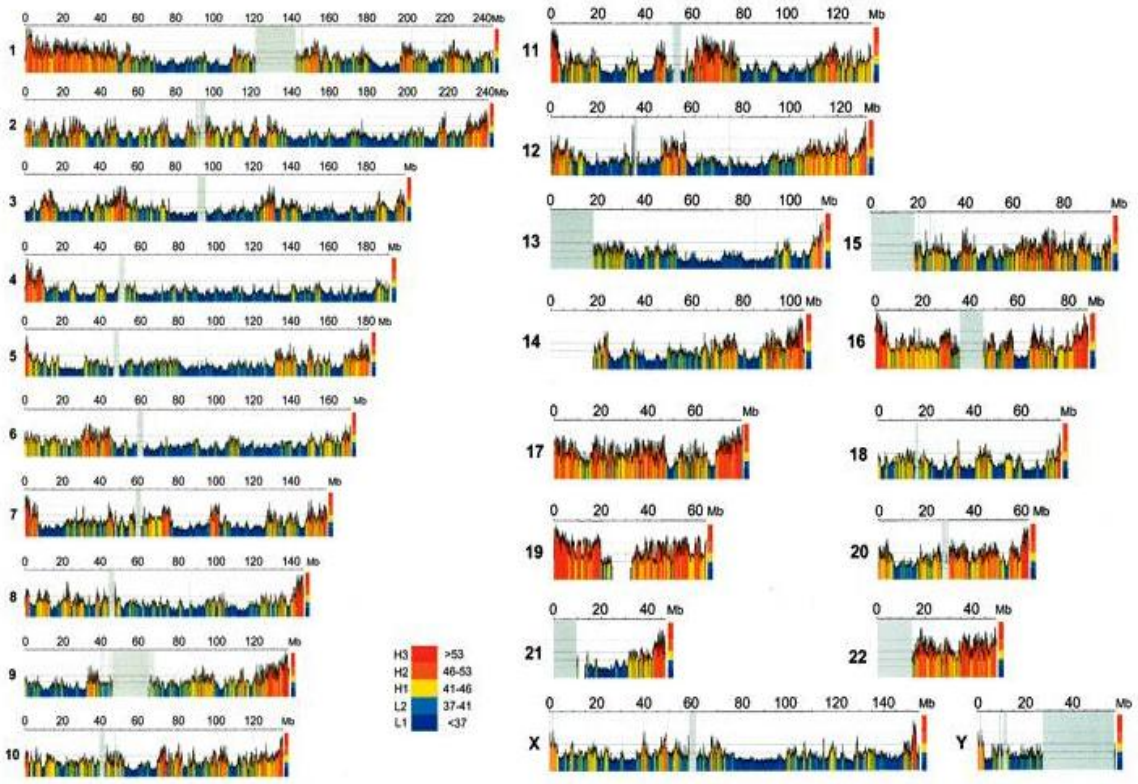


Figure 27 Isochore map for Human genome as generated by Constanini et al.

Isochores are large regions of DNA (>300 kilobases) with a high degree of uniformity in guanine (G) and cytosine (C) content. There are five families of isochores, L1, L2, H1, H2, H3 which vary in terms of the %GC and location throughout the genome. Isochores have been implicated in biological functions such as gene density, recombination and replication timing. Studies aimed at elucidating the boundaries between the individual families have suggested a positive trend in SNP density in isochores of high GC content [107].It is not unusual to find CpG islands within a much larger isochore.

The reader should note however a lack of potential methyl SNPs on chromosome 14; this was probably due to an issue with imputation accuracy (see previous section) or

it may be due to lower number of isochores occurring on this chromosome. The map above (figure 27) shows chromosomes 12, 13, and 14 to have long regions where the GC content is low < 37%.

A second way in which imputed SNPs were assessed was through their potential effect on the secondary structure of annotated miRNAs. A Perl script was written that mapped the location of imputed SNPs to miRNAs currently documented in miRBase. To be considered, SNPs should fall within the mature or the precursor sequence. Once mapped the script then used RNAfold to predict the secondary structure of the miRNA and to predict the overall minimum free energy. If the minimum free energy changes by greater than one degree, the SNP was considered to have a significant impact on the regulatory function of the molecule. Studies targeting genes such as AVPR1 have shown that a one degree change in the miRNA can have a significant impact on the miRNA function [119]. Pictured in figure 28 is example output from this script, demonstrating the potential effects of SNP introduction on miRNA secondary structure. Hsa-mir-1324 pictured in figure 28 was found to have a total of seven SNPs within its structure; six of these SNPs were imputed from the microarray. Four of these SNPs located on the 5’ leg of the hairpin function to raises the negative free energy over two degrees (-32.50˚ to -30.10˚) and to introduce additional bulges within the structure. In the picture, the miRNA is shown with the major allele express in all four of the SNPs of interest; on the right is the same miRNA with the minor allele.

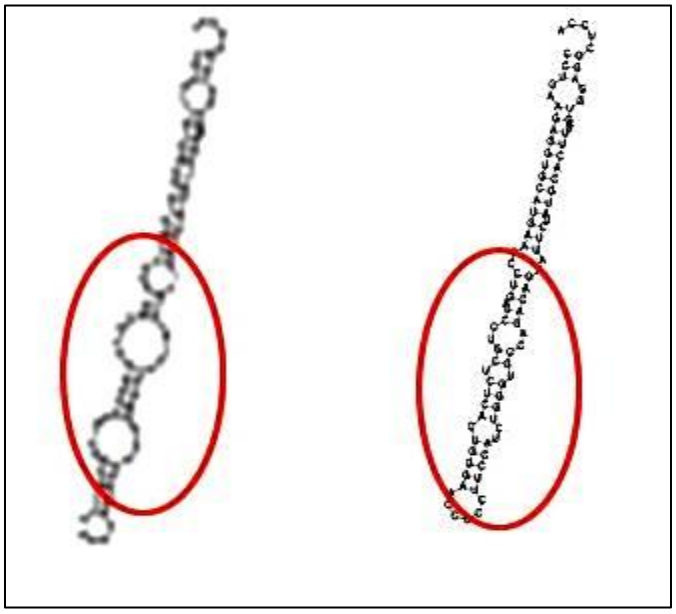


Figure 28 Example output from allele substitution script. The inclusion of four SNPs (circled in red) within the precursor structure of hsa-mir-1324 introduces multiple bulges within the structure and lowering the minimum free energy.

The number of SNPs affecting secondary structure varied dramatically according to chromosome with chromosome 1 having the largest number and chromosomes 11,12, and 14-16 showing none. The reason for this is unclear as to why some of the largest chromosomes, having the greatest number of annotated miRNA are underrepresented. It is possible that the lack of SNPs on chromosomes 11, 12, and 14-16 may be due to the success of the imputation process as a similar underrepresentation is present among the methylated SNPs. The number of miRNA related SNPs already present on the platform (N =25) however showed a similar distribution to that of imputed SNPs (table below).

Chromosome	Sequence Length (in base pairs)	# of miRNAs	version 5 SNPs	Imputed SNPs
1	245203898	135	5	15
2	243315028	101	0	5
3	199411731	77	0	4
4	191610523	57	0	0
5	180967295	69	0	2
6	170740541	56	0	3
7	158431299	70	0	6
8	145908738	71	0	7
9	134505819	72	3	2
10	135480874	67	2	11
11	134978784	76	3	0
12	133464434	62	0	0
13	114151656	37	1	2
14	105311216	89	1	0
15	100114055	58	4	0
16	89995999	56	0	0
17	81691216	88	2	1
18	77753510	31	0	1
19	63790860	111	1	0
20	63644868	39	1	2
21	46976537	19	0	1
22	49476972	38	2	4

Table 10 Genotyped and Imputed SNPs that potentially affect miRNA function through altering its structure.

In addition, when the distribution of the SNPs was examined, a majority of SNPs (85.3%) were found to be located within the precursor sequence, followed by 9.7%

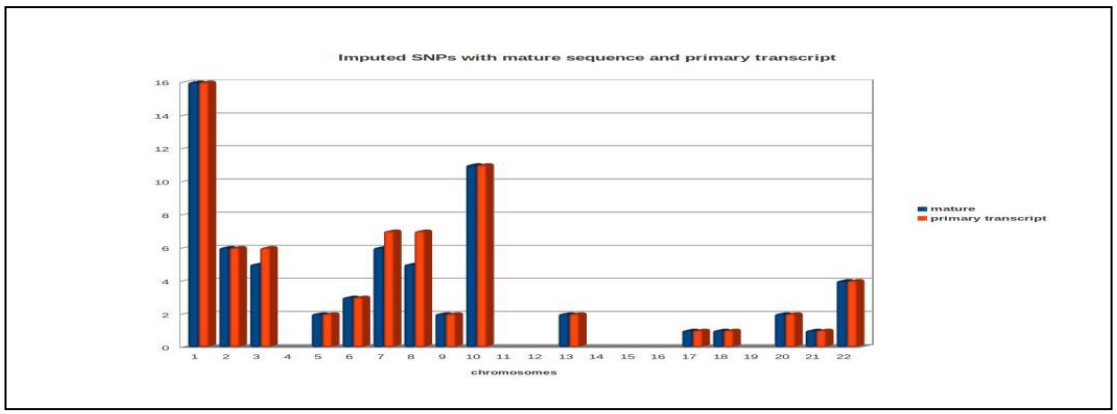


Figure 29 66 imputed SNPs were found to fall within the mature sequence or its precursor, affecting secondary structure. SNPs are included on this graph if their alleles affect a minimum free energy change greater than one degree.

within the mature sequence and 5% within the terminal loop. This finding was expected as SNPs that fall within the mature sequence or the terminal loop of the precursor are thought to have a greater impact on miRNA regulatory function and should be in lesser number.

Chapter Discussion

The GWAS SNP array is dense experimental device that uses the principles of linkage disequilibrium to assess common SNP variation. Linkage disequilibrium, i.e. the nonrandom association of two or more alleles at two or more loci enables even genomic coverage and eliminates the need for testing of every SNP. GWAS design however eliminates SNPs that may possess crucial function and importance. Therefore, imputation of pre-existing GWAS increases SNP resolution and expands the potential knowledge base of information for researchers.

In this project, we were able to use imputation to focus on two areas impacting miRNA regulatory function e.g. methylation and structural changes. We were able to impute a high number of SNPs that appear to fall within either CpG island/isochores and were able to identify sixty-six SNPs that affect miRNA secondary structures. In the first instance, the number of SNPs gained represents a substantial increase over the numbers of SNPs originally genotyped. In the genotyped SNPs, only 15% of probes having a high GC survived quality control filtering (call rate, HWEV). The unusually high number of SNPs that were identified may have been a combination of genotyped SNPs that had previously failed quality control but were “backwards imputed” using the program and novel SNPs. In this regard, imputation of the GWAS enabled us to overcome technical issues unique to high GC regions. In the second instance,

imputation generated an additional 66 SNPs that fell within miRNA structures and for whom there was differential effect based on respective alleles. Compared to the 25 genotype SNPs present on the array (table), this represents a modest increase. The distribution of these SNPs was unusual when one compares chromosome size and number of already present miRNAs.

Chapter Five: Global Discussion

The preceding chapters illustrate ways in which miRNAs can be evaluated experimentally to learn more about their role in the etiology of Schizophrenia and Bipolar disorder. Genetic epidemiology has yielded consistent evidence of an inherited component for each of these disorders and has provided adequate rationale for pursuing studies such as linkage analysis, candidate gene study, GWAS, and expression profiling to learn more about the genetic architecture of these disorders. These studies, however, have often created more questions than they have answered and have had issues of limited replication and consistency. Further, some studies such as linkage analysis were not really appropriate for use in studying these types of complex polygenic disorders. Other studies, i.e. candidate gene association, have been too limited in their approach to ever generate any findings of real value. Only GWAS have had the needed breadth and scale necessary to study Schizophrenia and Bipolar Disorder; the genes that have been replicated with any degree of consistency, e.g. TCF4 and NRGN, have been identified through large scale GWAS marshalling large subject numbers. One such Mega-GWAS performed by the PGC implicates one miRNA in Schizophrenia and Bipolar Disorder, hsa-mir-137 and provides some of the first substantial evidence that miRNAs could be implicated in Schizophrenia.

miRNAs may help to explain the phenotypic variability observed in Schizophrenia and Bipolar Disorder, but before it can integrated into the body of knowledge more must be learned about its biogenesis and relationship to target genes within the body. It is necessary, then, to clearly delineate the complete list of miRNAs

present in the human genome and to further document their relationship with individual gene targets. The human body possesses approximately 20,000 active genes, many of which are functional at only specific times in the human life span. By fine tuning transcription, miRNAs work to maintain homeostasis in the body and sharpen cellular responses to environmental crisis [190,191]. As an example, consider the interactions of methyl CpG-binding protein 2 (MeCP2), brain derived neurotrophic factor (BDNF) with the miRNAs hsa-mir-132 and hsa-mir-212. MeCP2 is a transcription factor which binds to methylated cytosines on CpG dinucleotides in DNA, recruiting transcriptional repressors and effectively stopping gene expression MeCP2 targets to the promoter III region of BDNF which represses expression of this gene. Hsa-mir-212, -132 target MeCP2 limiting its expression which in turn affects the functionality of BDNF. Ultimately, over or under expression of MeCP2 causes the neurodevelopmental defects seen in Rett Syndrome [191, 212].

Detection of miRNAs through Deep Sequencing

Chapter Two of this thesis summarizes the results of a deep sequencing study performed in our lab using the Illumina/Solexa. Deep sequencing and qPCR were used to create a miRNA profile of an affected population and a cell line model used as a proxy for human neural function. The goal of this project was the detection and identification of known and novel miRNAs in neuroblastoma cells. After initial detection, interesting novel candidates were validated in postmortem brain tissue of affected subjects from the Stanley Medical Research Institute. At the time the experiment was first performed, the available length for a typical read was small (~ 36 bases) but still adequate for the detection of miRNA candidates. We achieved our goal in this project of

identifying miRNAs within this cell line though the number was admittedly small. We believe that this small number was in part due to the stringency of the analytical software that was used. Subsequent comparison of popular software in a number of different types of data confirmed our suspicion. Excessive stringency is acceptable if one wishes to detect novel miRNAs for additional testing. Given the costs of validation, one cannot simply afford to waste time or resources on novel candidates that cannot be easily verified in the lab. Initially in our study, we identified 25 novel miRNA candidates in neuroblastoma cells. Seventeen of this number (68%) was found to be predicted by both miRDeep and miRanalyzer in a program comparison. Using that number of miRNAs (17), we were able to validate 12 successfully both in the original cell line and in post mortem brain tissue representing a 70.6% success rate overall. Comparison of the success rate of this multi-program approach to ones used in other studies is difficult as most do not choose to validate all of their predictions [63,210].

Our results from the deep sequencing study also suggested the possibility of additional novel miRNAs being derived from alternative biogenic sources such as snoRNA molecules. This finding proved simultaneously the strength of deep sequencing as a technology for discovery as well as the need for improved operationally defined criteria for the detection of novel miRNA candidates. In deep sequencing, a large number of reads are routinely set aside as unmapped by analytical software. This elimination of reads mapping to regions traditionally deemed unlikely to house miRNAs, i.e. snoRNAs, necessarily reduces the effective amount of information gained from deep sequencing technology. We used the eliminated portion of our data to predict the presence of these sno-derived novel candidates and were able to successfully confirm

their presence in the original cell line and a RNA tissue panel. So, by using an expanded definition of the possible biogenic pathways for miRNAs, we were able to better utilize the technology ultimately making it more efficient and effective in its intended purpose.

Also, in this deep sequencing study, we detected a novel miRNA (PRD5) which could potentially play a role in the regulation of the genes ZNF804a and C10orf26. Originating from traditional biogenic pathways, this novel miRNA was shown to be differentially expressed in the postmortem brain tissue of affected subjects from the SMRI. This finding must be replicated in other affected groups and its status should be expanded using other laboratory techniques such as a luciferase test and/or western blotting before it can be stated that PRD5 is a miRNA involved in Schizophrenia. Nevertheless, the correlation of the relative expression profiles of PRD5 with ZNF804A and C10orf26 suggests that this miRNA affects these genes in a different fashion, with ZNF804a being predicted to increase its expression and C10orf26 to exhibit a decrease. Target predictions generated by miRanda, Pita, and TargetScan for ZNF804A and C10orf26 intuitively mirror the observed differences in expression direction in the post mortem brain tissue. PRD5 was predicted by these programs to target the 5'UTR of ZNF804A and not the 3'UTR (see figure 18). Conversely, PRD5 was predicted to target the 3'UTR of C10orf26 and not the 5'UTR. There have only been a limited number of accounts suggesting that miRNAs target 5'UTR of genes but when they have been identified, increased mRNA expression was observed. A recent letter examining this lack of 5' targets suggests that it may be due to a predominant focus on cross species conservation and the exclusion of less-conserved sites as unreliable [114]. Further,

some assert that the strength of the effect is diminished if the miRNA targets the 5'UTR thereby making successful detection more difficult to achieve [192]. Regardless, the biological implication of 5' targeting of PRD5 on ZNF804A in this project is intriguing and makes this miRNA worthy of additional study.

Selection of predicted gene targets of individual miRNA

Chapter Three explores the issues encountered when selecting appropriate targets of specific miRNAs for experimental validation and testing. In this chapter, miRNA profiling of postmortem brain tissue detected seven differentially expressed miRNAs in subjects from the Stanley Medical Research Institute. Targets were predicted using miRanda against the 5' and 3'UTRs of all known genes (GRCh37). Filtering and selection of targets for each miRNA was first performed using an FDR threshold and then a second time using a more biologically attuned approach incorporation the effects of alternative splicing, target gene co-expression levels, and mRNA site accessibility into the process. Stringent FDR screening of these targets however generated inconsistent results for three miRNA out of the seven tested (hsa-mir-132, -212 and -154*) when one considers the known biology of each. First, the FDR results suggested that hsa-mir-154* had the highest proportion of credible targets in the predicted list. hsa-mir-154* is a star sequence which, because of its low native concentration, would be unlikely to possess so many targets. Most star sequences are degraded following their separation from the mature sequence and only are found in wild-type cells in trace levels [59]. Second, hsa-mir-132 and -212 share seed sequences which would suggest that a similar number of targets should survive FDR filtering. This was, however, not the result of FDR filtering, as the filtered targets for hsa-mir-212 were more numerous

than that of hsa-mir-132. The seed sequence is seen as an essential factor for miRNA:mRNA binding, so important that many think that the seed is all that is necessary for binding to occur (see chapter 3 for a longer description of this process). Many prediction algorithms such as miRanda weigh more heavily the degree to which the seed sequence maps to the prospective binding site than the rest of the mature sequence. Therefore, if two miRNAs share a seed sequence one would expect a similar number of predicted targets to be generated.

The algorithms used in current miRNA target prediction attempt to replicate the interaction between the mature sequence and the gene's untranslated region (5' or 3' end). The success with which they do this is however questionable as most algorithms are thought to have extremely high false positive rates. The miRNA research community has attempted to compensate for this inadequacy by introducing a greater number of biological parameters into the target prediction process [51, 53, 151]. Unfortunately, the number of experimentally documented interactions is currently too limited on which to build credible target prediction algorithms with which we can evaluate using statistics. Further, the sheer number of possible targets for any one miRNA is cost prohibitive to allow for complete experimental validation and strict assessment of false positive rates.

One of the underlying assumptions regarding the use of FDR is independence of individual tests; in this case, this assumption was most certainly violated by the occurrence of multiple binding sites within a single gene. Multiple binding sites can occur within a gene making those sites interdependent and inflating the FDR statistic [114].

Nevertheless, uniform bias free approaches to target prediction are needed if the

field is to advance. Other approaches can be employed including permutation or an empirical p value but these approaches can be costly to perform in the lab and will only be suited to use with specific miRNA. We suggest that a uniform bias free approach to target prediction can be routinely implemented only after more is known about the stochastic factors present within the cell that influence miRNA binding.

SNPs which impact miRNA functionality

Finally, chapter four examines how single nucleotide polymorphisms could potentially impact the regulatory function of a miRNA in two ways, through the promoter and through the structure of the precursor and the mature sequence. By imputing the HapMap3 reference panel onto the Affymetrix version 5 SNP framework, we were able to nearly double the number of markers potentially yielding information on methylation and miRNA structure. Over 2.5 million additional SNPs were imputed from the HapMap3 reference panel onto this framework, demonstrating the utility of imputation for increasing GWAS resolution. Despite a low number of arrays with which to work, we were able to identify 66 SNPs that appear to impact miRNA secondary structure. This number was in addition to the 25 SNPs that were already present on the platform. This number is encouraging given the assumed impact that a single SNP has on miRNA secondary structure and suggests that miRNAs might be more variable that previously thought.

Certainly, SNPs can affect miRNAs in a variety of other forms, e.g. SNPS within transcription factors, Drosha and Dicer enzymes but the two types of SNPs profiled in this project arguably has the greatest immediate impact on miRNA function. Indeed, changes in the structure of a single miRNA can radically shift the number of genes it

targets. For example, new gene targets can be created if SNPs fall within the seed sequence, by potentially shifting the sequence composition. Furthermore, the lack of experimentally verified miRNA: mRNA interactions make it difficult to assess the impact that SNPs within target sites might have. A limited number of SNPs in targets and miRNAs have been identified as being significantly associated in Schizophrenia. These studies have been hampered by a lack of experimentally proven targets, small sample size, and limited focus [59, 192]. These significant SNPs, as a result of these factors generally have not withstood correction for multiple testing. This is to be expected considering the small effect size individual SNPs has on the disorder. Evidence coming from previous genetic research, i.e. GWAS studies is quite clear about the size of the effect size of individual SNPs (median Odds ratio: ~1.33) [192,193]. It is no surprise, therefore, that SNPs associated with miRNAs in cases to also have similar effect sizes. It is necessary therefore to assess the role of SNPs genome-wide to truly determine their impact on miRNA function.

Closing Remarks

Corvin suggests that, in the future, genetic parameters will likely replace parts of the DSM-IV proven inefficient in the diagnosis of Schizophrenia and Bipolar disorder [95]. Psychiatric genetics however has only recently built a large enough knowledge base to accomplish this task. Previous techniques such as linkage analysis and candidate gene studies have only been able to provide a small piece of information necessary to fully comprehend the genetic structure. Even GWAS proponents admit that in order for the GWAS to be truly effective, the number of subjects tested must increase. We suggest, too, that in order to truly contribute to the current body of knowledge more

must be known about the function of miRNAs and in a greater number of subjects. Recent findings of alternative modes of biogenesis would suggest that we still do not know enough about its sources In addition, target prediction and validation is a slow process, hampered by algorithms with inherently high false positive rates. In order to improve current software and reduce these rates, we must simply know more about miRNA:mRNA interaction event, including indirect factors which might influence binding success. Low-cost, high throughput methods of target validation would be of tremendous use in this regard as they would allow

Limitations

Admittedly the projects described in Chapters 2, 3, and 4 have a number of clear limitations. First, MicroSeq data lacks a coherent system of normalization; and without it any statement of observed values cannot be made with any degree of certainty. Expression levels are simply not as straightforward as reads per kilobase per million (RPKM) and can be impacted by a number of experimental biases [20]. Further, without normalization, observed expression values cannot be compared between studies and laboratories. The lack of statistical quality control for MicroSeq is perhaps indicative of the newness of the field. Originally, the cost of this procedure was the rate limiting factor on the number of studies performed and any data generated from this technology was seen as precious. Lowering costs and advances in protocol have made MicroSeq more accessible; researchers are becoming aware of a need for a strong statistically based analytical foundation. Originally, recognizing the need for validation, researchers have employed qPCR to compensate for the deficiencies of deep sequencing. In this thesis in chapter two, qPCR was used in this capacity to determine novel miRNA levels and

validate presence [65].

A second limitation of this project is the high false positive rate of miRNA target prediction software. Indeed, this limitation is well known in the field of miRNA research overall. Early work in this field focused exclusively on the detection of probable candidates, setting aside the more important issue of functionality. In fact, early rules for the validation of a new miRNA candidate included:1) the presence of a hairpin structure (see figure 2,3), 2) the presence of the mature sequence within one leg of the hairpin and not within the loop of the structure and 3) the detection of a size appropriate fragment experimentally [93]. As a result, the miRNA research is now filled an excess of annotated miRNA without proven targets. In designing the original software, researchers had to rely on what they believed to be true about miRNA:mRNA interaction. With added experimentation, we are realizing that the picture is more complicated than previously thought and that factors such as co-expression, mRNA site accessibility play a role. A heavier emphasis on establishing targets for miRNA therefore will enable us to refine prediction software through more providing biological correlates that can be factored into prediction algorithms.

A third limitation to this project surrounds the probabilistic nature of SNP imputation. Certainly one can attempt to ensure that the genotypes on which SNP is based are as correct as possible but imputation is a probability and can be subject to error. Comparisons between programs used to generate imputation have demonstrated that it can be affected by features as linkage disequilibrium and minor allele frequency and that some programs perform better in certain situations than others. A small, limited reference panel and number of microarray most likely contributed to the inconsistency

of the findings documented in this chapter.

Unfortunately, one of the problems inherent with working postmortem brain tissue is its limited number of available samples. Consortiums similar to the ones devoted to GWAS studies (described in the first chapter of this thesis -genetic studies section) could conceivably be used to increase sample size and increase confidence in data findings. Other problems then must be addressed such as inter-site variability and batch effects. Imputation can be used to facilitate cross platform/cross group comparability but we should always be aware the effect of technical variation in this studies and work to create methods that minimize their effects.

Chapter Six: Future Directions

This work and others like it only present a glimpse of the potential that these technologies may yet have in the field of personalized genomics. A necessary part of this sort of research is vigorous experimental validation. High-throughput technologies such as deep sequencing constantly are being improved with better resolution and added read length. Third generation sequencing promises to eliminate potential biases created by PCR and library construction as well as providing larger read length. As this field matures, the analytical techniques will also mature, generating greater confidence in expression levels. Recent advances in this field include the realization that T4 ligase has a biased effect on library construction, artificially affecting read counts [20, 21, 65, 68].

Recent improvements to qPCR techniques include the addition of locked nucleic acid (LNA) based oligonucleotides to expression platforms. Originally created in 1998, LNA technology facilitates melting temperature (Tm) normalization, increases thermal stability of duplexes and increases target specificity. Faster acting and more efficient polymerases such as fusion taq polymerase have also been introduced that decrease reaction time, enabling researchers to assay more targets in a shorter amount of time. Also, additional means by which data are normalized such as quantile normalization, delta delta CT, and the rank invariant approach afford researchers a better set of tools with which to identify and compensate for experimental biases.

One future direction for this project is the need for additional testing of the identified novel miRNAs. In particular, there must be additional experimental work, i.e. luciferase assay and western blots, proving their effect on predicted targets. In the case

of the sno-derived miRNA candidates, one must also prove that they are actively processed similarly to the classically derived specimens. To do that, the sequence for the predicted mature miRNA, its precursor and the entire gene must be expressed into an appropriate cell line and the subsequent expression of each sequence type must be assessed. Alternatively, one might also isolate the individual cell components with ultracentrifugation and verify the native expression of molecules with qPCR. If these molecules are being processed by the cell's machinery as suspected, one might expect to see high levels of the active mature sequence in the cytoplasm.

As for the novel miRNA (PRD5), targets for this miRNA must be verified through luciferase assay and western blot assay. A future direction for this miRNA might not only be the validation of targets but also an exploration of the synergistic effects of this miRNA with miRNA such as hsa-mir-137. In TCF4, the predicted binding site for PRD5 is 1103 bases away from the site predicted for hsa-mir-137. Conceivably, PRD5 can be transfected both in combination with hsa-mir-137 and separately to see if it 1) targets the same genes with the same magnitude and 2) PRD5 and miR-137 act synergistically on their mutual targets.

This is a necessary step if the field is to progress toward a more realistic view of true mechanism of miRNA affecting gene pathways. Even with the findings generated by expression profiling, the general trend has been to investigate one miRNA at a time, validating targets and then focusing on pathways. Admittedly this is more efficient than examining each gene at a time, but it does not take into account that miRNA often function in consort, with multiple miRNA targeting a single gene simultaneously. By examining the effect of two or more miRNA on gene targets we can start to create a

more realistic picture of miRNA regulation on translation. Once the synergistic effects of miRNA are established in relevant cell lines, then their effects can be explored in animal models.

Lastly, one final future direction is the use of multivariate analytical approaches to identify common classes of miRNAs with similar functional patterns. The field is only now becoming technically advanced to permit shift from a simple detection to asking globally more important questions of functionality. Perhaps in the future, we will be able to successfully integrate miRNAs in other psychiatric genetic datasets to achieve a complete view of the structure of Schizophrenia and Bipolar Disorder.

Appendix 1: Databases Consulted in This Project

Database	location	Focus	Chapters
miRBase	www.mirbase.org	**miRNA database containing annotated sequences**	**2,3,4**
Genome Browser	www.genome.ucsc.edu	**General genomics database housed at University of California, Santa Cruz**	**2,3,4**
snoRNABase	www.snorna.biotoul.fr	**Human, Yeast snoRNAs database; Contains sequences**	**3**
Human Methylome	www.neomorph.salk.edu/human_methylome	**Housed at Salk Institute, methylation of fetal fibroblasts and H1 cells**	**4**
RFAM10.1	www.rfam.sanger.ac.uk	**RNA families, sequence, and covariance models**	**2**
Stanley Medical Research Institute Online Genomics Database	www.stanleygenomics.org	**Study of Schizophrenia and Bipolar Disorder. Holds 988 arrays across 6 different platforms**	**4**
SZGene	www.SZgene.org	**Variation associated with Schizophrenia; Holds 1727 studies and 287 meta-analyses**	**2**
PDGene	www.pdgene.org	**Variation associated with Parkinson's Disease; Holds 876 studies and 889 meta-analyses**	**2**
MSGene	www.msgene.org	**Variation associated with Multiple Sclerosis; Holds 789 studies and 324 meta-analyses**	**2**
Cancer 500	www.variantgps.nci.nih.gov/cgfseq/pages/snp500.do	**Re-sequencing of 102 reference samples from four ethic groups**	**2**
HapMap	www.hapmap.ncbi.nlm.nih.gov	**Single nucleotide polymorphisms**	**4**
Aceview	www.aceview.ncbi.nlm.nig.gov	**Alternative splicing based on EST database**	**3**
Biomart	www.biomart.org	**Central clearinghouse for multiple datasets and species**	**2,3,4**

Appendix 2: Programs Used in This Project

Program Name	Type	Focus	Chapters
miRDeep/miRDeep2	Executable	miRNA prediction	2
miRanalyzer	Web	miRNA prediction	2
DSAP	Web	miRNA prediction	2
Oligomap	Executable	mapping and alignment	2
Bowtie	Executable	mapping and alignment	2
Flux Capacitor	Executable	simulation of RNAseq datasets	2
BLAST	Web	mapping and alignment	2
RNAFold	Executable	Prediction of miRNA secondary structure	2,3
David	Web	Functional annotation of gene lists	2,3
IMPUTE2	Executable	SNP Imputation	4
GTOOL	Executable	Manipulation of imputed genotypes	4
Shape-IT	Executable	Genotype Phasing	4
Affymetrix PowerTools	Executable	Microarray quality control and genotype prediction	4
Beagle Call	Executable	Genotype prediction using haplotypes	4
SNPTest	Executable	SNP quality control and tests of association	4
miPred	Web	Prediction of hairpin (miRNA) status	2
miRanada	Executable	miRNA target prediction	2,3
PITA	Executable	miRNA target prediction, based on site accessibility	2,3
TargetScanS	Web	miRNA target prediction, based on seed conservation	2,3
StringDB	Web	Protein protein networks	2,3
LinRegPCR	Executable	Assessment of PCR specificity and quality	2,3
Perl	Executable	Free-form scripting language	2,3,4
R	Executable	Statistical scripting language	4
PLINK	Executable	Genotype manipulation, association test, Hardy Weinberg violations	4
SNPInfo	Web-based	Functional prediction of SNPs and their proxies	4
PolyPhen	Web-based	Allelic impact on protein structure	4
Haploview	executable	Identification of proxy SNPs through linkage disequilibrium	4

Appendix 3: Known miRNAs predicted by deep sequencing in Neuroblastoma

#miRNA	read count	#miRNA	read count
hsa-let-7a	661	hsa-miR-30a	2621
hsa-let-7b	635	hsa-miR-30c	226
hsa-let-7e	126	hsa-miR-30d	2840
hsa-let-7f	2472	hsa-miR-30e	921
hsa-miR-1	1893	hsa-miR-3200-5p	68
hsa-miR-101	141	hsa-miR-320a	2921
hsa-miR-103	4730	hsa-miR-330-3p	1227
hsa-miR-106a	1632	hsa-miR-342-3p	330
hsa-miR-106b*	78	hsa-miR-361-5p	65
hsa-miR-107	2560	hsa-miR-365*	384
hsa-miR-10a	393	hsa-miR-3662	68
hsa-miR-10b	64	hsa-miR-421	165
hsa-miR-1254	152	hsa-miR-423-3p	4428
hsa-miR-1255a	1146	hsa-miR-423-5p	9230
hsa-miR-1269	160	hsa-miR-503	113
hsa-miR-1285	108	hsa-miR-505*	63
hsa-miR-1292	61	hsa-miR-548h	211
hsa-miR-1301	968	hsa-miR-641	92
hsa-miR-140-3p	953	hsa-miR-7	35355
hsa-miR-146b-5p	258	hsa-miR-744	276
hsa-miR-148a	1443	hsa-miR-877	291
hsa-miR-148b	403	hsa-miR-9	386
hsa-miR-148b*	194	hsa-miR-92a-1*	1916
hsa-miR-151-3p	57	hsa-miR-92b*	385
hsa-miR-15a	94	hsa-miR-93	98
hsa-miR-17	177	hsa-miR-941	271
hsa-miR-17*	6206	hsa-miR-34c-5p	47
hsa-miR-181a	5288	hsa-miR-99b*	42
hsa-miR-181b	4121	hsa-miR-425	41
hsa-miR-181c	134	hsa-miR-548u	41
hsa-miR-181d	944	hsa-let-7c	40
hsa-miR-182	2064	hsa-miR-574-5p	40
hsa-miR-186	172	hsa-miR-16-2*	39
hsa-miR-191	5385	hsa-miR-148a*	38
hsa-miR-192	95	hsa-miR-3131	38
hsa-miR-193b*	99	hsa-miR-378	36
hsa-miR-194	86	hsa-miR-548j	36
hsa-miR-196a	467	hsa-miR-320b	35

hsa-miR-19b	252	hsa-miR-128	34
hsa-miR-21	458	hsa-miR-181a*	34
hsa-miR-2110	150	hsa-miR-25*	34
hsa-miR-221	2226	hsa-miR-30a*	33
hsa-miR-221*	1636	hsa-miR-3179	33
hsa-miR-222	22997		
hsa-miR-26a	2108		
hsa-miR-27b	734		

Appendix 4: Normalized Cq Values for novel miRNA validated in postmortem tissue of SMRI

subject/sample id	MiR-1	MiR-5	miR-6	miR-7	MiR-14
1	1.01894045	0.84629946	1.46977941	1.24387353	0.95019166
2	2.39271567	1.61832298	1.96196511	1.17405489	0.79900177
3	1.00594383	0.86830881	1.36958586	1.05946795	0.68493109
4	1.12914413	1.64978674	0.94386999	1.57730274	1.01970397
5	0.56817619	0.6803742	0.79879666	5.0727868	0.92610679
6	1.3428047	1.05268949	0.72083466	1.98731336	1.26838277
7	0.76332535	0.94386999	1.53732225	0.89089055	1.18191825
8	0.43473533	0.61119269	0.19750214	0.26876551	0.84811352
9	0.67136541	0.83015933	0.56845597	0.5540471	0.96246797
10	1.62794202	1.71456133	1.45103233	1.09402061	1.10134793
11	0.34439372	0.69806841	0.27700831	0.38925321	0.93207016
12	1.69185893	1.12970014	0.89662716	2.21643227	0.53188526
13	0.59048415	0.60226088	0.35580062	0.52294843	0.90205324
14	0.459721	0.5162782	0.29918735	0.50319191	1.02232534
15	0.65016155	0.82484799	1.09402061	9.2148328	1.11701015
16	0.69772483	0.65048171	0.89662716	1.13697449	1.07343165
17	0.72511918	0.65888582	0.73958111	1.0392624	0.8876941
18	0.3672236	0.49043731	0.17789052	0.40976277	0.81820405
19	0.44806827	0.62590709	0.1971303	0.57580032	0.95815332
20	0.49019593	0.59840763	0.349015	0.3030528	1.12204015
21	0.46268122	0.64218479	0.22127331	0.42041929	0.96061646
22	0.50294424	0.6339937	0.1744979	0.73014772	0.72148269
23	0.79840351	0.73014772	0.2142848	7.74861329	0.46006548
24	0.49019593	0.40194801	0.40976277	0.52970483	0.71273748
25	0.67568846	0.56845597	0.64218479	0.52294843	0.92018158
26	1.40451249	1.05946795	1.25994416	0.55761471	0.73080413
27	1.09348216	1.30103499	0.57211636	0.96841681	0.9837028
28	0.80871874	1.47924359	0.68475526	0.88519065	0.94653939
29	1	1.10815517	0.35809169	1.7591512	1.36467189
30	1.58667796	1.12247234	3.29975502	2.24506819	0.90611527
31	1.0454396	0.92586905	1.23591524	0.67602118	0.98749847
32	0.85132972	0.91405956	0.81432702	0.36977021	1.17661983
33	4.91015415	2.16025153	3.34238728	20.8211414	1.86185128
34	0.98675906	1.03261322	0.75396019	0.77356811	1.14605976
35	0.52944412	0.42585104	0.45994737	3.58690328	0.73977101
36	0.98866096	0.72547625	0.19586907	0.47190903	0.80414669
37	0.9785597	0.9024007	2.79258089	0.94994774	1.59604129
38	0.6093505	0.86275338	0.62993742	0.64631993	1.10488813
39	0.70855614	1.06629006	0.73014772	0.6803742	1.20721968
40	2.53663319	2.33321496	5.00808324	1.88784383	1.44768334
41	1.18863816	1.34346593	1.37840488	12.1437134	1.38319073

42	0.82391301	0.71622278	1.20458803	0.90821142	1.07619114
43	0.55484176	1	0.39937636	0.3445633	0.9786646
44	0.75358911	0.68916452	0.95606462	0.54698021	0.89743322
45	1.94845303	1.93694018	1.36958586	1.47924359	0.94532509
46	0.87909432	1.50800334	0.7443434	0.86275338	0.84380442
47	0.45444023	0.53654853	0.39682115	0.38925321	0.80363071
48	0.65016155	0.42859317	0.25978701	1.85183998	0.74262545
49	1.11474186	1.0392624	3.17509323	3.63324545	1.42098574
50	0.5829525	0.69806841	0.78356248	1.58745928	0.70636155
51	0.52944412	0.75396019	0.85174894	0.29727316	0.99258214
52	2.66002328	1.6711017	2.00011004	0.86830881	1.15863349
53	1.04008521	1.09402061	2.31828709	1.33487045	0.88201464
54	2.9007959	2.05212605	2.65280589	7.26689068	0.86242119
55	0.87684022	1.151664	1.0392624	0.4124013	1.2913841
56	2.59759569	2.02595111	2.06534006	1.1968811	1.21265589
57	0.78771408	0.71622278	1.11529078	1.8282197	1.39209735
58	1.01437264	0.78356248	0.71164039	0.68916452	1.15714709
59	2.97814659	2.55258526	3.23682403	3.34238728	1.10488813
60	1.01632776	0.77854926	1.31784414	9.45447904	0.96866548
61	0.82021949	1.03261322	0.40453623	0.59077492	0.95263433
62	0.79942908	1.0459544	0.46888976	2.92091227	1.19105682
63	1.93350321	2.0920239	1	0.9622209	1.89318293
64	0.72838446	0.81171782	0.87670909	0.34789671	1.20180784
65	0.32506289	0.79879666	0.72083466	0.43973941	1.03022994
66	1.45124935	1.11529078	1.92454767	0.72028984	0.727528
67	1.34539284	1.53732225	3.13459481	9.70035766	1.54761863
68	1.05082157	1.42793437	0.92884519	0.69583171	1.20026606
69	3.69968534	3.03559431	5.27962315	2.70438228	1.34208694
70	0.91949258	1.2201511	0.5172952	0.98092859	0.99321944
71	0.74397705	0.52970483	0.25978701	1.7591512	0.83894408
72	***5322437.7***	0.55761471	1.01291983	0.45408073	0.72519691
73	1.03973478	1.02009674	1.82939353	1.14503038	1.44026878
74	0.41193384	0.77356811	0.40976277	1	1.22753505
75	7.88524667	5.17141295	14.7223652	28.884767	2.98803989
76	3.07527354	2.10549482	5.54973332	1.78187913	1.2118778
77	1.057588	0.89662716	2.23070428	1	0.8208341
78	9.97970849	3.84930711	12.0660181	21.0901469	2.97655471
79	0.82709212	0.90821142	1.48876871	0.98724497	1.03380692
80	1.70494036	1.43252436	2.03899658	1.63923143	1.36493469
81	1.48136525	1.35211675	1.45103233	0.77356811	1.10836857
82	1.8652399	1.30941259	0.59077492	1.62874365	0.80296043
83	0.97417292	0.83550488	0.90821142	0.57580032	0.72575569
84	0.19817477	3.03559431	4.29309673	2.56902182	0.94508242
85	1.13714454	1.17405489	0.65048171	0.90821142	1.2629399
86	0.41325797	0.66739851	0.94386999	1.2201511	1.1165084
87	0.44749345	0.72547625	1	4.49038343	0.64985574
88	2.812788	1.64978674	5.58546906	14.534581	2.14367665

89	2.51555533	2.55258526	2.13269745	1.77047869	1.140556
90	1.32908512	0.61397016	2.3938939	0.80911697	0.86619355
91	0.99566599	0.96841681	1.58745928	1.36958586	1.22289524
92	0.66493265	0.71622278	0.3030528	0.45117553	1.18275302
93	1.7673371	1.30103499	3.25766653	2.61896926	1
94	1.0474546	0.97465262	1.25188305	0.75881508	1.10425006
95	6.02197255	3.40737065	3.9748454	6.47400423	0.84239744
96	1.0454396	0.79879666	0.80394026	0.49677368	0.85607882
97	7.07656863	4.0610273	7.56879777	10.5676586	1.27935458
98	4.825794	1.86376432	3.70388374	2.84687481	0.82347261
99	1.87123557	1.26805717	0.98724497	0.94994774	1.22517363
100	0.77219547	0.64631993	1.62874365	0.65888582	1.04958598
101	2.10175876	1.32632997	2.47196648	1.08006634	0.71824836
102	1.30960843	1.06903119	2.91162806	1.45476252	1.44582613
103	0.31078113	0.9024007	1.54730874	1.34346593	0.97302746
104	1.71481777	1.37840488	1.91223445	1.4052041	0.65420822
105	2.0275551	1.11529078	0.10782637	0.84088485	1.16909183

Appendix 5: Example code for functions performed in thesis

A. Code written in R

1. Linear Regression

```
library(MASS)
data<- read.table (“CQvalues.txt”)
summary (data)
summary(ols <-lm(RIN~CQ + diagnosis, data = cdata)
opar<-par(mfrow = c(2,2), oma = c(0,0,1.1, 0))
plot(ols,las =1)
par(opar)
summary(rr.huber <-rlm(RIN~CQ + diagnosis, data = cdata))
#huber weights
Hw<-data.frame(state=cdata$subject, resid = rr.huber$resid, weight = rr.huber$w)
Hw2 <- hweights[order(rr.huber$w), ]
```

2. False Discovery Method: Benjamini Hochberg Method

```
data<-read.table (“pvalues.txt”)
p.adjust(data, methods = “BH”, n = length(p))
```

B. Code written in Perl

1. SNP Haplotype Phasing and Imputation

```
use strict; use warnings;
my $genfile = shift; my $chromosomenum = shift; my $intervalsize = shift;  my
    $chromosomelength = shift;
if (!defined ($chromosomelength)){
    die "name of gen, chromosome, intervalsize, and chromosomelength\n";
    }

my ($start,$div,$end, $pos);
$div = int("$chromosomelength"/"$intervalsize");
my $j = 0; $start = $j +1;
$pos = $chromosomelength - $intervalsize;
open OUT, ">chromosize.txt";
    while ($j < $div){
            if ($start > $pos){
            print OUT "$pos\t$start\n";
            }else{
            print OUT "$start\t$pos\n";
            }
            $start = $pos; $pos = $pos -$intervalsize; $j = $j + 1;
            }
```

```
    close OUT;
my $i = 0;
open FH, "chromosize.txt";
while ($i < $div){
    while (my $line = <FH>){
    chomp $line;
    if($line =~/(\S+\s+\S+)/){
    my $in = $1;
    my $section = "$i$genfile";
    print "\nphasing chr$chromosomenum $in\n";
`impute2 -phase -m chr15.1000K.map -g $genfile -strand_g chr15.strand -int $in -o
    $section `;
    $i = $i +1;
    }}}

use strict; use warnings;
my $map = shift; my $ref = shift; my $leg = shift; my $int = shift; my $out = shift;
if (!defined($out)){
    die "map,ref,legend,intervals,out\n";
}
my ($start, $end, $diff);my $j = 1;my $pattern = "chr15.calls_haps";
open VAL, $int;
    while (my $line = <VAL>){
    chomp $line;
            if ($line =~ /(\d+)\s+(\d+)/){
            $start = $1;
            $end = $2;
            $diff = ($end - $start);
            if ($diff eq 5000000){
                    #print "$start\t$end\n";
                    print "$j$pattern\n";
            `impute2 -m $map -known_haps_g "$j$pattern" -h $ref -l $leg -int $start
    $end -o "$j$out" `;
             $j = $j + 1;
    }else{
}}}
```

2. miRNA Target Prediction

```
use strict;use warnings;
my $file = shift; my $header =
    "query\tgene\tMFE\tmatchlen\tpercent\tqueryseq\treferenceseq\n";
print "$header"; my ($queryname, $genename, $matchlen, $querymatchper,
    $refmatchper, $MFE, $querysequence, $refsequence);
open FH, $file;
    while (my $line = <FH>){
```

```
        chomp $line;
        if ($line =~/^\>(\S+)\s+\d+\:.*(ENSG\d+):\d+:\d+\s+\d+\.\d+\s+.*\%/){
                print "\n$1\t$2\t";
                }elsif ($line
=~/Forward:\s+Score\:\s+(\d+\.\d+)\s+Q\:\d+\s+to\s+\d+.*Align\s+Len\s+\((\d+)\)\s
+\((\d+\.\d+)\%\)\s+\((\d+\.\d+)\%\)/){
                print "$1\t$2\t$3\t$4\t";
                }elsif ($line =~/Energy:(.*)/){
                print "$1\t";
               }elsif ($line =~/Query.*\s+(3'.*)/){
                print "$1\t";
                }elsif ($line =~/Ref.*\s+(5'.*)/){
                print "$1\t";
                }else{}}
```

3. Assessment of SNPs within CPG islands

```
use strict;use warnings; my $cpgisland = shift; my $mirna = shift; my $chromosome
    = shift;
if (!defined($chromosome)){
    die "name of cpgisland file, mirna file, and target chromosome\n";
    }
open MIR, $mirna;
    while (my $line = <MIR>){
        chomp $line;
        if ($line =~/(hsa\-mir\-\S+)\s+(\S+)\s+(\S+)\s+(\S+)\s+(\S+)/){
            my $mir = $1;
            my $mirchrom = $2;
        if ($mirchrom eq $chromosome){
            my $start = $3;
            my $end = $4;
            open CPG, $cpgisland;
            while (my $line = <CPG>){
                chomp $line;
                if ($line =~/(\S+)\s+(\S+)\s+(\S+)/){
                my $chrom = $1;
                my $cpgstart = $2;
                my $cpgend = $3;
        if (($mirchrom eq $chrom) && ($start > $cpgstart) && ($start < $cpgend)){
        #if (($mirchrom eq $chrom) && ($start < $cpgstart) && ($end > $cpgstart)
    && ($end < $cpgend)){
        # if (($mirchrom eq $chrom) && ($start > $cpgstart) && ($start < $cpgend)
    && ($end > $cpgend)){
                print
    "$mir\t$mirchrom\t$start\t$end\t$chrom\t$cpgstart\t$cpgend\n";
                                                }}}}}}
```

Appendix 6: Known miRNAs identified in Neuroblastoma

#miRNA	read count	#miRNA	read count
hsa-let-7a	**661**	**hsa-miR-30a**	**2621**
hsa-let-7b	**635**	**hsa-miR-30c**	**226**
hsa-let-7e	**126**	**hsa-miR-30d**	**2840**
hsa-let-7f	**2472**	**hsa-miR-30e**	**921**
hsa-miR-1	**1893**	**hsa-miR-3200-5p**	**68**
hsa-miR-101	**141**	**hsa-miR-320a**	**2921**
hsa-miR-103	**4730**	**hsa-miR-330-3p**	**1227**
hsa-miR-106a	**1632**	**hsa-miR-342-3p**	**330**
hsa-miR-106b*	**78**	**hsa-miR-361-5p**	**65**
hsa-miR-107	**2560**	**hsa-miR-365***	**384**
hsa-miR-10a	**393**	**hsa-miR-3662**	**68**
hsa-miR-10b	**64**	**hsa-miR-421**	**165**
hsa-miR-1254	**152**	**hsa-miR-423-3p**	**4428**
hsa-miR-1255a	**1146**	**hsa-miR-423-5p**	**9230**
hsa-miR-1269	**160**	**hsa-miR-503**	**113**
hsa-miR-1285	**108**	**hsa-miR-505***	**63**
hsa-miR-1292	**61**	**hsa-miR-548h**	**211**
hsa-miR-1301	**968**	**hsa-miR-641**	**92**
hsa-miR-140-3p	**953**	**hsa-miR-7**	**35355**
hsa-miR-146b-5p	**258**	**hsa-miR-744**	**276**
hsa-miR-148a	**1443**	**hsa-miR-877**	**291**
hsa-miR-148b	**403**	**hsa-miR-9**	**386**
hsa-miR-148b*	**194**	**hsa-miR-92a-1***	**1916**
hsa-miR-151-3p	**57**	**hsa-miR-92b***	**385**
hsa-miR-15a	**94**	**hsa-miR-93**	**98**
hsa-miR-17	**177**	**hsa-miR-941**	**271**
hsa-miR-17*	**6206**	hsa-miR-34c-5p	47
hsa-miR-181a	**5288**	hsa-miR-99b*	42
hsa-miR-181b	**4121**	hsa-miR-425	41
hsa-miR-181c	**134**	hsa-miR-548u	41
hsa-miR-181d	**944**	hsa-let-7c	40
hsa-miR-182	**2064**	hsa-miR-574-5p	40
hsa-miR-186	**172**	hsa-miR-16-2*	39
hsa-miR-191	**5385**	hsa-miR-148a*	38
hsa-miR-192	**95**	hsa-miR-3131	38
hsa-miR-193b*	**99**	hsa-miR-378	36
hsa-miR-194	**86**	hsa-miR-548j	36

hsa-miR-196a	**467**	hsa-miR-320b	35
hsa-miR-19b	**252**	hsa-miR-128	34
hsa-miR-21	**458**	hsa-miR-181a*	34
hsa-miR-2110	**150**	hsa-miR-25*	34
hsa-miR-221	**2226**	hsa-miR-30a*	33
hsa-miR-221*	**1636**	hsa-miR-3179	33
hsa-miR-222	**22997**		
hsa-miR-26a	**2108**		
hsa-miR-27b	**734**		

References

1. Probst, A. V., Dunleavy, E., & Almouzni, G. (2009). Epigenetic inheritance during the cell cycle. Nature Reviews.Molecular Cell Biology, 10(3), 192-206.

2. Mill, J., Tang, T., Kaminsky, Z., Khare, T., Yazdanpanah, S., Bouchard, L., et al. (2008). Epigenomic profiling reveals DNA-methylation changes associated with major psychosis. American Journal of Human Genetics, 82(3), 696-711.

3. Petronis, A. (2004). The origin of schizophrenia: Genetic thesis, epigenetic antithesis, and resolving synthesis. Biological Psychiatry, 55(10), 965-970.

4. Petronis, A., Gottesman, I. I., Kan, P., Kennedy, J. L., Basile, V. S., Paterson, A. D., et al. (2003). Monozygotic twins exhibit numerous epigenetic differences: Clues to twin discordance? Schizophrenia Bulletin, 29(1), 169-178.

5. Berger, S. L., Kouzarides, T., Shiekhattar, R., & Shilatifard, A. (2009). An operational definition of epigenetics.23, 781-783.

6. Waddington CH. Organisers and Genes. Cambridge, UK: Cambridge Univ. Press, 1940.

7. Santarelli, D.M., Beveridge, N.J., Tooney, P.A. & Cairns, M.J. (2011) Upregulation of dicer and microRNA expression in the dorsolateral prefrontal cortex brodmann area 46 in schizophrenia. Biological Psychiatry, 69(2), 180-187.

8. Kim, A.H. Reimers, M., Maher, B., Williamson, V., McMichael, O., McClay, J.L. et al. (2010). MicroRNA expression profiling in the prefrontal cortex of individuals affected with schizophreni and bipolar disorder. Schizophrenia Research, 124 (1-3), 183-191.

9. Ruby, J.G., Jan, C. H., & D. P. Bartel. (2007) Intronic microRNA precursors that bypass Drosha processing. Nature, 448 83-86.

10. Westholm, J.O. & E.Lai. (2011) Mirtrons: microRNA biogenesis via splicing. Biochimie 93 (11) , 1897-1904.

11. Alawi, F., & Lin, P. (2010). Loss of dyskerin reduces the accumulation of a subset of H/ACA snoRNA-derived miRNA. Cell Cycle (Georgetown, Tex.), 9(12), 2467-2469.

12. Brameier, M., Herwig, A., Reinhardt, R., Walter, L., & Gruber, J. (2011). Human box C/D snoRNAs with miRNA like functions: Expanding the range of regulatory RNAs. Nucleic Acids Research 39(2), 675-686.

13. Ender, C., Krek, A., Friedlander, M. R., Beitzinger, M., Weinmann, L., Chen, W., et al. (2008). A human snoRNA with MicroRNA-like functions. Molecular Cell 32 (4), 519-528.

14. Kiss, T. (2002). Small nucleolar RNAs: An abundant group of noncoding RNAs with diverse cellular functions. Cell, 109(2), 145-148.

15. Ono, M., Scott, M. S., Yamada, K., Avolio, F., Barton, G. J., & Lamond, A. I. (2011). Identification of human miRNA precursors that resemble box C/D snoRNAs. Nucleic Acids Research, 39(9), 3879-3891.

16. Politz, J. C. R., Hogan, E. M., & Pederson, T. (2009). MicroRNAs with a nucleolar location. RNA,15 (9), 1705-1715.

17. Saraiya, A. A., & Wang, C. C. (2008). snoRNA, a novel precursor of microRNA in giardia lamblia. PloS Pathology 4(11), e1000224.

18. Scott, M. S., Avolio, F., Ono, M., Lamond, A. I., & Barton, G. J. (2009). Human miRNA precursors with box H/ACA snoRNA features. PloS Computational Biology 5(9), e1000507.

19. Taft, R. J., Glazov, E. A., Lassmann, T., Hayashizaki, Y., Carninci, P., & Mattick, J. S. (2009). Small RNAs derived from snoRNAs. RNA. 15, 1233-1240.
20. Mortazavi, A, Williams, B.A., McCue, K., Schaeffer, L. & B. Wold. (2008) Mapping and quantifying mammalian transcriptomes by RNAseq. Nature Methods 5(7) 621-628

21. Wan, L., Yan, X., Chen,T., & F. Sun (2012) Modeling RNA degradation for RNAseq with applications. Biostatistic (Oxford, England).

22. Finn, R.D., Gardener, P.P., & A. Bateman. (2012) Making your database available through wikipedia: The pros and cons. Nucleic Acids Research 40 (1), D9-D12.

23. Rogelj, B. (2006) Brain-specific small nucleolar RNAs. Journal of Molecular Neuroscience. 28(2) 103-109.

24. Brazma, A., Hingamp, P., Quackenbush, J., Sherlock, G., Spellman, P., Stoeckert, C., et al. (2001). Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. Nature Genetics, 29(4), 365-371.

25. Rother, K., Potrzebowski, W., Puton, T., Rother, M., Wywial, E., & Bujnicki, J. M. (2012). A toolbox for developing bioinformatics software. Briefings in Bioinformatics, 13(2), 244-257.

26. Williamson, V., Kim, A., Xie, B., McMichael, G., Gao, Y., and V. Vladimirov. (2012)

Detecting miRNAs in deep-sequencing data: a software performance comparison and evaluation. Briefings in Bioinformatics, 13(1).

27. Perkins, D. O., Jeffries, C., & Sullivan, P. (2005). Expanding the 'central dogma': The regulatory role of nonprotein coding genes and implications for the genetic liability to schizophrenia.10, 69-78.

28. Perkins, D. O., Jeffries, C. D., Jarskog, L. F., Thomson, J. M., Woods, K., Newman, M. A., et al. (2007). microRNA expression in the prefrontal cortex of individuals with schizophrenia and schizoaffective disorder. Genome Biology, 8(2), R27.

29. Howie, B., Marchini, J., Stephens, M., & Chakravarti, A. (2011). Genotype imputation with thousands of genomes. G3, 1(6), 457-470.

30. Marchini, J., & Howie, B. (2010). Genotype imputation for genome-wide association studies. Nature Reviews.Genetics, 11(7), 499-511.

31. Howie, B. N., Donnelly, P., & Marchini, J. (2009). A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genetics, 5(6), e1000529.

32. Pei, Y. F., Zhang, L., Li, J., & Deng, H. W. (2010). Analyses and comparison of imputation-based association methods. PloS One, 5(5), e10827.

33. Wang, Z., Jacobs, K. B., Yeager, M., Hutchinson, A., Sampson, J., Chatterjee, N., et al. (2011). Improved imputation of common and uncommon SNPs with a new reference set. Nature Genetics, 44(1), 6-7.

34. Pei, Y. F., Li, J., Zhang, L., Papasian, C. J., & Deng, H. W. (2008). Analyses and comparison of accuracy of different genotype imputation methods. PloS One, 3(10), e3551.

35. Zheng, J., Li, Y., Abecasis, G. R., & Scheet, P. (2011). A comparison of approaches to account for uncertainty in analysis of imputed genotypes. Genetic Epidemiology, 35(2), 102-110.

36. Halperin, E., & Stephan, D. A. (2009). SNP imputation in association studies. Nature Biotechnology, 27(4), 349-351.

37. Li, Y., C. J. Willer, J. Ding, P. Scheet, and G. R. Abecasis, (2010) MaCH: using sequence and genotype data to estimate haplotypes and unobservedgenotypes. Genetic Epidemiology, 34: 816–834.

38. Browning, B and Z. Yu. (2009) Simultaneous genotype calling and haplotype phase inference improves genotype accuracy and reduces false positive associations for

genome-wide association studies. The American Journal of Human Genetics 85:847-861.

39. Sim, X., Ong, R., Suo, C., Tay, W-T., Liu, J., Ng, D., Boehnke, M., Chia, K-S., Wong, T-Y., Seielstad, M., Teo, Y-Y., and E-S. Tai. (2011) Transferability of Type 2 Diabetes Implicated Loci in Multi-Ethnic Cohorts from Southeast Asia. PLoS Genetics 7(4) e10011363,doi :10.1371/journal.pgen.1001363

40. Uh, H-W., Deelen, J., Beekman, M., Helmer, Q., Rivadeneira, F., Hottenga, J-J., Boomsma, D., Hofman, A, Uitterlinden, A., Slagboom, P., Bohringer, S., and J. Houwing-Duistermaat (2011). How to deal with the early GWAS data when imputing and combining different arrays is necessary. European Journal of Human Genetics 1-5.

41. Jiang, P., Wu, H., Wang, W., Ma, W., Sun, X., & Lu, Z. (2007). MiPred: Classification of real and pseudo microRNA precursors using random forest prediction model with combined features. Nucleic Acids Research, 35(Web Server issue), W339-44.

42. Wang, Z., Jacobs, K. B., Yeager, M., Hutchinson, A., Sampson, J., Chatterjee, N., et al. (2011). Improved imputation of common and uncommon SNPs with a new reference set. Nature Genetics, 44(1), 6-7.

43. Saunders, M. A., Liang, H., & Li, W. H. (2007). Human polymorphism at microRNAs and microRNA target sites. Proceedings of the National Academy of Sciences of the United States of America, 104(9), 3300-3305.

44. Thomas, L. F., Saito, T., & Saetrom, P. (2011). Inferring causative variants in microRNA target sites. Nucleic Acids Research, 39(16), e109.

45. Barenboim, M., Zoltick, B. J., Guo, Y. J., & Weinberger, D. R. (2010). MicroSNiPer: A web tool for prediction of SNP effects on putative microRNA targets.31, 1223-1232.

46. Berezikov, E., Thuemmler, F., van Laake, L. W., Kondova, I., Bontrop, R., Cuppen, E., et al. (2006). Diversity of microRNAs in human and chimpanzee brain. Nature Genetics, 38(12), 1375-1377.

47. Kertesz, M., Iovino, N., Unnerstall, U., Gaul, U., and E. Segal (2007) The role of site accessibility in microRNA target recognition. Nature Genetics 39 1278-1284.

48. Lai, C. Y., Yu, S. L., Hsieh, M. H., Chen, C. H., Chen, H. Y., Wen, C. C., et al. (2011). MicroRNA expression aberration as potential peripheral blood biomarkers for schizophrenia. PloS One, 6(6), e21635.

49. Mishra, P. J., & Bertino, J. R. (2009). MicroRNA polymorphisms: The future of pharmacogenomics, molecular epidemiology and individualized medicine. Pharmacogenomics, 10(3), 399-416.

50. Xiao, F., Zuo, Z., Cai, G., Kang, S., Gao, X., & Li, T. (2009). miRecords: An integrated resource for microRNA-target interactions. Nucleic Acids Research, 37(Database issue), D105-10.

51. John, B., Enright, A. J., Aravin, A., Tuschl, T., Sander, C., & Marks, D. S. (2004). Human MicroRNA targets. PLoS Biology, 2(11), e363.

52. Ripke, S., Sanders, A. R., Kendler, K. S., Levinson, D. F., Sklar, P., Holmans, P. A., et al. (2011). Genome-wide association study identifies five new schizophrenia loci. Nature Genetics, 43(10), 969-976.

53. Griffiths-Jones, S., Saini, H. K., van Dongen, S., & Enright, A. J. (2008). miRBase: Tools for microRNA genomics.36, D154-D158.

54. Ritchie, W., Flamant, S., & Rasko, J. E. (2009). Predicting microRNA targets and functions: Traps for the unwary. Nature Methods, 6(6), 397-398.

55. Clop, A., Marcq, F., Takeda, H., Pirottin, D., Tordoir, X., Bibe, B., et al. (2006). A mutation creating a potential illegitimate microRNA target site in the myostatin gene affects muscularity in sheep. Nature Genetics, 38(7), 813-818.

56. Yu, Z. B., Li, Z., Jolicoeur, N., Zhang, L. H., Fortin, Y., Wang, E., et al. (2007). Aberrant allele frequencies of the SNPs located in microRNA target sites are potentially associated with human cancers.35, 4535-4541.

57. Campayo, M., Navarro, A., Vinolas, N., Tejero, R., Munoz, C., Diaz, T., et al. (2011). A dual role for KRT81: A miR-SNP associated with recurrence in non-small-cell lung cancer and a novel marker of squamous cell lung carcinoma. PloS One, 6(7), e22509.

58. Rotunno, M., Zhao, Y., Bergen, A. W., Koshiol, J., Burdette, L., Rubagotti, M., et al. (2010). Inherited polymorphisms in the RNA-mediated interference machinery affect microRNA expression and lung cancer survival. British Journal of Cancer, 103(12), 1870-1874.

59. Sun, G., Yan, J., Noltner, K., Feng, J., Li, H., Sarkis, D. A., et al. (2009). SNPs in human miRNA genes affect biogenesis and function. RNA (New York, N.Y.), 15(9), 1640-1651.

60. Yang, H., Dinney, C. P., Ye, Y., Zhu, Y., Grossman, H. B., & Wu, X. (2008). Evaluation of genetic variants in microRNA-related genes and risk of bladder cancer. Cancer Research, 68(7), 2530-2537.

61. Wang, W. X., Wilfred, B. R., Baldwin, D. A., Isett, R. B., Ren, N., Stromberg, A., et al. (2008). Focus on RNA isolation: Obtaining RNA for microRNA (miRNA) expression

profiling analyses of neural tissue. Biochimica Et Biophysica Acta, 1779(11), 749-757.

62. Smirnova, L., Grafe, A., Seiler, A., Schumacher, S., Nitsch, R., & Wulczyn, F. G. (2005). Regulation of miRNA expression during neural cell specification. The European Journal of Neuroscience, 21(6), 1469-1477.

63. Lee, P. R., & Fields, R. D. (2009). Regulation of myelin genes implicated in psychiatric disorders by functional activity in axons. Frontiers in Neuroanatomy, 3, 4.

64. Hakak, Y., Walker, J. R., Li, C., Wong, W. H., Davis, K. L., Buxbaum, J. D., Haroutunian, V., and Fienberg, A. A. (2001). Genome-wide expression analysis reveals dysregulation of myelination-related genes in chronic schizophrenia. Proc. Natl. Acad. Sci. U.S.A. 98, 4746–4751.

65. Risso, D., Schwartz, K., Sherlock, G., & Dudoit, S. (2011). GC-content normalization for RNA-seq data. BMC Bioinformatics, 12(1), 480.

66. Hackenberg, M., Rodriguez-Ezpeleta, N., & Aransay, A. M. (2011). miRanalyzer: An update on the detection and analysis of microRNAs in high-throughput sequencing experiments. Nucleic Acids Research, 39(Web Server Issue), W132-8.

67. Langmead, B., Trapnell, C., Pop, M., & Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biology, 10(3), R25.

68. Robinson, M. D., & Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biology, 11(3), R25.

69. Bar M FAU - Wyman, Stacia,K., FAU, W. S., FAU, F. B., Qi J FAU - Garg, Kavita,S., FAU, G. K., FAU, P.R., et al. (1128).MicroRNA discovery and profiling in human embryonic stem cells by deep sequencing of small RNA libraries.

70. Meyer, S. U., Pfaffl, M. W., & Ulbrich, S. E. (2010). Normalization strategies for microRNA profiling experiments: A 'normal' way to a hidden layer of complexity? Biotechnology Letters, 32(12), 1777-1788.

71. Abelson, J.F., et al. (2005) Sequence variants in SLITRK1 are associated with Tourette's syndrome. Science 310, 317-320.

72. Zhang, Qipeng, Lu, Ming, and Cui, Qinghua. (2008) SNP analysis reveals an evolutionary acceleration of the human-specific microRNAs. Nature Precedings. http://hdl.handle.net/10101/npre.2008.2127.1.

73. Lestrade, L., & Weber, M. J. (2006). snoRNA-LBME-db, a comprehensive database of human H/ACA and C/D box snoRNAs. *Nucleic Acids Research, 34*(Database issue),

D158-62.

74. Muinos-Gimeno, M., Guidi, M., Kagerbauer, B., Martin-Santos, R., Navines, R., Alonso, P., et al. (2009). Allele variants in functional MicroRNA target sites of the neurotrophin-3 receptor gene (NTRK3) as susceptibility factors for anxiety disorders. *Human Mutation, 30*(7), 1062-1071.

75. Bergen, S. E., & Petryshen, T. L. (2012). Genome-wide association studies of schizophrenia: Does bigger lead to better results? *Current Opinion in Psychiatry, 25*(2), 76-82.

76. Owen, M. J., Craddock, N., & O'Donovan, M. C. (2010). Suggestion of roles for both common and rare risk variants in genome-wide studies of schizophrenia. *Archives of General Psychiatry, 67*(7), 667-673.

77. Spencer, C. C., Su, Z., Donnelly, P., & Marchini, J. (2009). Designing genome-wide association studies: Sample size, power, imputation, and the choice of genotyping chip. *PLoS Genetics, 5*(5), e1000477.

78. Gilmore, S. A., Durgun, M. B., & Sims, T. J. (1996). Schwann cell-neuron relationships in spinal cord gray matter. *Glia, 18*(4), 261-268.

79. Wilkins, A., Majed, H., Layfield, R., Compston, A., & Chandran, S. (2003). Oligodendrocytes promote neuronal survival and axonal length by distinct intracellular mechanisms: A novel role for oligodendrocyte-derived glial cell line-derived neurotrophic factor. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience, 23*(12), 4967-4974.

80. Allaman, I., Belanger, M., & Magistretti, P. J. (2011). Astrocyte-neuron metabolic relationships: For better and for worse. *Trends in Neurosciences, 34*(2), 76-87.

81. Hiard, S., Charlier, C., Coppieters, W., Georges, M., & Baurain, D. (2010). Patrocles: A database of polymorphic miRNA-mediated gene regulation in vertebrates. *Nucleic Acids Research, 38*(Database issue), D640-51.

82. Ziebarth, J. D., Bhattacharya, A., Chen, A., & Cui, Y. (2012). PolymiRTS database 2.0: Linking polymorphisms in microRNA target sites with human diseases and complex traits. *Nucleic Acids Research, 40*(Database issue), D216-21.

83. Friedlander, M. R., Chen, W., Adamidi, C., Maaskola, J., Einspanier, R., Knespel, S., et al. (2008). Discovering microRNAs from deep sequencing data using miRDeep.*26*, 407-415.

84. Friedlander, M. R., Mackowiak, S. D., Li, N., Chen, W., & Rajewsky, N. (2012). miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven

animal clades. *Nucleic Acids Research, 40*(1), 37-52.

85. Huang, P., Liu, Y., Lee, C., Lin, W., Gan, R. R., Lyu, P., et al. (2010). DSAP: Deep-sequencing small RNA analysis pipeline.*38*, W385-391.

86. Tabares-Seisdedos, R., & Rubenstein, J. L. (2009). Chromosome 8p as a potential hub for developmental neuropsychiatric disorders: Implications for schizophrenia, autism and cancer.*14*, 563-589.

87. Berninger, P., Gaidatzis, D., van Nimwegen, E., & Zavolan, M. (2008). Computational analysis of small RNA cloning data.*44*, 13-21.

88. Sewer, A., Paul, N., Landgraf, P., Aravin, A., Pfeffer, S., Brownstein, M. J., et al. (2005). Identification of clustered microRNAs using an ab initio prediction method.*6*

89. Szklarczyk, D., Franceschini, A., Kuhn, M., Simonovic, M., Roth, A., Minguez, P., et al. (2011). The STRING database in 2011: Functional interaction networks of proteins, globally integrated and scored.*39*, D561-D568.

90. Lewis, B. P., Burge, C. B., & Bartel, D. P. (2005). Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell, 120*(1), 15-20.

91. Gejman,P., Sanders, A. & Duan, J. (2010). The role of genetics in the etiology of Schizophrenia. The Psychiatric clinics of North America 33(1) 35-66.

92. Gejman, P., Sanders, A.R & K. S Kendler. (2011) Genetics of schizophrenia: new findings and challenges. Annual Review of Genomics and Human Genetics, 12:121.
93. Ambros V, Bartel B, Bartel DP, Burge CB, Carrington JC, Chen X, Dreyfuss G, Eddy SR, Griffiths-Jones S, Marshall M, Matzke M, Ruvkun G, Tuschl T. (2003) A uniform system for microRNA annotation. RNA, 2003, 9(3), 277-279.

94. Sklar, P., Ripke, S., Scott, L. J., Andreassen, O. A., Cichon, S., Craddock, N., et al. (2011). Large-scale genome-wide association analysis of bipolar disorder identifies a new susceptibility locus near ODZ4. *Nature Genetics, 43*(10), 977-983.
95. Corvin, A. P. (2011). Two patients walk into a clinic...a genomics perspective on the future of schizophrenia. *BMC Biology, 9*, 77.

96. Pienaar, E., Theron, M., Nelson, M., & Viljoen, H. (2006). A quantitative model of error accumulation during pcr amplification. *Computational Biology and Chemistry, 30*(2), 102-111.
97. Lieberman, J. A. (1999). Is schizophrenia a neurodegenerative disorder? A clinical and neurobiological perspective. *Biological Psychiatry, 46*(6), 729-739.

98.International Schizophrenia Consortium, Purcell, S. M., Wray, N. R., Stone, J. L.,

Visscher, P. M., O'Donovan, M. C., et al. (2009). Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature, 460*(7256), 748-752.

99. Dohm, J. C., Lottaz, C., Borodina, T., & Himmelbauer, H. (2008). Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. Nucleic Acids Research, *36(16)*: e105.

100. Owen, M. J., Craddock, N., & Jablensky, A. (2007). The genetic deconstruction of psychosis. *Schizophrenia Bulletin, 33*(4), 905-911.

101. Schulze, T. G. (2010). Genetic research into bipolar disorder: The need for a research framework that integrates sophisticated molecular biology and clinically informed phenotype characterization. *The Psychiatric Clinics of North America, 33*(1), 67-82.

102. Moreau,M.P.;Bruse,S.E.;David-Rus,R.;Buyske,S.;Brzustowicz,L.M. (2011) Altered microRNA expression profiles in postmortem brain samples from individuals with Schizophrenia and Bipolar Disorder. Biological Psychiatry 69 (2) 188.

103. Beveridge, N.J. and M.Cairns. (2012) MicroRNA dysregulation in schizophrenia . Neurobiology of Disease 46:263-271.

104. Khanna, A. and S. Stamm. (2010) Regulation of alternative splicing by short non-coding nuclear RNAs. RNA Biology 7(4) 480-485.

105. Bonnet, E., Wuyts, J., Rouze, P., and Y. Van de Peer (2004) Evidence that microRNAs precursors, unlike other non-codingRNAs have lower folding energies than random sequences Bioinformatics 20: 2911-2917.

106. Trikalinos,T., Salantis, G., Khoury, M., and J. Ioannidis. (2006). Impact of Violations and Deviations in Hardy-Weinberg Equilibrium on Postulated Gene-Disease Associations. American Journal of Epidemiology 163:300-309.

107. Costantini, M., Cammarano, R., and G. Bernaldi. (2009) The evolution of isochore patterns in vertebrate genomes. BMC Genomics 10:146.

108. Choi, K. H., Higgs, B. W., Wendland, J. R., Song, J., McMahon, F. J., & Webster, M. J. (2011). Gene expression and genetic variation data implicate PCLO in bipolar disorder. *Biological Psychiatry, 69*(4), 353-359.

109. Higgs, B. W., Elashoff, M., Richman, S., & Barci, B. (2006). An online database for brain disease research. *BMC Genomics, 7*, 70.

110. Torrey, E. F., Webster, M., Knable, M., Johnston, N., & Yolken, R. H. (2000). The stanley foundation brain collection and neuropathology consortium. *Schizophrenia*

Research, 44(2), 151-155.
111. Voellenkle, C., Rooij, J., Guffanti, A., Brini, E., Fasanaro, P., Isaia, E., et al. (2012). Deep-sequencing of endothelial cells exposed to hypoxia reveals the complexity of known and novel microRNAs. *RNA (New York, N.Y.), 18*(3), 472-484

112. Inukai, S., de Lencastre, A., Turner, M., and F. Slack (2012) Novel MicroRNAs Differentially Expressed during Aging in the Mouse Brain. Plos One 7(7); e40028. Doi:10.1371/journal.pone.0040028.

113. Sartorius, N., Jablensky, A and R. Shapiro.(1978) Cross-cultural differences in the short-term prognosis of schizophrenic psychoses. Schizophrenia Bulletin 4, 102-113.

114. Lee, I., Ajay, S. S., Yook, J. I., Kim, H. S., Hong, S. H., Kim, N. H., et al. (2009). New class of microRNA targets containing simultaneous 5'-UTR and 3'-UTR interaction sites. *Genome Research, 19*(7), 1175-1183.

115. Miller, B., Zeier, Z., Li, X., Lanz, T., et al. (2012) MicroRNA-132 dysregulation in schizophrenia has implications for both neurodevelopment and adult brain function. Proceedings of the National Academy of Sciences of the United States of America, 109(8), 3125-3130.

116. Benjamini, Y and Y Hochberg. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J Roy Statis Soc Ser B 57: 289-300.

117. Lister, R., Pelizzola, M., Dowen, R.H., Hawkins, R.D., Hon,G., et al. (2009) Human DNA methylome s at base resolution show widespread epigenomic differences Nature 462(7271) 315-322.

118. Gardiner-Garden, M., and M. Fromme. (1987) CpG islands in vertebrate genomes. J. Molecular Biology 196 (2) 261-282.

119. Maher. B.S., Vladimirov, V. I., Latendresse, S.J., Thiselton, D.L., McNamee, R., Kang, M., Bigdeli, T.B. et al. (2011) The AVPR1A gene and substance use disorders: association, replication, and functional evidence. Biological Psychiatry 70(6) 519-527.

120. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L. et al. (2007) PLINK: a toolset for whole genome association and population-based linkage analysis. American Journal of Human Genetics 81(3) 559-575.

121. Delaneau, O., Coulonges, C and J.F. Zagury. (2008) Shape-IT: a new rapid and accurate algorithm for haplotype inference. BMC Bioinformatics 9 540-554.

122. Sullivan, E., Kendler, K.S. and M. C. Neale (2003) Schizophrenia as a complex trait: evidence from a meta-analysis of twin studies. Arch Gen Psychiatry 60 1187-1192.

123. Maher, B., and J. LoTurco. (2012) Disrupted-in-Schizophrenia (DISC1) Functions Presynaptically at Glutamartergic Synapses. Plos One 7(3) , e 34053 doi:10.1371/journal.pone.0034053.

124. Bustin, S., Benes, V., Garson, J., Hellemans, J. et al. (2009) The MIQE Guidelines-Minimum Information for Publication of Qualitative Real-Time PCR Experiments. Clinical Chemistry 55(4)611-622.
125. Creighton, C.J. Reid, J., and P. Gunaratne. (2009) Expression profiling of microRNAs by deep sequencing. Briefings in Bioinformatics 10(5) 490-497.
126. Han, Y., Chen, J., Zhao, X., Liang, C. et al. (2011) MicroRNA Expression Signatures of Bladder Cancer Revealed by Deep Sequencing. PlosOne 6(3) e18286: doi10.1371/journal.pone 0018286.

127. Sullivan, P. (2005) The Genetics of Schizophrena. PlosOne Med 2(7): e212.doi:10.1371/journal.pmed.0020212.

128. Craddock, N., O'Donovan, M.C., and M.J.Owen. (2005) The genetics of schizophrenia and bipolar disorder: dissecting psychosis. Journal of Medical Genetics 42 193-204.

129. Noble, W. (2009) How does multiple testing correction work? Nat. Biotechnol. 27(12) 1135-1137.

130. Melios, N., and M. Sur. (2012) The Emerging Role of microRNAs in Schizophrenia and Autism Spectrum Disorder. Frontiers in Psychiatry 3 doi:10.3389/fpsyt.2012.00039.

131. Wanet, A., Tacheny, A., Arnould,T., and P. Renard. (2012) miR-212/132 expression and functions: within and beyond the neuronal compartment. Nucleic Acids Research 40(11) doi:10.1093/nar/gks151.

132. Marin, R., and J. Vanicek. (2012) Optimal Use of Conservation and Accessibility Filters in MicroRNA Target Prediction. PLosOne 7(2) e32208 doi:10.1371/journal.pone.032208.

133. Higdon, R., van Belle, G., and E. Koeller (2008) A note on the false discovery rate and inconsistent comparisons between experiments. Bioinformatics 24(10) 1225-1228.

134. Jablensky, A., and N. Sartorius. (2008) What did the WHO Studies Really Find? Schizophrenia Research 34 (2) 253-255.

135. Cannon, M., and P.Jones. (1996) Schizophrenia. Journal of Neurology, Neurosurgery, and Psychiatry 61 604-613.

136. Beveridge, N.J., Gardiner, E., Carroll, A.P., Tooney,P.A., and M.J. Cairns. (2010) Schizophrenia is associated with an increase in cortical microRNA biogenesis. Mol

Psychiatry 15(12)1176-1189.

137. Melios, N., Huang,H.S., Grigorenko, A., Rogaev, E., and S.Akbarian. (2008) A set of differentially expressed miRNAs, including miR-30a-5p, act as post-transcriptional inhibitors. Biological Psychiatry 65(12) 1006-1014.

138. Beveridge, N.J., Tooney, P.A., Carroll, A.P., Gardiner, E., Bowden, N., Scott, R.J., Tran, M., et al. (2008) Dysregulation of miRNA 181b in the temporal cortex in schizophrenia. Human Molecular Genetics 17(8) 1156-1168.

139. Riley, B. and K. Kendler. (2006) Molecular genetics of Schizophrenia. European Journal of Human Genetics 14 669-680.

140. Cardno, A.G. and I.I. Gottesman (2000) Twin studies of schizophrenia: from bow-and-arrow concordances to star wars Mx and functional genomics. AM J SLTMed Genetics 97(1) 12-17.

141 Wahlberg, K.E., Wynne, L., Oja, H., Keskitalo, P. et al. (1997) Gene-Environment Interaction in Vulnerability to Schizophrenia: Findings From the Finnish Adoptive Family Study of Schizophrenia. Am J Psychiatry 1997; 154:355–362.

142. Thapar, A., Harold, G., Rice, F., Langley, K. and M. O'Donovan (2007). The contribution of gene–environment interaction to psychopathology. Development and Psychopathology,19, pp 989 - 1004 doi:10.1017/S0954579407000491

143. Chakravarti, A. (1999) Population genetics—making sense out of sequence. Nature Genetics supplement 21 56-60.

144. Pritchard, J.K., and N Cox. (2002) The allelic architecture of human disease genes: common disease – common variant … or not? Human Molecular Genetics 11(20) 2417-2423.

145. Shih, P., Belmonte, P., and P. Zandi. (2004) A review of the evidence from family, twin, and adoption studies for a genetic contribution to adult psychiatric disorders. International Review of Psychiatry 16(4) 260-283.

146. Kendler, K.S., Pedersen, N.L., Neale, M.C., and A.A. Mathe. (1995) A pilot Swedish twin study of affective illness hospital- and population ascertained sub-samples: results of model fitting. Behavior Genetics, 25(3) 217-232.

147. Heston, L.L. (1966) Psychiatric disorders in foster home reared children of schizophrenic mothers. British Journal of Psychiatry 112(489) 819-825.

148. Bray, N. (2008) Gene expression in the Etiology of Schizophrenia. Schizophrenia Bulletin 34(3) 412-418.

149. Mimics, K., Middleton, F.A., Marquez, A., et al. (2000) Molecular characterization of schizophrenia viewed by microarray analysis of gene expression in prefrontal cortex. Neuron 28(1) 53-67.

150. Altar C.A., Jurata, L.W., Charles, V., Lemire, A., et al. (2005) Deficient hippocampal neuron expression of proteasome, ubiquitin, and mitochondrial genes in multiple schizophrenia cohorts. Biological Psychiatry 58(2) 85-96.

151. Maziere, P., and A.J. Enright (2007) Prediction of MicroRNA targets. Drug Discovery Today. 12(11) 452-458.

152. Gottesman, I.I. (1991) Schizophrenia Genesis: The Origns of Madness. W.H. Freeman & Co. New York.

153. Cohen E, Chow EW, Weksberg R, Bassett AS (1999) Phenotype of adults with the 22q11 deletion syndrome: A review. *Am J Med Genet* 86(4)**:**359-365.

154. Chen X, Wang X, Chen Q, Williamson V, van den Oord E, Maher BS, O'Neill FA, Walsh D, Kendler KS. (2008) MEGF10 association with Schizophrenia. Biol Psychiatry. 63(5):441-448.

155. Chen X, Dunham C, Kendler S *et al (2004)* Regulator of G-protein signaling 4 (RGS4) gene is associated with schizophrenia in Irish high density families. Am J Med Genet 2004; 129B: 23–26.

156. Glatt SJ, Faraone SV, Tsuang MT (2003) Association between a functional catechol *O*-methyltransferase gene polymorphism and schizophrenia: meta-analysis of case–control and family-based studies. Am J Psychiatry 2003; 160: 469–476.

157. Thiselton DL, Webb BT, Neale BM *et al (2004)* No evidence for linkage or association of neuregulin-1 (NRG1) with disease in the Irish study of high-density schizophrenia families (ISHDSF). Mol Psychiatry 2004; 9: 777–783

158. Shi J, Levinson DF, Duan J, et al. (2009) Common variants on chromosome 6p22.1 are associated with schizophrenia. Nature 2009; 460:753–757

159. Purcell SM, Wray NR, Stone JL, et al. (2009) Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature 2009; 460:748–752.

160. Steinberg S, de Jong S, Andreassen OA, et al. Common variants at VRK2 and TCF4 conferring risk of schizophrenia. Hum Mol Genet 2011; 20:4076–4081.

161. Smrt, R.D., Szulwach, K.E., Pfeiffer, R. L., Li, X., Guo, W., et al. (2010) MicroRNA miR-137 regulate neuronal maturation by targeting ubiquitin ligase mind bomb-1. Stem

Cells 28(6) 1060-1070.

162. Ginsberg, S. D., Hemby, S.E., and J.F. Smiley (2012) Expression profiling in neuropsychitric disorders: emphasis on glutamate receptors in bipolar disorder. Pharmacol Biochem Behav. 100(4) 705-11.

163. Mirnics, K., Middleton, F., Marquez, A., Lewis, D., and P. Levitt (2000) Molecular Characterization of Schizophrenia Viewed by Microarray Analysis. Neuron 28 53–67.

164. Lin, C.Y., Sawa, A., and H. Jaaro-Peled. (2012) Better understanding of mechanisms of schizophrenia and bipolar disorder: from human gene expression profiles to mouse models. Neurobiol Dis. 45(1) 48-56.

165. Bromet, E., Andrade, L.H., Hwang, I., et al. (2011) Cross-national epidemiology of DSM-IV major depressive episode. BMC Medicine 9 doi:10.1186/1741-7015-9-90.

166. Dweep, H., Sticht, C., Pandey, P., and N. Gzretz (2011) miRWalk – database: prediction of possible miRNA binding sites by “walking” the genes of 3 genome. Journal of Biomedical Informatics doi: 10.1016/j.jbi.2011.05.002

167. Xao, F., Zuo, Z., Cai, G., Kang, S., Gao, X., and T. Li (2009) miRecords : an integrated resource for microRNA-target interactions. Nucleic Acids Research 37: D105-D110.

168. Chen, X., Williamson, V.S., An, S.S., Hettema, J.M., et al (2008) Cannabinoid receptor 1 gene association with nicotine dependence. Archives of General Psychiatry 65(7) 816-824.

169. Kwon, J., and A. Goate. (2000) The Candidate Gene Approach. Alcohol Research and Health. 24(3) 164-168.

170. Rucker, J., Newman, S., Gray, J., Gunasinghe, C. et al. (2011) OPCRIT+: an electronic system for psychiatric diagnosis and data collection in clinical and research settings. British Journal of Psychiatry199 151-155.

171. World Health Organization. (1993) The ICD-10 Classification of Mental and Behavioural Disorders: Diagnostic Criteria for Research.

172. American Psychiatric Association. (2000) Diagnostic and Statistical Manual of Mental Disorders (4th edn, text revision) (DSM–IV–TR). APA.

173. McGuffin P, Farmer A, Harvey I. (1991) A polydiagnostic application of operational criteria in studies of psychotic illness. Development and reliability of the OPCRIT system. Arch Gen Psychiatry 48 764–70.

174. Carbonelli, J., Alloza, E., Arce, P., Borrego, S., et al. (2012) A map of human microRNA variation uncovers unexpectedly high levels of variability. Genome Medicine 4(62) doi:10.1186/gm363.

175. Dick, D., Riley, B., and K. Kendler (2010). Nature and nurture in neuropsychiatric genetics: where do we stand? Dialogues Clinical Neuroscience 12 7-23.

176. Jorgensen, T., Ruczinski, I., Kessing, B., Smith, M. et al (2009) Hypothesis-Driven Candidate Gene Association Studies: Practical Design and Analytical Considerations American Journal of Epidemiology 170(8) 986 – 993.

177. Lachman HM, Morrow B, Shprintzen R, *et al.* (1996). Association of codon 108/158 catechol-o-methyltransferase gene polymorphism with the psychiatric manifestations of velo-cardio-facial syndrome.. *Am J Med Genet* **67** (5): 468-72.

178. Vrijenhoek T, Buizer-Voskamp JE, van der Stelt I, et al. (2008) Recurrent CNVs disrupt three candidate genes in schizophrenia patients. *Am. J. Hum. Genet.* 83 504–10.

179. Rodriguez-Murillo,L., Gogos, J., and M. Karayiorgou (2012) The Genetic Architecture of Schizophrenia: New Mutations and Emerging Paradigms. Annual Reviews Medicine. 2012. 63:63–80.

180. Harrison, P., and A. Law. (2006) Neuregulin 1 and Schizophrenia: Genetics, Gene Expression and Neurobiology. Biological Psychiatry 60 132-140.

181. Riley, B., and K. Kendley (2011) Classical genetic studies of schizophrenia. In D Weinberger and Paul Harrison. (Eds.), Schizophrenia 245-268. United Kingdom: Wiley-Blackwell.

182. Murphy, K.C., Jones, L. A. and M. J. Owen (1999) High rates of schizophrenia in adults with velo-cardial-facial syndrome. Archives of General Psychiatry 56 940-945.

183. Karayiorgou, M., Simon, T., and J. Gogos. (2010) 22q11.2 microdeletions: linking DNA structural variation to brain dysfunction and schizophrenia. Nature Reviews Neuroscience 11 402-416.

184. Dahary, D., Shalgi, R., and Y. Pilpel. (2011) CpG Islands as a putative source for animal miRNAs: evolutionary and functional implications. Molecular and Biological Evolution 28(5) 1545-1551.

185. Sachs, N.A., Sawa, A., Holmes, S.E., Ross, C.A., et al (2005) A frameshift mutation in Disrupted in Schizophrenia 1 in an American family with schizophrenia and schizoaffective disorder. Molecular Psychiatry 10 758-764.

186. Talkowski, M.E., Seltman, H., Basset, A.S. et al. (2006) Evaluation of a susceptibility gene for schizophrenia: genotype based meta-analysis of RGS4 polymorphisms from thirteen independent samples. Biological Psychiatry 60 152-162.

187. Kirov, G., Zaharieva, I., Georgieva, L., et al . (2009) A genome-wide association study in 574 schizophrenia trios using DNA pooling. Molecular Psychiatry 14 796-803.

188. O'Donovan, M.C., Craddock,N., Norton, N., et al. (2008) Identification of novel schizophrenia loci by genome-wide association and follow-up. Nature Genetics 40 1053-1055.

189. Lencz, T., Morgan, T.V., Athanasiou, M., et al. (2007) Converging evidence for a pseudoautosomal cytokine receptor gene locus in schizophrenia. Molecular Psychiatry 12 572-580.

190. Grueter, C., van Rooij, E., Johnson, B. et al. (2012) A Cardiac MicroRNA Governs Systemic Energy Homeostasis by Regulation of MED13. Cell 149(3) 671-683.

191. Ebert, M., and P. Sharp (2012) Roles for MicroRNAs in Conferring Robustness to Biological Processes. Cell 149(3) 515-524.

192. Hansen, T., Olsen, L., Lindow, M., Jakobsen, K., et al. (2007) Brain expressed microRNAs implicated in Schizophrenia etiology. PLos One 2(9) e873 doi:10.1371/journal.pone.00000873.

192. Stringer, S., Wray, N., Kahn, R., E. Derks. (2011) Underestimated Effect Sizes in GWAS: Fundamental Limitations of Single SNP Analysis for Dichotomous Phenotypes. PLos One 6(11) e27964: doi10.1371/journal.pone.0027964.

193. Hindorff, L.A., Sethupathy, P., Junkins, H.A., Ramos, E.A., et al. (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci U.S.A. 106(23) 9362-9367.

194. Albus, M. (2012) Clinical Courses of Schizophrenia. Pharmacopsychiatry 45 (Suppl 1) S31-S35.

195. Greenwood, T.A., Light, G., Swerdlow, N.R., Radant, A., and D. Braff (2012) Association Analysis of 94 Candidate Genes and Schizophrenia-related Endophenotypes. PLosOne 7(1) e29630. Doi:10.1371/journal.pone.0029630.

196, Braff, D., Schork, N., and I. Gottesman. (2007) EndophenotypingSchizophrenia. Anerican Journal of Psychiatry 164 705-707.

197. Keller, W., Fischer, B., and W. Carpenter. Revisting the Diagnosis of Schizophrenia: Where have we been and where are We going? (2011) CNS Neuroscience and Therapeutics 17 83-88.

198. Gupta, M., Bhatnagar, P., Grover, S., et al. (2009) Association studies of catechol-0-methyltransferase (COMT) gene with schizophrenia and response to antipsychotic treatment. Pharmacogenomics 10(3) 385-397.

199. Palmatier, M.A., Kang, A.M., and K.K. Kidd. (1999) Global variation in the frequencies of functionally different catechol-O-methyltransferase alleles. Biological Psychiatry 46 557-567.

200. Illi, A., Kampman, O., Hanninen, K. et al. (2007) Catechol-0-methyltransferase Val108/158 Met genotype and response to antipsychotic medication in schizophrenia. Human Psychopharmacology 22 211-215.

201. Stefansson, H., Sarginson, J., Kong, A., et al (2003) Association of Neuregulin 1 with Schizophrenia Confirmed in a Scottish Population. American Journal of Human Genetics. 72(1) 83-87.

202. Yang, J.Z., Si, T.M., Ruan, Y., Ling, Y. et al. (2003) Association study of neuregulin gene with schizophrenia. Molecular Psychiatry 8 706-709.

203. Ortega, M.C., Bribian, A., Peregrin, S. et al. (2012) Neuregulin-1/ErbB4 signaling controls the migration of oligodendrocyte precursors cells during development. Experimental Neurology 235 (2) 610-620.

204. Liu, Y-L., Fann, C., Liu, C-M., Wu, J-Y. et al. (2006) Evaluation of RGS4 as a Candidate Gene for Schizophrenia. American Journal of Medical Genetics Part B (Neuropsychiatric Genetics)141B 418-420.

205. Lipska, B., Peters, T., Hyde, T., Halim, N. et al . (2006) Expression of DISC1 binding partners is reduced in schizophrenia and associated with DISC1 SNPs. Human Molecular Genetics 15(8) 1245-1258.

206. St Clair, D., Blackwood, D., Muir, W., et al (1990) Association within family of a balanced autosomal translocation with major mental illness. Lancet 336 13-16.

207. Ono, K., Kuwabara, Y., and J. Han (2011) MicroRNAs and cardiovascular diseases 278(10) 1619-1633.

208. Li, Y., Lin, L., and Jin, P. (2008) The microRNA pathway and fragile X mental retardation protein. Biochim Biophys Acta 1779(11) 702-705.

209. Chen, Q., Chen, X., Zhang, M., et al. (2011) miR-137 is frequently down-regulated in gastric cancer and is a negative regulator of Cdc42. Digestive Diseases and Sciences 56(7) 2009-2016.

210. Vaz, C., Ahmad, H., Sharma, P., Gupta, R. et al. (2010) Analysis of microRNA transcriptome by deep sequencing of small RNA libraries of peripheral blood. BMC Genomics 11: 288 doi: 10.1186/1471-2164-11-288.

211. Williams, R. (2012, September 12). Diagnostic Criteria for Schizophrenia.. Retrieved from http://biopsychinstitute.com/psychiatric-disorders/schizophrenia

212. Im., H-I., Hollander, J., Bali, P., and P. Kenny. (2010) MeCP2 control BDNF expression and cocaine intake through homeostatic interactions with microRNA-212 Nature Neuroscience 13 1120-1127.

213. Callicot, J.H., Straub, R.E., Pezawas, L.,, Egan, M. et al (2005) Variation in DISC1 affects hippocampal structure and function and increases risk for schizophrenia Proceedings of the National Academy of Sciences. 102 8627-8632.

214. Walsh, T., McCellan, J., McCarthy, S., Addington, A., et al. (2008) Rare Structural Variants Disrupt Multiple Genes in Neurodevelopmental Pathways inSchizophrenia. Science 320(5875) 539-543.

215. Logan, C.V., Lucke, B., Pottinger, C., Abdelmaed, Z.A. el at. (2011) Mutations in MEGF10, a regulator of satellite cell myogenesi, cause early onset myopathy, areflexia, respiratory distress and dysphagia (EMARDD). Nature Genetics 43(12) 1189-1192.

216. Pawel, S., and J. Lupski (2010) Structural Variation in the Human Genome and Its Role in Disease. Annual Review of Medicine 61 437-455.

217. Filipowicz, W., Bhattacharyya, S., and N. Sonenberg. (2008) Mechanisms of post-transcriptional regulation by microRNAs: are the answers in sight? Nature reviews genetics 9 102-114.

218. Smith, T.F., and M.S. Waterman (1981) Identification of common molecular subsequences. Journal of Molecular Biology 147 195-197

219. Rice, P., Longden, I., and A. Bleasby. (2000) EMBOSS: The European Molecular Biology Open Software Suite. Trends in Genetics 16(6) 276-277

220. Li, T., Li, Z., Chen, P., Zhao, et al. (2010) Common Variants in Major Histocompatibility Complex region and TCF4 Gene Are Significantly Associated with Schizophrenia in Han Chinese. Biological Psychiatry 68(7) 671-673.

221. Stefansson, H., Ophoff, R.A., Steinberg, S., Andreassen, O.A. et al. (2009) Common variants conferring risk in Schizophrenia. Nature 460 744-747.

222. Daughtery, L.C., Seal, R.L., Wright, M.W et al (2012) Gene family matters: expanding the HGNC resource. Human Genomics 6: 4-10.

223. Kapushesky, M., Emam, I., Holloway, E., Kurnosov, P. et al (2010) Gene Expression Atlas at the European Bioinformatics Institute. Nucleic Acids Research 38 (suppl 1) D690-D698.

224. Wei, X., GuoJun, C., and S. NingSheng. (2009) Progress in miRNA target prediction and identification. Sci China Ser C-Life Sci 52(12) 1123-1130.

225. Jacewicz, R., Galecki, P., Florkowski, A., and J. Berent (2008) Association of the tyrosine hydroxylase gene polymorphism with schizophrenia in the population of central Poland. Psychiatr. Pol. 42(4) 583-593.

226. Chen, C., Ridzon, D., Broomer, A., Zhou, Z., et al. (2005) Real-time quantification of microRNAs by stem-loop PCR. Nucleic Acids Research 30(20) doi:10.1093/nar/gni178.

227. Vester, B., and J. Wengel. (2004) LNA (Locked Nucleic Acid): High-Affinity Targeting of Complementary RNA and DNA. Biochemistry. 43(42) 13233-13241.

228. Vester, B., and J. Wengel. (2003) LNA: a versatile tool for therapeutics and genomics. Trends in Biotechnology. 21(2) 74-81.

229. Gerstein, M., Bruce, C., Rozowsky, J., Zheng, D., et al. (2007) What is a gene, post-ENCODE? History and updated definition. Genome Research 17 669-681.

230. Pearson, H. (2006) What is a gene? Nature 441: 398-401.

231. Harrow, J., Nagy, A., Reymond, A., Alioto, T., et al (2009) Identifying protein coding genes in genomic sequences. Genome Biology 10 (210) doi:10.1186/gb-2009/10/1/201.

232. Aparicio, S. (2000) How to count … human genes. Nature Genetics 25 129-130.

233. Ebert, M. and P. Sharp. (2010) Emerging Roles for Natural MicroRNA Sponges Current Biology 20 R858-R861.

234. Poliseno, L., Salmena, L., Zhang, J., Carver, B. et al. (2010) A coding independent function of gene and pseudo gene mRNAs regulate tumor biology. Nature 465 1033-1038.

235. Hermeking, H. (2012) MicroRNAs in the p53 network: micromanagement of tumor suppression. Nature Reviews 12 613-626.

Vita

Vernell Seay Williamson was born on June 17, 1969. She received a Bachelors of Science in Anthropology, with a minor in theater from Longwood University in 1991. She also has a Master's of Arts in Anthropology from Wake Forest University and a Master's of Science in Biology from Virginia State University. From 2006-2007, she taught Biology at John Tyler Community College and Germanna College.