# Development of a parallel tagged sequencing assay to study disease-related sequence variation in post-transcriptional gene regulation.

## MASTERARBEIT

Zur Erlangung des akademischen Grades

"Master of Science in Engineering"

Studiengang:

**"Umwelt- Verfahrens- und Biotechnik"**

Management Center Innsbruck

**angefertigt im Labor von Prof. Dr. Thomas Tuschl**

Betreuer Rockefeller University:

**Dr. Neil Renwick / Dr. Kemal Akat**

Begutachtende:

**Dr. Katrin Bach**

Verfasser:

**Dipl.-Ing. (FH) Carlo Bäjen**

Matrikelnummer: 0810352035

# Declaration

"‘Ich erkläre hiermit an Eides statt, dass ich die vorliegende Masterarbeit selbständig angefertigt habe. Die aus fremden Quellen direkt oder indirekt übernommenen Gedanken sind als solche kenntlich gemacht.
Die Arbeit wurde bisher weder in gleicher noch in ähnlicher Form einer anderen Prüfungsbehörde vorgelegt und auch noch nicht veröffentlicht.”

ODER

DECLARATION IN LIEU OF OATH.

I hereby declare, under oath, that this master thesis has been my independent work and has not been aided with any prohibited means. I declare, to the best of my knowledge and belief, that all passages taken from published and unpublished sources or documents have been reproduced whether as original, slightly changed or in thought, have been mentioned as such at the corresponding places of the thesis, by citation, where the extent of the original quotes is indicated. The paper has not been submitted for evaluation to another examination authority or has been published in this form or another.

New York, August 1, 2010 ______________________

# Acknowledgments

I am deeply grateful to Dr. Katrin Bach for supervising this thesis, and her support throughout my studies. I would also like to thank Prof. Gerhard Hillmer and his exceptional team at the Management Center Innsbruck for enabling these studies.

I owe my most sincere gratitude to Prof. Thomas Tuschl who gave me the opportunity to work in his laboratory at Rockefeller University. His boundless energy and enthusiasm motivates his trainees, including me. I warmly thank Ashley Searles and Cherin Sohn for their helpful administrative support.

I would like to thank my great advisors and friends, Dr. Neil Renwick and Dr. Kemal Akat. Neil co-supervised my benchwork and thesis preparation. Kemal provided computer support. One could not wish for better nor friendlier mentors.

Special thanks to Stefanie Großwendt and Volker Hovestadt (both ETH Zürich). Steffi initiated PCR optimization in this complex project and Volker kindly wrote perlscripts and provided bioinformatic support.

I wish to thank Prof. Markus Stoffel (ETH), Dr. Pablo Landgraf (HHU Düsseldorf), and David Keller-Gymnich (Rutgers University), for providing clinical information and DNA samples. I also wish to thank Dr. Agnes Viale (MSKCC) for 454 sequencing and Xuning Wang (Rockefeller University) for administering the RNAworld server.

I wish to extend my warmest thanks to all those in the Tuschl lab who gave me support while doing my research, including Dr. Manuel Ascano Jr., Dr. Pavol Cekan, Dr. Thalia Farazi, Dr. Markus Hafner, Dr. Stefan Juranek, Dr. Klaas Max, and Dr. Jessica Spitzer.

I also wish to thank Sean McGeary, Aleksandra Mihailovic, Jason Miller, Jeff Nusbaum, and Sara Rouhanifard for their technical support and friendship. Leonida Fleming tirelessly cleaned and prepared necessary laboratory equipment.

Many thanks to Carlos Oliveira, who gave me a home here in New York. Thanks buddy.

My deepest gratitude to my family and friends for their unflagging love and support. I am indebted to my parents and grandparents for their care and love. Thanks also to all other colleagues and friends, including my Grimma and MCI buddies.

I owe my loving thanks to my wonderful girlfriend Anna. I appreciate her support during my overseas research experience. Without her encouragement and understanding, it would have been impossible for me to undertake this work.

I very much appreciate the generous financial support from the Austrian Marshall Plan Foundation (Vienna) and Rockefeller University (New York) that enabled me to obtain an exceptional research experience prior to my PhD studies.

# Table of Contents

# Abstract

Post-transcriptional regulation of gene expression is a complex process, mediated by RNA-binding proteins (RBPs) and microRNAs (miRNAs), that ensures optimal distribution and usage of messenger RNAs (mRNAs); these molecules bind as trans-acting factors directly to cis-elements within transcripts. Sequence variations or mutations in RNA-binding domains (RBDs) of RBPs, in miRNAs, or RBP or miRNA target sites potentially alter transcript regulation, causing phenotypic changes and disease.

In our study, we are interested in studying the molecular basis for the significant clinical overlaps between Fragile X syndrome (FXS), neurofibromatosis type 1 (NF1), and autism. FXS and NF1 are single-gene disorders, respectively caused by mutations in the genes that encode for the Fragile X Mental Retardation Protein (FMRP) and the NF1 protein. We hypothesize that a mutated FMRP or miRNA binding site in a shared regulatory pathway could contribute to the phenotypic overlap of these disorders.

To investigate this hypothesis, we designed a nested multiplexed PCR assay followed by barcoded 454 pyrosequencing to assess multiple genomic regions in hundreds of individual patients. We first applied this method to study sequence variations in 15 FMRP and 9 miRNA target sites in the NF1 gene of 400 persons with and without autism. We modified the assay for increased sample throughput, and for a separate study, analyzed 17 highly expressed, multicopy miRNA genes in 768 persons with chronic lymphocytic leukemias, Type 2 Diabetes mellitus, and autism.

We developed a bioinformatic pipeline for data management and biocomputational analyses. We used Oligomap for barcode sequence identification and read trimming. We identified all target regions in similar proportions and found similar abundances of barcoded primer combinations for each patient group. For sequence alignment and single base pair polymorphism discovery, we used the MOSAIK/GigaBayes and BWA/Samtools pipeline, and the GS Amplicon Analyzer.

We identified 40 sequence variations in FMRP target regions (i2161, i2163, i2174, i2181, i2193, i2209, i2231, i2233, i2241/42, and i2249) and 4 variations in miRNA target sites (miRNA103/107, miRNA- 10a/b, and miR 30a-5p/b/c/d/e-5p). We detected 114 variations in regions flanking FMRP target regions and 21 variations in regions flanking miRNA target sites.

Sequence variation is likely an underestimated source of regulatory and pathogenetic changes in the human genome. Our assay is broadly applicable to all genomic regions and shows much promise for detection of mutations that perturb post-transcriptional regulation of gene expression.

# Kurzbeschreibung

Post-transkriptionelle Genregulation ist ein komplexer Prozess, in dem Spleißen, Translation, Stabilität und Lokalisierung von messenger RNA (mRNA) durch RNA-bindende Proteine (RBPs) und microRNAs (miRNAs) reguliert werden. RBPs und miRNAs binden als trans-agierende Faktoren an regulatorische cis-Elemente des Transkripts. Mutationen in miRNAs und ihren Bindesequenzen sowie in RNA-bindenden Domänen von RBPs beinflussen, instabilisieren und zerstören die transkriptionell Stabilität sowie Translation, was schlussendlich zur Veränderung des Phänotyps und Krankheiten führen kann.

Die neurologischen Erkrankungen Fragiles-X-Syndrom (FXS) und Neurofibromatose Typ 1 (NF1) haben ähnliche Krankheitsbilder und sind mit Autismus assoziiert. FXS und NF1 sind monogene Krankheiten, die typischerweise durch Mutationen im Gen für das Fragile X Mental Retardation Protein (FMRP) bzw. für das NF1 Protein verursacht werden.

Wir vertreten die Hypothese, dass mutierte RBP- und miRNA-Bindesequenzen in gemeinsamen molekularen Regulationswegen zu Überlappungen in den Phänotypen von FXS, Autismus und NF1 beitragen können. Wir entwickelten eine Methode zur raschen und parallelen Untersuchung von genetische Variation in RBP- und miRNA-Bindesequenzen; basierend auf multiplexed und nested PCRs, gefolgt von paralleler, Barcode-kodierter 454-Pyrosequenzierung. Wir erprobten die Methode bei der Untersuchung von genetischen Variationen in 15 FMRP- und 9 miRNA-Bindesequenzen im NF1 Gen von 400 Personen mit und ohne Autismus. Wir modifizierten den Assay für einen höheren Probendurchsatz und analysierten 17 hochexprimierte miRNA Gene in 768 Patienten mit chronischer lymphatischer Leukämie, Diabetes mellitus vom Typ 2 und Autismus.

Wir etablierten eine Bioinformatik-Pipeline für bioinformatische Analysen und Datenmanagement; bestehend aus Oligomap für Barcode-Identifikation und der MOSAIK/GigaBayes- bzw. BWA/Samtools-Pipeline sowie dem GS Amplicon Analyzer für Sequenzealignments und Variationsanalysen. Alle Genbereiche und Barcodekombinationen konnten dabei identifiziert werden.

Vierzig Variationen in den FMRP-Bindesequenzen (i2161, i2163, i2174, i2181, i2193, i2209, i2231, i2233, i2241/42 und i2249) und vier Variationen in den miRNA-Bindesequenzen (miRNA 103/107, miRNA10a/b und miR 30a-5p/b/c/d/e-5p) wurden detektiert. Weitere 114 bzw. 21 Variationen um eine FMRP- bzw. miRNA-Bindesequenzen konnten identifiziert werden.

Unsere Methode ist uneingeschränkt auf allen Abschnitte des Genoms anwendbar und vielversprechend für die Detektion von Mutationen, die Störungen in der post-transkriptionellen Genregulation hervorrufen können.

# 1 Introduction

## 1.1 Post-transcriptional regulation of gene expression

Post-transcriptional regulation of gene expression is a complex mechanism that enables cells to adapt to environmental conditions, respond to stress, maintain homeostasis, and differentiate (Yang et al., 2003; Gama-Carvalho et al., 2006; Sonenberg and Hinnebusch, 2009). This type of gene regulation is mainly mediated by RNA-binding proteins (RBPs) and microRNAs (miRNAs) and enables control of all steps that connect transcription to protein translation (Moore, 2005; Keene, 2007).

Throughout their lifecycle, messenger ribonucleic acids (mRNAs) are part of regulatory ribonucleoprotein complexes. The principal steps in the transcript lifecycle are processing (transcription), nuclear export, cytoplasmic existence (translation), and degradation (Moore, 2005; McKee and Silver, 2007; Piao et al., 2010). The removal of transcripts from the translationally active pool is done by the cellular decay machinery, mainly the RNA-induced silencing complex (RISC) (van den Berg et al., 2008; Piao et al., 2010).

Due to the large numbers of RBPs and miRNAs, the potential combinations of ribonucleoprotein complexes (RNPs) and their influence on post-transcriptional gene regulation is staggering (Keene and Tenenbaum, 2002; McKee and Silver, 2007; Sonenberg and Hinnebusch, 2009). The roles of these regulatory complexes remain largely unexplored in human biology and disease (Martin and Ephrussi, 2009).

### 1.1.1 RNA-binding proteins and RNA-binding domains

Greater than 600 RBPs, possessing several well-defined RNA-binding domains (RBDs) are predicted to be encoded in the human genome (Venter, 2003). Currently 373 RBPs are listed in the Human Protein Reference Database[1] (Rozen and Skaletsky, 2000; Prasad et al., 2009). Structural diversity of RBPs and target recognition of RNA-binding domains (RBDs) are a function of the type, number, and arrangement of RBDs, helping RBPs to attain specificity (high affinity) for an RNA sequence (Lunde et al., 2007).

Several RBDs are currently recognized including the RNA-recognition motif (RRM), Zinc finger, K-homology (KH) domain, and double-stranded RBD (dsRBD). RRMs are the most abundant RBDs in higher vertebrates and are found in at least 0.5-1 % of human genes (Venter et al., 2001); they are typically composed of 80 to 90 amino acids (Query et al., 1989; Maris et al., 2005) and,

[1] http://www.hprd.org/; available at May 23, 2010

to date, twenty different structures of RRM:RNA complexes have been descibed (Lunde et al., 2007). The KH domain consists of approx. 70 amino acids, binds single-stranded (ss)DNA and ssRNA, and has a functionally important signature sequence near its centre; a mutation in the KH domain of Fragile X Mental Retardation Protein (FMRP) causes Fragile X syndrome (Lewis et al., 2000; Backe et al., 2005). Another domain contains repeats of an Arg-Gly-Gly motif, termed the RGG box, and also binds mRNA. (Schaeffer et al., 2001).

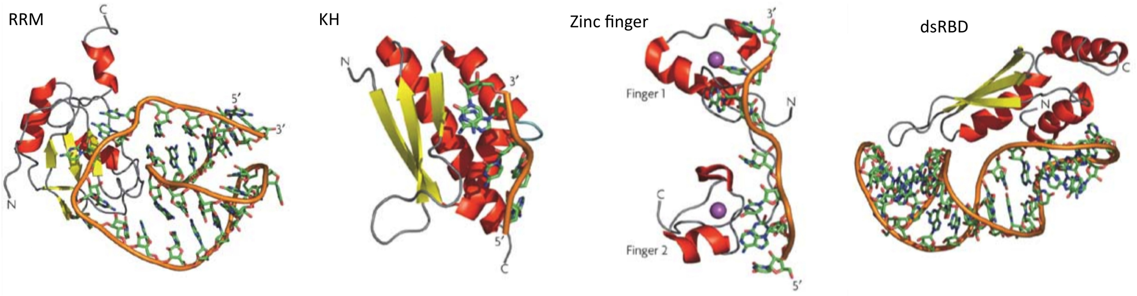


Figure 1.1: Example of known RBD structures (Lunde et al., 2007). Depicted are an RRM (the N-terminal RNA-recognition motif of human U1A), K-homology-3 (KH3) domain of Nova-2, two zinc fingers of TIS11d, and yeast Rnt1 dsRBD.

The binding of RBDs to RNA targets is one of the molecular bases of post-transcriptional RNA regulation and can be modulated precisely through combinations of RBDs, other enzymatic domains, and different intra-molecular interactions including electrostatic interactions, shape complementarity, and hydrogen bonding. Different combinations of RBDs and linker sequences between RBDs enable the recognition of RNA target sites over long distances, even allowing RBDs to bind to separated target sites. Linker length has functional relevance; longer linkers makes the recognition of several target sites possible, whereas a shorter linker permits binding only within a contiguous stretch of nucleotides (Lunde et al., 2007).

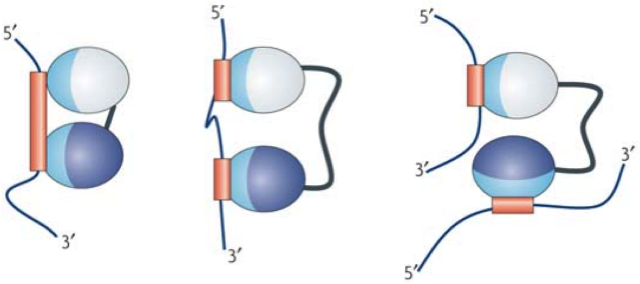


Figure 1.2: Scheme of modular RNA-binding domains (Lunde et al., 2007); (left) combined to recognize a long RNA sequence, (centre) separated target sites, and RNAs that belong to different molecules (right).

Current methods to identity RNA targets of RBPs include cross-linking and immunoprecipitation (CLIP) and RNP immunoprecipitation-microarray (RIP-Chip) protocols (Ule et al., 2005; Keene et al., 2006). Recently a new method for direct identification of RNA-binding motifs and RNA targets of mRBPs, referred to as photoactivatable ribonucleoside-enhanced crosslinking and immunoprecipitation (PAR-CLIP), was developed in the Tuschl laboratory (Hafner et al., 2010).

### 1.1.2 miRNAs and their regulatory function

microRNAs (miRNAs) are a class of endogenous small non-coding RNAs (19-24 nucleotides in length), which are initially transcribed as long primary transcripts (pri-miRNAs). After processing into stem-loop precursors (pre-miRNAs) of approx. 70 nucleotides (nt) by the nuclear enzyme Drosha, they are processed into mature miRNAs by the cytoplasma enzyme Dicer (Lee et al., 2003; Han et al., 2004). Following loading into the Argonaute protein, mature miRNAs guide the silencing complex to complementary mRNA target sites and regulate transcript cleavage or translational repression (see Figure 1.3) (Meister and Tuschl, 2004; Yekta et al., 2004; Bartel, 2009).

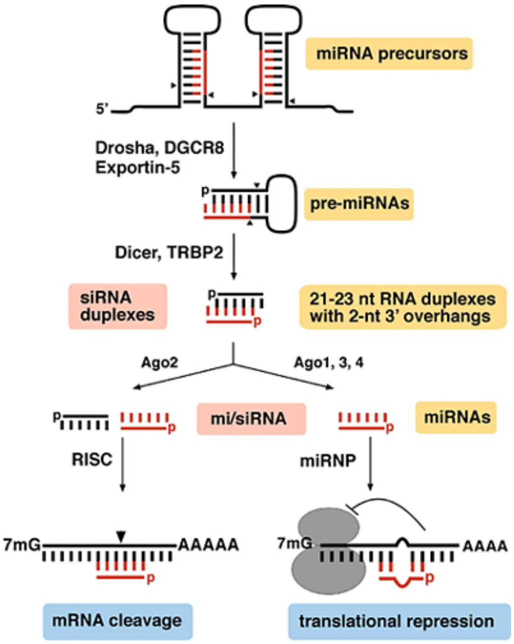


Figure 1.3: Simplified scheme of post-transcriptional gene regulation. Single-stranded pre-miRNA (hairpin structure) is processed by ribonuclease III. The resulting short miRNA (or siRNA) duplexes are separated, where the mature strand is assembled into an effector Argonaute (Ago) protein complex, while the star sequence is degraded. Depending on type of Ago protein, two different regulatory pathways are resulting in: (i) mRNA cleavage and (ii) translational repression by binding of miRNAs on mRNA. Figure is adapted from Meister and Tuschl, 2004.

Currently 677 human miRNAs are recognized (http://www.microRNA.org; last update 2008-09-08), targeting approx. 60 % of all human genes (John et al., 2004; Friedman et al., 2009). Single miRNAs can target multiple mRNAs, and transcripts may contain multiple target sites for the same and/or different miRNAs (Landgraf et al., 2007; Ruby et al., 2007).

To maintain their regulatory function, a remarkable fraction of miRNA target sites, corresponding to the miRNA seed sequence (nucleotides 2-7), is conserved. Most (> 45,000) target sites are located in 3' untranslated region (3'UTR) (Bartel, 2004; Friedman et al., 2009); open reading frames (ORFs) and 5'UTRs are rarely targeted, possibly due to removal of the silencing complexes by the translation apparatus (Bartel, 2004). miRNA target sites are not conserved in cell types or tissues where the cognate miRNA is absent (Farh et al., 2005). Appoaches like barcoded small RNA sequencing, miRNA in situ hybridization and Ago-CLIP allow the identification of miRNAs and their target sites in different tissues (see Figure 1.4), helping to define specificities of their expression and regulation.

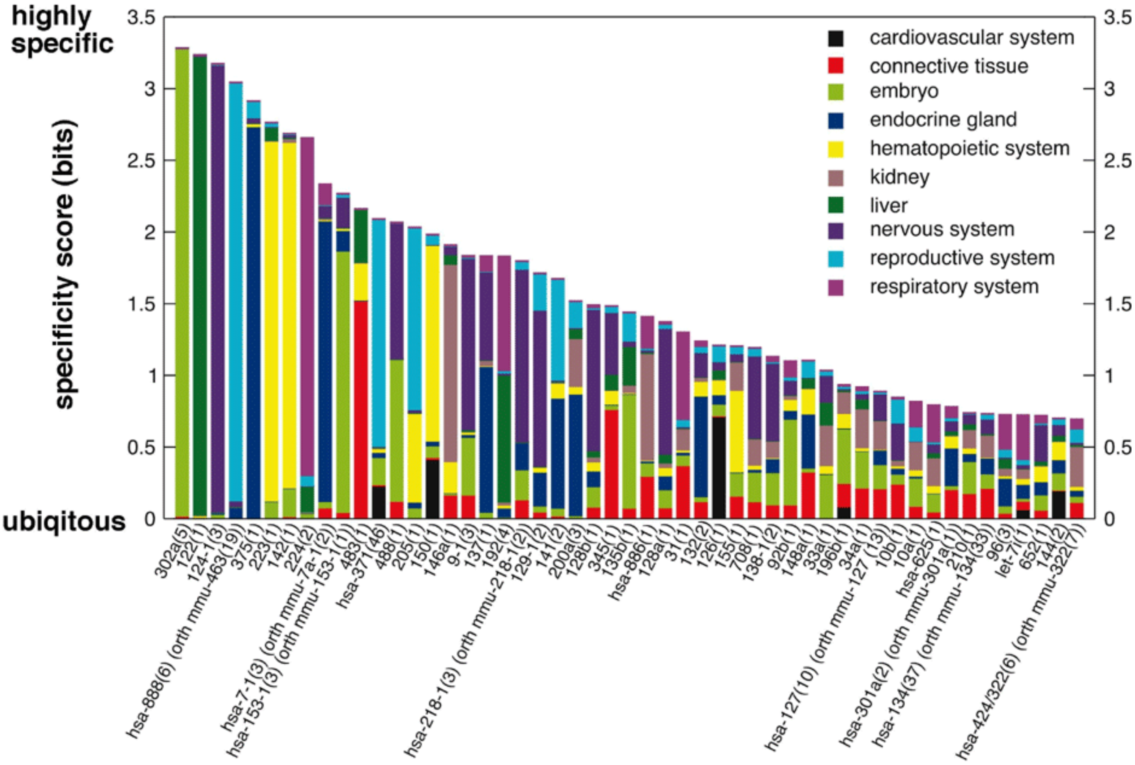


Figure 1.4: The 51 most specific miRNA precursor clusters for human cells and tissues (x-axis). Different tissue types are given (colored). Tissue specificity is indicated by the total height of each bar (specificity score) on the y-axis. The relative heights for each of the tissues are proportional to sequence reads of a miRNA precursor cluster in a given tissue type relative to all tissue types. Graph adapted from Landgraf et al., 2007.

The binding efficiency and efficacy of miRNAs to RNA targets depends partly on sequence contexts, and on a ranked hierarchy of seed sequence binding: 8mer»7mer>7mer-A1»6mer (Grimson et al., 2007; Nielsen et al., 2007). In general, binding efficiency is improved when RNA target sequences are located (i) within the 3'UTR and at least 15 nt away from the stop codon; (ii) in an AU-rich environment or (iii) close to each other (Bartel, 2004; Grimson et al., 2007). Target sites located within 40 nt, but not closer than 8 nt, tend to act cooperatively and significantly enhance repression over the level expected from their independent contributions (Saetrom et al., 2007).

## 1.2 Mutations in genetically-encoded components of post-transcriptional regulatory networks cause disease

Theoretically, mutations in genes encoding RBPs and miRNAs may perturb post-transcriptional regulatory networks, resulting in phenotypic changes or disease (Nicoloso et al., 2010). Alternatively, mutations of the transcripts and target sites, respectively, could alter RBP:RNA interaction, resulting in a partially overlapping phenotype.

Several neurological disorders, including Fragile X Syndrome (FXS), Frontotemporal Lobar Dementia (FTLD), and Amyotrophic Lateral Sclerosis (ALS) are caused by mutations in genes encoding RBPs, namely FMRP, TAR DNA Binding Protein (TDP43), and the Fused in Sarcoma (FUS) protein (Pieretti et al., 1991; Licatalosi and Darnell, 2006; Keene, 2007; Sreedharan et al., 2008; Kwiatkowski et al., 2009).

Mutations in miRNA genes also cause disease. Mutations in the miR-96 seed region cause non-syndromic progressive hearing loss (Mencía et al., 2009; Friedman and Avraham, 2009). miRNA 15a and 16 are deleted in more than half of B cell chronic lymphocytic leukemias (CLL) (Calin et al., 2008). miR-375, a regulator of glucose-induced insulin secretion, may be affected in Type 2 Diabetes mellitus (Poy et al., 2004). Disease-associated miRNA profiles are also reported for Alzheimer and Parkinson disease (Fiore et al., 2008).

Sequence variations or mutations in miRNA target sites potentially destroy or modify target sites with variable effects on transcript stability and translation (Lewis et al., 2005; Brennecke et al., 2005; Li et al., 2010; Rahman et al., 2010). We also expect that SNPs have a significant effect on RNA regulation; approximately ten million SNPs, corresponding to one variant per 3,000 bases, are estimated to constitute 90 % of the variation in the human genome. SNPs may abrogate, strengthen or weaken RBP:target and miRNA:target RNA interactions.

Remarkably, the prevalence of autism is 15-30 % in FXS patients and 4 % in NF1 patients (Kelleher and Bear, 2008); conversely, the prevalence rates for FXS and NF1 in persons with autism are 2-3 % and 0-4 %, respectively (Gillberg and Forsell, 1984; Mouridsen et al., 1992; Fombonne et al., 1997). Since FMRP is absent in FXS and results in autism, mutation of a target transcript, such as NF1, could result in a similar endophenotype.

In this thesis, we are interested in the significant clinical overlaps of the single-gene disorders FXS and neurofibromatosis type 1 (NF1) with the multigenic disorder autism (Laxova, 1994; North et al., 1997; Berry-Kravis, 2002; McConkie-Rosell et al., 2005; Budimirovic et al., 2006; Saemundsen et al., 2007).

### 1.2.1 Fragile X syndrome and FMRP

Fragile X syndrome is the most frequent form of heritable mental retardation, with a characteristic profile of autistic-like features, including deficits in imaginative play, repetitive speech, hand biting and scratching, as well as, gaze aversion and social anxiety (Turk and Cornish, 1998). FXS is typically caused by abrogation of FMRP expression, an RNA-binding protein encoded by the fragile X mental retardation 1 (FMR1) gene (Verkerk et al., 1991; Verheij et al., 1993). The majority of FXS cases (99 %) are caused by an increased number of CGG trinucleotide segments (> 200 repeats) in the 5' end of the FMR1 gene located on the X-chromosome (McConkie-Rosell et al., 2005). However a single point mutation (substitution) in the KH domain of FMRP causes FXS by interfering with RNA-binding (Boulle et al., 1993).

FMRP has three RBDs (two KH domains and one RGG box) and recognizes up to 4 % of mRNAs in human fetal brains through direct interaction with G quartet sequences, U-rich sequences and so called ”kissing complexes”, which are relatively big secondary structures of at least 60 nucleotides (Ashley et al., 1993; Brown et al., 1998; Chen et al., 2003; Darnell et al., 2005). The main function of FMRP is regulatory repression of translation by increasing the affinity for capped mRNA of its binding partner CYFIP1 and eIFAE (Laggerbauer et al., 2001; Napoli et al., 2008). FMRP is a critical requirement for the regulation of localization of a subset of dendritic mRNAs, which may influence the quality and efficacy of mRNA dynamics (Antar et al., 2006; Bassell and Warren, 2008; Dictenberg et al., 2008) and synaptic plasticity (Bear et al., 2004; Muddashetty et al., 2007; Waung et al., 2008; Castillo et al., 2008).

### 1.2.2 Neurofibromatosis type 1

Neurofibromatosis is a genetic disorder of neural crest-derived cells in neuroectodermal and mesenchymal tissues (Gabhane et al., 2010). This disorder is caused by mutations in the 288 kb NF1 gene, located on chromosome 17, that encodes the protein Neurofibromin. This protein is expressed in many tissues, including brain, kidney, spleen, or thymus, and acts mainly as a tumor suppressor and limits cell growth.

Mutations of the NF1 gene lead to cell overgrowth and tumor development; especially malignant peripheral nerve sheath tumors, neuroendocrine tumors, and nonneurogenic gastrointestinal stromal tumors (Cohen and Shuper, 2010; Cavallaro et al., 2010). The majority (82 %) of NF1-causing mutations are either frameshift or nonsense mutations (Shen et al., 1996); approximately 1 to 5 % of NF1 patients have large deletions (> 700 kb) (Tonsgard et al., 1997).

### 1.2.3 Autism spectrum disorder

Autism spectrum disorders (ASD) are highly variable neurodevelopmental diseases that are subclassified as autism, Asperger syndrome and atypical pervasive developmental disorders (PDD) (Mayes et al., 1993; Muhle et al., 2004; Myers et al., 2007; Geschwind, 2008). ASD is characterized by widespread abnormalities of (i) language, communication and imaginative play, (ii) social interactions, and (iii) restricted interests and activities (Noens et al., 2006; Landa, 2007; Rogers, 2009).

In contrast to FXS and NF1, the causes of autism are predominantly unkown. At present many genes (> 100) and several chromosomal regions have been investigated for their involvement in autism, however most mutations are still not identified (Losh et al., 2008a,b). SNPs and de novo copy number variations (CNVs) have been found in 24 % of autism patients and likely cause a subset of autism (Sebat et al., 2007; Allen-Brady et al., 2009).

Autism is likely caused by a series of de novo mutations (Sebat et al., 2007). Reported mutations potentially affect cell-adhesion, as well as synaptogenesis, leading to abnormal formation of synapses and dendritic spines and synaptic dysfunctions (Persico and Bourgeron, 2006; Kelleher and Bear, 2008; Levy et al., 2009). Candidate autism genes encode for serotonin transporters (i.e. SLC6A4), neuroligins (NLGN3/ 4), or tumor suppressors (PTEN, TSC1 and TSC2) (Jamain et al., 2003; Coutinho et al., 2004; Butler et al., 2005; Tavazoie et al., 2005; Buxbaum et al., 2007).

## 1.3 Sequencing human genomes and bioinformatics

Next-generation sequencing (NGS) technologies have transformed our experimental approaches and the understanding of genetic (Wang et al., 2006).

Many different high-throughput genotyping approaches, consisting of various combinations of different allele-discrimination chemistries and signal detection methods, have been developed. These approaches have been widely used to discover SNPs and CNVs, for epigenotype analysis, and for the discovery of new genetic polymorphisms (Tsuchihashi and Dracopoli, 2002; Corona and Toffoli, 2004). They have replaced classic approaches like restriction fragment length polymorphism (RFLP) analysis, using Southern blotting, and pre-amplification by PCR (Shumaker et al., 1996; Germer and Higuchi, 1999; Kwok, 2000), and are better than classical genotyping platforms based on microarrays, genechips, Real-time PCR assays, and mass spectrometry (Meyer et al., 2009; Araujo, 2009; Dettman et al., 2010). Most of these approaches are still limited by relatively low sensitivity and efficiency (Li et al., 2007); accuracy, speed and costs are further limiting factors (Jenkins and Gibson, 2002).

NGS enables analyses of genomic sequences with much higher coverage, accuracy and efficiency (Kidd et al., 2008; Tucker et al., 2009; Futschik and Schlötterer, 2010), and will likely be the method used to identify the causes of diseases such as autism. Several massively parallel sequencing devices are currently available; namely the Illumina Genome Analyzer[2], the Applied Biosystems SOLiD Sequencer[3], and the Roche 454 Genome Sequencer[4]. Advantages of these are (i) massive automated parallel performance, (ii) the detection of minor alleles, and (iii) non-reliance on cloning and amplification steps.

Bioinformatic approaches are also rapidly evolving to handle large volumes of sequencing data (Mardis, 2008; Voelkerding et al., 2009). To improve the workflow from data generation through to analysis and publication, respectively, streamlined pipelines are being developed, including multiple sequence alignment algorithms, mapping tools, and tools for converting sequencing data into an uniform format (Moore et al., 2010; Koboldt, 2010).

[2] http://www.illumina.com/; available at March 3, 2010

[3] http://www.appliedbiosystems.com/absite/us/en/home/applications-technologies/solid-next-generation-sequencing.html; available at March 4, 2010

[4] http://www.454.com/; available at March 3, 2010

## 1.4 Scope of this thesis

The goal of this project is to develop a rapid screening method to analyze genetic variation in RBP and miRNA target sites using patient and control DNA samples, massive parallel 454 sequencing, and bioinformatic approaches. The assay will be validated by investigating sequence variations in FMRP and miRNA target sites of the NF1 gene in DNA obtained from persons with and without autism. We will then modify the assay to analyze highly expressed miRNA genes using DNA samples from hundreds of patients with chronic lymphocytic leukemias, Type 2 Diabetes mellitus and autism. Should no variation be identified we still have a useful method to assess variations in other genomic regions. Should variations be found, the knowledge of the abundance of variation in these evolutionally conserved and functional regions will enable us to calculate how many samples are necessary for subsequent disease association studies. Mutated genes may reveal which genomic regions, pathways etc. are responsible for the clinical overlap between diseases. Out study will enhance disease classification and diagnostics, assist in elucidating pathogenesis, and delineate potential targets for therapeutic intervention.

# 2 Materials

## 2.1 Clinical materials and demographic data

Samples from persons with and without autism (or ASD) were obtained from the Centre for Collaborative Genetic Sudies, National Institute of Mental Health[1] (NIMH), Bethesda, USA. Four hundred persons were selected at random from a total population of approximately 3500 persons. The study group consisted of 134 persons with autism and 266 participants without autism (control participants); most were non-affected relatives. Specific demographic information for each patient or relative was provided including year of birth (age), gender, race, age of disorder onset, and progeny status (twin: monozygotic or dizygotic). The 400 participants were drawn from 149 families; an example of pedigree information is shown in Figure 2.1.

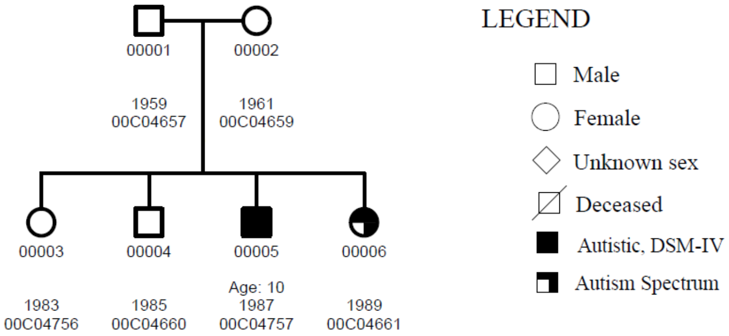


Figure 2.1: Typical pedigree and demographic information for each family. NIMH subject ID, age at assessment, year of birth, and sample ID are indicated in descending order below each symbol.

DNA samples from 768 patients with B cell chronic lymphocytic leukemias (CLL) and Type 2 Diabetes mellitus were obtained from the Stoffel Laboratory, Institute of Molecular Systems Biology (ETH Zurich, Switzerland), and Dr. Pablo Landgraf (HHU Düsseldorf, Germany). All samples were either provided in barcoded tubes or in 96-well plates at a concentration of 200 ng/$\mu$l.

[1] http://www.nimh.nih.gov/index.shtml; available at March 16, 2010.

## 2.2 Oligonucleotides

Primer pairs were designed and optimized using the Primer3 algorithm[2] v. 0.4.0 (Rozen and Skaletsky, 2000). Primers were synthesized by Integrated DNA Technology, Inc. (Coralville, USA) and delivered in 96-well plates in water at a concentration of 100 $\mu$M. Primers were diluted to a working concentration using distilled water. Primer sets used in this thesis for the NF1 and miRNA gene project are listed in Appendix 1 and 2.

Seventeen regions of the NF1 gene were amplified in a multiplexed PCR format using 17 target-specific primer pairs (Table 6.1). Twenty-four regions of the NF1 gene were amplified in nested PCR reactions using 24 target-specific primer pairs (Table 6.2).

Seventeen regions encoding 17 miRNA genes were amplified in a multiplexed PCR format using target-specific primer pairs (Table 6.3). Seventeen regions encoding 17 miRNA genes were amplified in nested PCR reactions using target-specific primer pairs (Table 6.4).

## 2.3 Chemicals

A list of chemicals used in this thesis provided in Table 2.1.

Table 2.1: List of chemicals used in this thesis. Described are the chemical and commercial names; the list is separated by manufacturer.

| Chemical | Commercial name | Manufacturer |
|---|---|---|
| Glycerol | Glycerol | Fisher Scientific |
| KCl | Potassium chloride, purity = 99.4 % | |
| NaCl | Sodium chloride | |
| Tris base | Tris (hydroxymethyl) aminomethane | |
| Triton X-100 | $C_8H_{17}(C_6H_4)(OCH_2CH_2)XOH$ | |
| DDT | Dithiothreitol, purity = 99.0 % | Sigma |
| EDTA | Ethylenediaminetetraacetic acid, purity = 99.995 % | |
| HCl | Hydrochloric acid | |
| $MgCl_2$ | Magnesium chloride hexahydrate, purity = 99.0 % | |
| $MgSO_4$ | Magnesium sulfate purity = 99.0 % | |
| Agarose | NuSieve® GTG® Agarose | Lonza |
| Lmp-Agarose | Seakem® LE Agarose | |

[2] http://frodo.wi.mit.edu/primer3; available at February 20, 2010

## 2.4 Reagent kits and enzymes

All reagent systems used in this thesis are listed in Table 2.2.

Table 2.2: List of reaction systems (Kits) and enzymes used in this thesis.

| **Designation** | **Manufacturer** |
|---|---|
| AccuPrime$^{TM}$ Reaction Mix | Invitrogen Life Science, USA |
| *Pfx* polymerase | Invitrogen Life Science, USA |
| *Taq* polymerase | Tuschl Laboratory, USA |

## 2.5 Buffers and solutions

All buffers and solutions used in this thesis are listed in Table 2.3.

Table 2.3: Composition of buffers and solutions.

| **Buffer/ solution** | **Composition** |
|---|---|
| PCR reaction buffer (10x) | 10 mM b-mercaptoethanol |
| | 500 mM KCl |
| | 20 mM $MgCl_2$ |
| | 100 mM Tris HCl (pH 8) |
| | 1 % Triton-X-100 |
| *Taq* stocking buffer | 100 mM NaCl |
| | 0.1 mM EDTA |
| | 5 mM DTT |
| | 50 mM Tris HCl (pH 8) |
| | 50 % Glycerol (v/v) |
| | 1 % Triton-X-100 (v/v) |
| TBE buffer (1x) | 89 mM Tris base |
| | 20 mM EDTA (pH 8) |
| | 8 mM boric acid |
| 5x Loading Buffer | 0.2 % bromophenol blue |
| | 0.2 % xylene cyanol FF |
| | 50 mM EDTA (pH) 8 |
| | 20 % Ficoll type 400 |

## 2.6 DNA ladders

Commercial DNA ladders (New England Biolabs, Inc., UK.) were used to assess the size and concentration of PCR products and are shown in Figure 2.2.

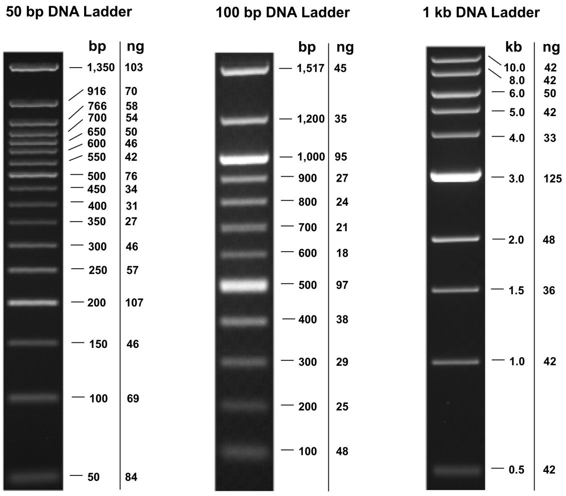


Figure 2.2: Commercial DNA ladders as visualized using ethidium bromide staining and agarose gel electrophoresis.

## 2.7 Equipment

Equipment used in this thesis is listed in Table 2.4.

Table 2.4: Equipment used in this thesis.

| Equipment | Manufacturer |
|---|---|
| EpMotion 5075® Automated Pipetting Station | Eppendorf, Germany |
| Eppendorf Vacufuge$^{TM}$ vacuum centrifuge | Eppendorf, Germany |
| LAS-3000 imaging system | Fujifilm Corporation, Japan |
| Mastercycle® ep384 thermal cycler | Eppendorf, Germany |
| Mastercycle® epgradient S thermal cycler | Eppendorf, Germany |
| OWL Easycast$^{TM}$ electrophoresis chamber | Thermo Scientific, USA |
| PCR plates Thermo-Fast 96, Non-Skirted, 0.2 mL | Thermo Scientific, USA |

Table 2.4: – continued on next page

Table 2.4: – continued from previous page

| Equipment | Manufacturer |
|---|---|
| PowerPac 3000 | Bio-Rad Laboratories, USA |
| SmartSpec$^{TM}$ Plus Spectrophotometer | Bio-Rad Laboratories, USA |
| Spectroline® UV Transilluminator | Spectroline, USA |
| Table centrifuge Sorvall Biofuge fresco | Thermo Scientific, USA |
| Thermomixer | Eppendorf, Germany |
| Vortex-Genie 2 Model G560 | Scientific Industries, USA |

## 2.8 Software and databases

Different software tools and databases were used for bioinformatic applications. Each software and database, including field of application (function) and manufacturer or reference, is listed in Table 2.5.

Table 2.5: Software and databases used in these studies.

| Software/ database | Manufacturer | Function |
|---|---|---|
| Burrows-Wheeler Aligner | (Li and Durbin, 2009, 2010) | Aligner |
| EpBlue$^{TM}$ | Eppendorf, Germany | Automated pipetting |
| GigaBayes | Boston College, USA | SNP discovery |
| GS Amplicon | Roche, Switzerland | Variation discovery |
| GS Mapper | Roche, Switzerland | Mapper/ aligner |
| LAS-3000 ImageReader Pro | Fujifilm Corporation, Japan | Gel imaging/ documentation |
| miRBase | (Griffiths-Jones et al., 2008) | Database (miRNA) |
| MOSAIK applications | Boston College, USA | Assembler/ aligner |
| Oligomap | (Berninger et al., 2008) | RNA sequence annotation |
| Primer3 | (Rozen and Skaletsky, 2000) | Primer design |
| PuTTY | Tatham Team, England | SSH and telnet client |
| PyroBayes | Boston College, USA | Rewrite/ transformation |
| SAMtools-0.1.7 | (Li et al., 2009) | Multi sequence alignment |

# 3 Methods

## 3.1 PCR strategy

To assess sequence variation in multiple genomic regions in hundreds of individuals, we designed a nested multiplexed PCR assay with barcoded pyrosequencing. The first round of PCR amplification consisted of multiplexed PCR in which amplicons of approximately 1 kb in size were generated. These products were diluted 10-fold and served as templates for subsequent singleplex nested PCR reactions in which patient-specific barcodes were introduced. Barcoded (patient-specific) amplicons, of approx. 200 bp in size, were pooled and sequenced.

We designed assays for 2 projects to assess sequence variations in (i) RBPs and miRNA target sites in the NF1 gene and (ii) miRNA gene sequences. The first project, referred to as the NF1 project, was established in a 96-well plate format using an 18-multiplexed PCR, followed by 24 nested singleplex PCRs for each of the 400 DNA samples obtained from persons with and without autism. Ninety-six-well PCR reactions were performed using a Mastercycle® epgradient S thermal cycler (Eppendorf, Germany).

The second project, referred to as the miRNA gene project, was established in a 384-well plate format to enable higher sample throughput using an 17-multiplexed PCR, followed by 17 nested singleplex PCRs for each of the 768 DNA samples obtained from persons with chronic lymphocytic leukemias, Type 2 Diabetes mellitus and autism. Three hundred and eighty four well PRC reactions were performed using a Mastercycle® ep384 thermal cycler (Eppendorf, Germany).

PCR reactions and cycling conditions were optimized using DNA (100 ng/$\mu$l) from human tumor raji cell lines and the Mastercycle®ep384 and epgradient S thermal cycler. Two positive and two negative controls were present on each PCR plate; DNA from human tumor raji cell lines functioned as positive control, whereas water served as negative control. Both positive and negative controls were amplified using the forward (F1) and reverse primer (R1) for each specific target.

## 3.2 Fusion primer design

For the NF1 project, twenty-four target sites were evaluated; 15 FMRP target sites were identified through PAR-CLIP analysis and 9 miRNA target sites were detected using Target Scan analysis (personal communication Ascano and Renwick).

The reference sequence for primer design was obtained from the NCBI[1] database (NG_009018.1; GI:213385299; 01-NOV-2009). For the miRNA gene project, seventeen miRNA genes were selected from the miRBase[2] (Griffiths-Jones, 2004; Griffiths-Jones et al., 2006, 2008) database.

First, target-specific primers were designed using the Primer3 algorithm[3] v. 0.4.0 (Breslauer et al., 1986; Rozen and Skaletsky, 2000) with an optimum length of 20 nt, annealing temperature (Tm) at 60 °C, GC content between 20 - 80 %, a maximum allowable local alignment score for self-complementarity of eight, and a maximal mononucleotide repeat (poly-X) score of five.

Following successful amplification of a PCR product of expected size, fusion primers, consisting of a 454 sequence adapters (see Figure 3.2), patient-specific barcode sequences, and target-specific primer sequences, were designed (Table 3.1). All fusion barcode sequence combinations are shown in Figure 4.1; each target-specific primer sequence is listed in Chapter 2.

Table 3.1: Fusion primer design; 454-adapter sequence (underlined), barcode sequence (italicized) and target-specific primer sequence.

| | **454-adapter** | **Barcode** | **Target-specific sequence** |
|---|---|---|---|
| i2231-F | 5' -GCCTCCCTCGCGCCATCAG | *AACCGCAT* | AACTGTCACAGCCCGACTCT- 3' |
| i2231-R | 5' -GCCTTGCCAGCCCGCTCAG | *TAATCCGG* | AGCAACAAACCCCAAATCAA- 3' |

## 3.3 Multiplexed PCR

The multiplexed (18-plex) PCR for the NF1 project was prepared in 96-well plates. Each well contained a total volume of 50 $\mu$l, containing 100 ng DNA, 1.25 U *Pfx* DNA Polymerase, Reaction Mix (10 mM $\mathrm{MgSO_4}$, 0.3 mM dNTP), and a total primer concentration of 720 nM (20 nM each).

The multiplexed (17-plex) PCR for the miRNA gene project was prepared in 384-well plates. Each well contained a total volume of 10 $\mu$l, containing 100 ng DNA, 1.25 U *Pfx* DNA Polymerase, Reaction Mix (10 mM $\mathrm{MgSO_4}$, 0.3 mM dNTP), and a total primer concentration of 3.4 $\mu$M (100 nM per primer).

For each plate two positive and two negative controls were added. Optimized PCR reaction and cycle conditions are shown in Table 3.2 and Table 3.3, respectively.

[1] National Center for Biotechnology Information; http://www.ncbi.nlm.nih.gov/; available for entire periode

[2] http://www.mirbase.org/; available for entire periode

[3] http://frodo.wi.mit.edu/primer3; available at February 20, 2010

Table 3.2: Multiplexed PCR reaction conditions for 96-well and 384-well formats.

| **Components** | **Volume 96-well plates** | **Volume 384-well plates** |
|---|---|---|
| Reaction Mix (10x) | 529 $\mu$l | 520 $\mu$l |
| *Pfx* DNA Polymerase | 53 $\mu$l | 52 $\mu$l |
| Primer Mastermix | 48 $\mu$l (720 nM) | 260 $\mu$l (3.4 $\mu$M) |
| $dH_2O$ | 4662 $\mu$l | 1508 $\mu$l |
| | 5292 $\mu$l (for 100 + 8 wells) | 2340 $\mu$l (for 256 + 4 wells) |

Table 3.3: Multiplexed PCR cycle conditions.

| **Conditions** | **Guidelines** |
|---|---|
| **Denaturation** | Temp: 95 °C. Time: 2:00 min on first cycle; 45 s remainder |
| **Annealing** | Temp: 55 °C. Time: 1:15 min |
| **Extension** | Temp: 68 °C. Time: 1:30 min; 5 min on last cycle |
| **Number of cycles** | 35 cycles for NF1 targets; 38 cycles for miRNA project |

## 3.4 Nested PCR

The nested PCR step consisted of 24 singleplex PCR reactions for the NF1 project and 17 singleplex PCR reactions for the miRNA gene project. Singleplex PCRs for the NF1 project were prepared in 96-well plates. Each well contained a total volume of 25 $\mu$l, containing PCR reaction mix (Table 3.4), 1 $\mu$l DNA template (multiplexed PCR product; 1:20 diluted in $dH_2O$), and the barcoded primer pair (Fx and Rx) at a total concentration of 200 nM.

Singleplex PCRs for the miRNA gene project were prepared in 384-well plates. Each well contained a total volume of 10 $\mu$l, containing PCR reaction mixture, containing 1 $\mu$l DNA template (multiplex PCR product; 1:20 diluted in $dH_2O$), PCR reaction mix, and the barcoded primer pair (Fx and Rx) at a total concentration of 200 nM.

Reaction buffer and *Taq* polymerase were prepared in the Tuschl laboratory (Rockefeller University, New York). The reaction buffer (10x) contained 10 mM $\beta$-mercaptoethanol, 500 mM KCl, 20 mM $MgCl_2$, 1 % Triton-X-100 and 100 mM Tris HCl (pH 8). The *Taq* polymerase was stored at -80 °C in a buffer consisting of 1 % Triton-X-100, 100 mM NaCl, 5 mM DTT, 0.1 mM EDTA, 50 % Glycerol and 50 mM Tris HCl (pH 8). Optimized PCR reaction and cycle conditions are shown in Table 3.4 and Table 3.5, respectively.

Table 3.4: Singleplex PCR reaction conditions for 96-well and 384-well formats.

| Components | Volume 96-well plates | Volume 384-well plates |
|---|---|---|
| Reaction Buffer (10x) | 270 $\mu$l | 240 $\mu$l |
| *Taq* DNA Polymerase (0.6 U) | 54 $\mu$l | 48 $\mu$l |
| 2 mM dNTPs (10x) | 270 $\mu$l | 240 $\mu$l |
| $dH_2O$ | 1458 $\mu$l | 1292 $\mu$l |
|  | 2052 $\mu$l (for 100 + 8 wells) | 1820 $\mu$l (for 256 + 4 wells) |

Table 3.5: Nested PCR cycling conditions.

| Conditions | Guidelines |
|---|---|
| **Denaturation** | Temp: 95 °C. Time: 2:00 min on first cycle; 45 s remainder |
| **Annealing** | Temp: 55 °C. Time: 1:15 min |
| **Extension** | Temp: 72 °C. Time: 1:00 min; 5 min on last cycle |
| **Number of cycles** | 35 cycles |

## 3.5 Automated pipetting system

To increase sample throughput, we used an EpMotion 5075® Automated Pipetting Station (Eppendorf, Germany) to assemble PCR reactions, dispense barcoded primer combinations, and pool PCR products. Pipetting procedures were performed using the EpBlue™ software (Eppendorf, Germany). Reaction mixtures were dispensed using a 300 $\mu$l multi-dispense pipette with filter tip change after 8 aspirations. Primers were dispensed using a 50 $\mu$l pipette with filter tip change after each aspiration. Diluted PCR amplification products were added to plates using an 8-channel multipipette and 50 $\mu$l filter tips, followed by 3 mixing cycles of one half of total volume at a speed of 7.0 mm/s. Reaction mixtures in 96/ 384-well plates were incubated at 4 °C on Thermostat Plus® thermoblocks (Eppendorf, Germany). An example of a worktable layout is shown in Figure 3.1.

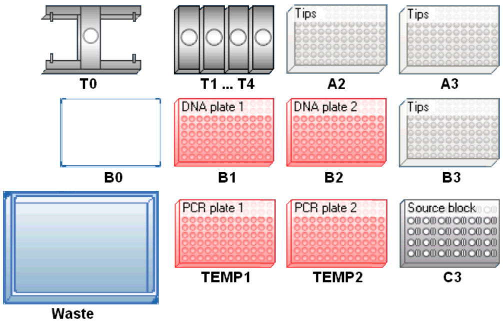


Figure 3.1: Worktable layout of the EpMotion 5075® Automated Pipetting Station (Eppendorf, Germany). Labware: A2/A3/B3 = 50 $\mu$l filter tips; B1/B2 = EP TT PCR racks for 96/384-well plates; C3 = 1.5 ml rack; TEMP1/2 = EP TT PCR destination 96/384-well racks, temperature-controlled; T0 = plate mover; T1-4 = pipette holder for TM_50 ($\mu$l) , TM_50_8 ($\mu$l, multichannel), TM_300 ($\mu$l), TM_1000 ($\mu$l); and waste can. Position B0 was not used.

## 3.6 Agarose gel electrophoresis

PCR products were visualized using ethidium bromide staining and agarose gel electrophoresis; Seakem® LE agarose in 1x TBE buffer (89 mM Tris base, 20 mM EDTA (pH 8), 8 mM boric acid) and ethidium bromide staining (0.01 mg/ml final concentration). Running conditions were commonly 150 to 200 V for 30-45 minutes in an OWL Easycast$^{\mathrm{TM}}$ electrophoresis camber (Thermo Scientific, USA). PCR products were visualized on the LAS-3000 imaging system (Fujifilm Holdings Corporation, Japan).

## 3.7 Generation of quantitative marker and estimation of pooled PCR product concentrations

We generated a quantitative size marker to estimate pooled PCR product concentrations for each target region prior to the pooling step. We selected a PCR reaction that generated a product of 235 bp (region i2209 plus fusion primers), without visible primer dimerization, and performed 50 cycles of amplification to ensure consumption of PCR primers. Using a primer concentration of 200 nm in a 50 $\mu$l reaction, we generated a product at a concentration of 15.5 ng/$\mu$l; marker concentration was 13 ng/$\mu$l following addition of 5x Loading Buffer. Ten microliter aliquots of each purified nested PCR product were run on a 2 % agarose gel; concentrations were estimated visually by comparing band intensities with marker bands corresponding to 10 - 60 ng amounts of DNA. The intensities of our self-made marker (smM) was consistent with those ssen for a commercially prepared 100 bp DNA ladder (New England BioLabs Inc).

## 3.8 Gel purification of pooled PCR products

Following nested PCR, 2 $\mu$l of each patient-specific PCR reaction were pooled per nested PCR region. The pooled products of expected size were gel-purified. Forty microliter aliquots of each pool were run on a 3 % low-melting point NuSieve®GTG®agarose gel (Lonza, Switzerland). DNA was visualized using a Spectroline® Transilluminator set at a wavelength of 365 nm. Each band was excised in a gel slice weighing approximately 100 mg and transferred to a 1.5 ml reaction tube.

### 3.8.1 Phenol-chloroform extraction

Five hundred microliters NaCl (0.4 M) were added to each gel slice. Following incubation for 10 min at 70 °C, we added 500 $\mu$l of pre-heated buffered water-saturated phenol (pH 7.8). Following vigorous vortexing, phases were immediately separated in a tabletop centrifuge for 5 min at maximum speed (11,000 rpm) and 4 °C. The aqueous upper phase was collected, extracted with 500 $\mu$l basic phenol/chloroform/isoamylalcohol (25:24:1), and vigorously vortexed. The aqueous upper phase was collected and 500 $\mu$l chloroform was added.

### 3.8.2 Ethanol precipitation

After removing the chloroform phase, nucleic acids were precipitated by adding one tenth volume of 3 M sodium acetate buffer (pH 5.2; store at 4 °C) and two volumes of absolute ethanol (stored at -20 °C), and incubated either at -20 °C overnight or at 80 °C for 25 min. Samples were centrifuged for 30 min at 11,000 rpm in a pre-cooled centrifuge (4 °C). The supernatant was removed carefully. Pellets were washed in 200 $\mu$l 70 % (v/ v) ethanol, followed by centrifugation at 11,000 rpm for 15 min. The remaining liquid was removed using a pipette. Pellets were vacuum dried using an Eppendorf Vacufuge$^{TM}$ vacuum centrifuge (approx. 30 min at 30 °C), resuspended in 40 $\mu$l of $dH_2O$, and stored at -4 °C.

## 3.9 Quantitation

Concentrations of purified pooled products were measured by using a SmartSpec$^{TM}$ Plus Spectrophotometer (Bio-Rad Laboratories, USA). Band intensities and the quality of purification were assessed using quantitative agarose gel electrophoresis (Section 3.7).

## 3.10 Next-generation (454) sequencing

Bidirectional sequencing of pooled and purified nested PCR amplicons was performed using massive parallel 454 sequencing (454 Life Sciences)[4], through the Genome Research Center at the Memorial Sloan-Kettering Cancer Center (MSKCC, New York). An overview of this sequencing approach is provided in Figure 3.2.

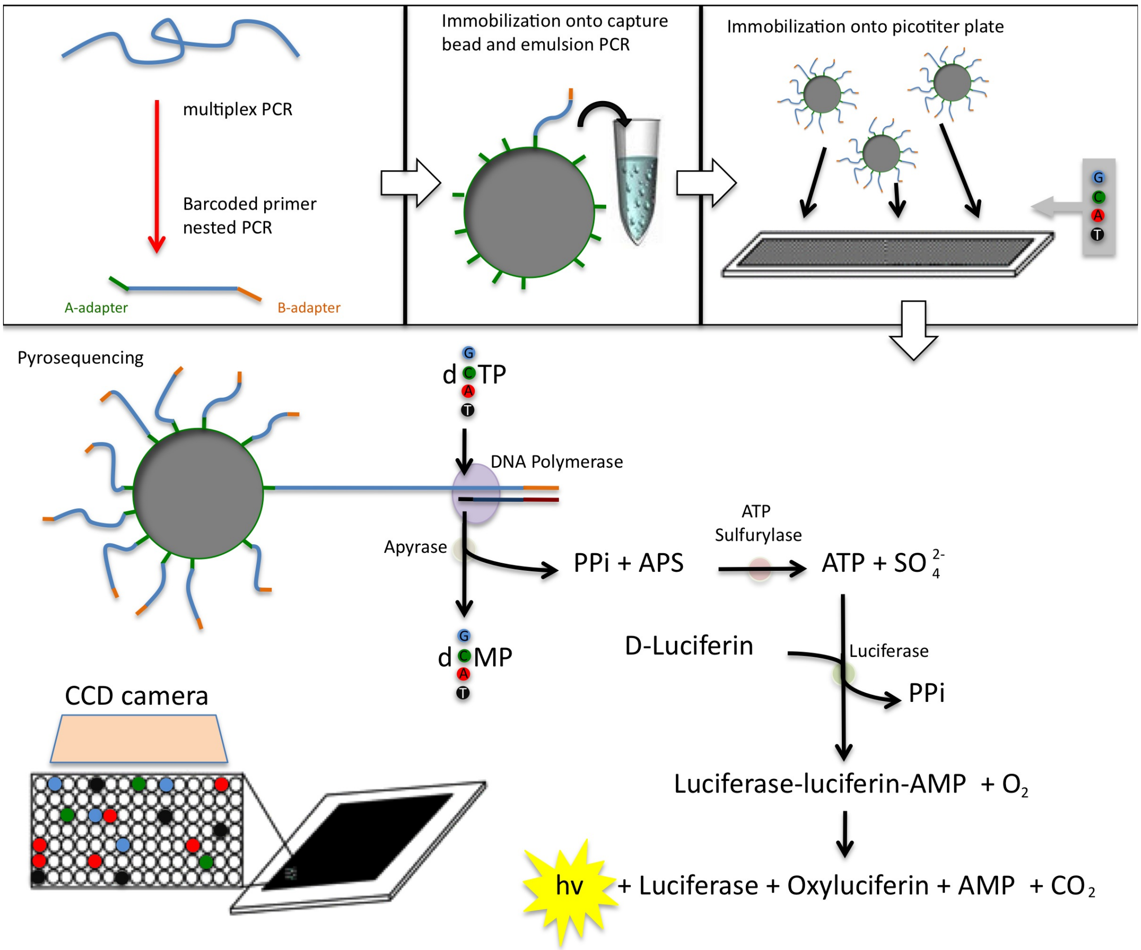

multiplex PCR
Barcoded primer nested PCR
A-adapter
B-adapter
Immobilization onto capture bead and emulsion PCR
Immobilization onto picotiter plate
G
C
A
T
Pyrosequencing
dGCATTP
DNA Polymerase
Apyrase
PPi + APS
ATP Sulfurylase
ATP + SO$_4^{2-}$
dGCATMP
D-Luciferin
Luciferase
PPi
CCD camera
Luciferase-luciferin-AMP + O$_2$
hv
+ Luciferase + Oxyluciferin + AMP + CO$_2$

Figure 3.2: Scheme of massive parallel 454 sequencing. Steps include generation of PCR products for sequencing, immobilization onto beads and emulsion PCR, immobilization of products onto picotiter plates, pyrosequencing, signal detection (e.g. through a CCD camera), and generation of sequence flowgram format (.SFF) files.

[4] http://www.454.com/; available at March 3, 2010

## 3.11 Data management and bioinformatics

Most computer systems (e.g. LINUX and Mac OS) are based on the computer operating platform UNIX because of its simplicity, speed and reliability (Ritchie and Thompson, 1983). We used several methods for data management and bioinformatic analyses; mainly based on UNIX. First, we used the Genome Sequencer (GS) application "Reference Mapper" (454 Life Sciences Corporation) to analyze and assess 454 sequencing .SFF files. To assess the quality of data produced by our nested multiplexed PCR assay, based on the presence of target-specifc primers, target sequences and barcode sequences, a bioinformatic pipeline was developed; consisting of (i) PyroBayes, (ii) Oligomap, and (iii) the MOSAIK package. Specific perl scripts (written by Volker Hovestadt) were used to link each of these applications and to generate trimmed sequences and statistics. Preliminary SNP and short-INDEL polymorphism analyses were performed using the GigaBayes application and the GS "Amplicon Variant Analyzer", respectively. Bioinformatic workflow is summarized in Figure 3.3.

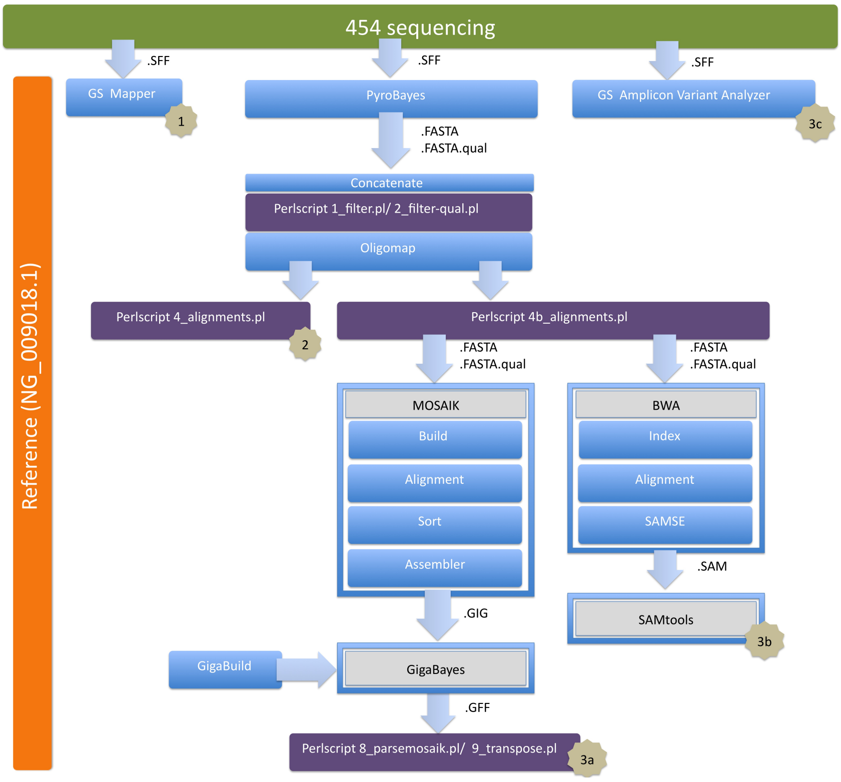


Figure 3.3: Workflow for data management and bioinformatic analyses. Numbered stars indicate key steps: (1) analysis of .SFF sequencing files, (2) distribution of target-specifc primers, target sequences and barcode sequences, and (3) alignment and variation results through various algorithms/ tools (a,b,c). Arrows indicate the direction of workflow; data file formats are indicated next to these arrows.

### 3.11.1 Roche Genome Sequencer Tools

Sequencing trace files from 454 pyrosequencing were provided in a standard flowgram format (.SFF) for subsequent analysis using the Genome Sequencer (GS) software package (454 Life Sciences Corporation). This system is accessible using a Command Line Interface (CLI) and Graphical User Interface (GUI). In this thesis two GS applications were used: (i) GS Reference Mapper and (ii) GS Amplicon Variant Analyzer (AVA).

The GS Reference Mapper aligns sequencing reads against a reference sequence with or without associated annotations, generates a consensus sequence, and computes read statistics. The mapper computation was done with the following settings for overlap detection; seed step = 12, seed length = 16, seed count = 1, and hit-per-seed limit = 70. The minimum overlap length was 40 with 100 % overlap identity. Additionally a repeat score threshold of 12 was applied. The Homo sapiens neurofibromin 1 gene (NG_009018.1; GI:213385299; 01-NOV-2009) was used as reference sequence.

To quantitate and identify both novel as well as known variant sequences, an amplicon library was designed by the AVA. After generating a multiplexer matrix, an association between each patient sample and the patient-specific barcode combination (Multiplex Identifier, named MIDs) was determined.

### 3.11.2 PyroBayes

To analyze 454 sequencing files using non GS software tools, we converted the .SFF format to .FASTA and .FASTA.qual standards (Pearson and Lipman, 1988) using the PyroBayes base caller, developed in the Marth Lab in the Department of Biology at the Boston College (Chestnut Hill, USA)[5] (Quinlan et al., 2008). The input commands are listed below; the additional option [-o] enables renaming of original files; the number of reads can be specified using option `[-n X]`.

```
⊐ ./pyroBayes -i <in.sff> -n X -o <new name>
```

### 3.11.3 Oligomap and Grep

Oligomap[6] is an algorithm for identifying short DNA sequences in high-throughput data sets, avoiding complex computational algorithms (like TargetScan or BLAST) (Carninci et al., 2005; Berninger et al., 2008).

[5] http://bioinformatics.bc.edu/marthlab; available at March 5, 2010

[6] http://www.mirz.unibas.ch/software/oligomap/oligomap_101.tar; available at June 13, 2010

In this thesis, Oligomap formed part of the bioinformatic pipeline to find target-specific primer sequences within the sequencing data. The input protocol is listed below. The additional option [-d] allows scanning of all .fa files in one directory. By using command [-r], a final report was created with a maximum [-m] number of hits for one query.

To search and filter a defined pattern out of a specific file and to identify sequences of either primer or short target sequences, the UNIX based regular expression matcher Grep[7] (global/ regular expression/ print) was used; command lines for both functions are listed below.

```
⊐ ./oligomap target.fa query.fa [-d] [-r new_filename] [-m maxhits]

⊐ ./grep [options] PATTERN [file]
```

### 3.11.4 Multiple sequence alignment

Multiple sequence alignments were performed using (i) the MOSAIK software tool and (ii) the Burrows-Wheeler Aligner, and compared against the alignments generated using the 454 GS tool; see Section 3.11.1.

#### 3.11.4.1 MOSAIK

MOSAIK and its subprograms (MosaikBuild, MosaikAligner, MosaikSort, and MosaikAssembler) were used for downstream SNPs and short insertion/deletion (INDEL) discovery. The resulting .GIG file was loaded into the GigaBayes application (see Section 3.11.6), and evaluated through perlscripts written by Hovestadt. The MOSAIK source code is shown below; the additional arguments [-p] or [-mm] regulated the specified number of CPUs and allowed mismatches, respectively.

```
⊐ MosaikBuild -fr <in.db.fa> -oa <db.out>

⊐ MosaikBuild -fr <combined.fa> -fq <combined.qual> -out <combined.out> -st '454'

⊐ MosaikAligner -ia <db.out> -in <combined.out> -out <combined.align> -p 31 -rur
<combined.notalign> -mm 10

⊐ MosaikSort -in <combined.align> -out <combined.sort>

⊐ MosaikAssembler -ia <db.out> -in <combined.sort> -out <combined.gig> -f 'gig'
```

[7] http://unixhelp.ed.ac.uk/CGI/man-cgi?grep; available at March 5, 2010

#### 3.11.4.2 Burrows-Wheeler Aligner

The sequencing data used was in .FASTA and .FASTA.qual format. Additional options [-...] for BWA are listed at http://bio-bwa.sourceforge.net/bwa.shtml[8]. The final indexed alignment file .SAI was used to create a .SAM format files; source codes are described below.

```
❑ bwa aln [-n maxDiff] [-o maxGapO] [-e maxGapE] [-d nDelTail] [-i nIndelEnd] [-k
maxSeedDiff] [-l seedLen] [-t nThrds] [-cRN] [-M misMsc] [-O gapOsc] [-E gapEsc] [-q
trimQual] <in.db.FASAT> <in.query.fq> > <out.sai>
```

```
❑ bwa samse <in.db.FASTA> <in.sai> <in.query.fq> > <out.sam>
```

The Sequence Alignment/ Map (SAM) format is an alignment format for visualization and storage of sequence alignments or gene maps against reference sequences, the binary equivalent of SAM is called Binary Alignment/Map (BAM) (Li et al., 2009). SAM supports different sequencing platforms with diverse alignment formats, and reads up to 128 Mbp. A set of output formats are discretionary, including ISIZE (Interred Insert Size), MAPQ (Mapping Quality), and CIGAR (Compact Idiosyncratic Gapped Alignment Report).

### 3.11.5 SAMtools and Pileup

SAMtools-0.1.7 was used to convert the .SAM to .BAM format and predict genetic variations (Li et al., 2009). It contains many useful converters (blast2sam.pl, bowtie2sam.pl, ect.) and tools, which can be used for various applications. In this thesis the Pileup tool was used to desribe the base-pair information at each chromosomal position.

```
❑ samtools import <in.db.fa.fai> <in.sam> <out.bam>
❑ samtools sort <in.bam>
❑ samtools index <in.sorted.bam>
Pileup:
❑ samtools pileup -f <in.db.fa> <in.sorted.bam> -c > <out.pileup>
❑ perl samtools-0.1.7a/misc/samtools.pl varFilter -d 10 <in.pileup> > <out.pileup-10X>
```

[8] available at March 19, 2010

### 3.11.6 GigaBayes

To discover short-read SNPs and short-INDELs, we used the GigaBayes polymorphism discovery package, consisting of gigaBuild and gigaBayes, both programs were developed in the Marth Lab in the Department of Biology at the Boston College (Chestnut Hill, USA)[9] (Marth et al., 1999).

The gigaBuild program requires DNA sequence reads or a reference file in a .FASTA format, and builds a binary file (.GIG format), containing complete assembly information required by the polymorphism discovery program gigaBayes.

The gigaBayes program analyses of a large number of assembled high-throughput sequencing data sets against an entire reference genome (Hillier et al., 2008). Different input formats can be applied; in case of this thesis mainly the .GIG format, created by MOSAIK Assembler (see Section 3.11.4.1), was used; input commands are shown below. The additional arguments [–log] enables the creation of a log file for program execution, whereas the [–indel] switch enables single-base INDEL detection. The command [–ploidy 'X'] instructs the program that the sequences come from a haploid or diploid organism, and the [–O] argument controls how much detail is printed in the output file about the candidate polymorphism. Option [–CRL] sets the minimum number of aligned bases that must be present at a given genome position before it is considered for full Bayesian analysis[10].

```
❏ gigaBuild -fd <in.fasta> [or -ace <in.ace>] -gig <out.gig>
❏ gigaBayes -gig <in.gig> -gff <out.gff> -log <out.log> -indel -ploidy ’diploid’ -sampleDel
’|’ - O 3 -CRL 5
```

The main output was presented in the standard tab-separated .GFF format, providing the reference sequence name (contig), annotation type (INDEL or SNP), start/ end coordinates of the variation, a Bayesian algorithm score, and a polymorphism candidate descriptor to provide information about individual genotypes, polymorphic alleles, and read-specific base calls.

[9] http://bioinformatics.bc.edu/marthlab; available at March 5, 2010

[10] GigaBayes Polymorphism Discovery Software Documentation; last revised: March 2, 2009

# 4 Results

We designed a nested multiplexed PCR assay with barcoded pyrosequencing to assess sequence variations in multiple genomic regions in hundreds of individuals. Following fusion primer design and optimization of PCR conditions and workflow formats, we applied this method to assess sequence variations in the FMRP RBP target regions in the coding sequence and miRNA target sites in the 3' UTR of the NF1 gene. We subsequently modified the assay for increased sample throughput, analyzing 17 highly expressed miRNA genes in hundreds of patients with chronic lymphocytic leukemia, Type 2 Diabetes mellitus, and autism. To analyze sequence data generated through this assay, we developed a bioinformatic pipeline and used several methods for data management and biocomputational analyses.

## 4.1 Target-specific primer design

Most target-specific primer pairs generated products of expected size following nested multiplexed PCR. When the expected product was absent, we resolved the problem using alternative primer pairs or optimizing PCR cycling conditions using gradient PCR. Eighteen multiplexed and 24 nested primer pairs were designed for the NF1 project. Seventeen multiplexed primer pairs and 17 nested primer pairs were designed for the miRNA gene project. Primer sets for both projects are listed Appendix 1 and 2.

## 4.2 Barcoded fusion primer design

After establishing that products of expected size were generated using target-specific primers, barcoded (patient-specific) primers were designed to amplify each target for 100 patients in the NF1 project and 256 persons in the miRNA gene project. Each fusion primer had a modular structure, consisting of 454 A/B sequencing adapters, barcode sequences with a length of eight nucleotides, and target-specific primer sequences. Four hundred and eighty barcoded primers were designed for the NF1 project and 544 barcoded primers were designed for the miRNA gene project. A scheme for pipetting barcoded primers to generate patient-specific PCR products is shown in Figure 4.1.

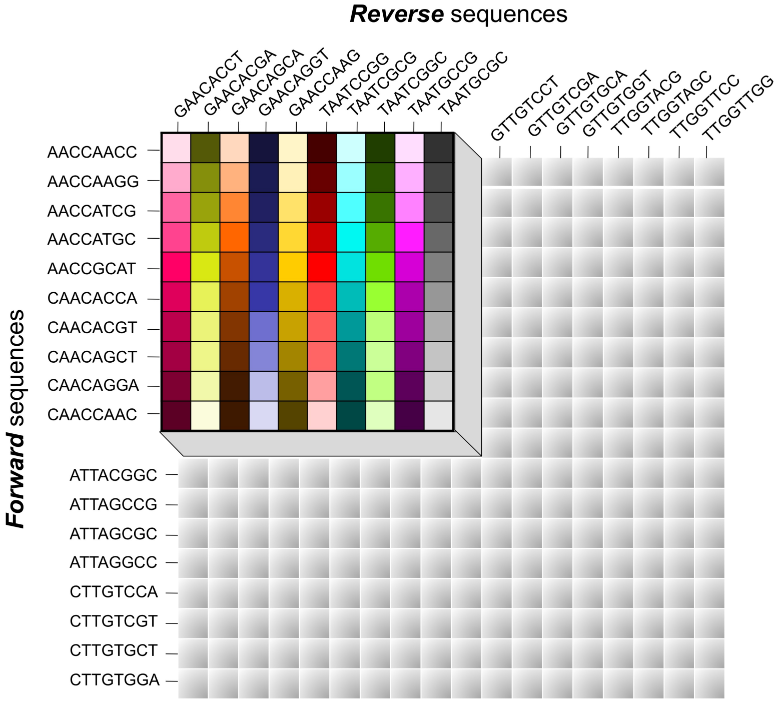


Figure 4.1: Scheme for pipetting barcoded primers to generate patient-specific PCR products in nested PCR. The elevated grid indicates combinations of 10 barcoded forward and 10 barcoded reverse primers used to generate 100 patient-specifc PCR products for each target in the NF1 project. Additional barcode sequences, corresponding to the non-elevated grid, were used in the miRNA gene project to increase the sample throughput up to 256 patient-specific PCR products.

## 4.3 Optimization of PCR reaction and cycling conditions

### 4.3.1 PCR optimization for the NF1 project

Visible amplification products for all regions were generated following 30 cycles of amplification in nested PCR (Figure 4.2 A). The yield of nested PCR products was not improved by increasing the number of PCR cycles beyond 35 in multiplexed PCR reactions (Figure 4.2 B). We checked annealing times in both multiplexed and nested PCR, comparing annealing times of 45 s and 1 min 15 s. Longer annealing times improved PCR amplification (Figure 4.2 C). Consequently we selected 35 PCR cycles and an annealing time of 1 min 15 s for our nested PCR reaction conditions. We also tested a range of diluted multiplexed products to assess their efficiency as template for nested PCR reactions, and established that a 1:20 dilution was sufficient (data not shown).

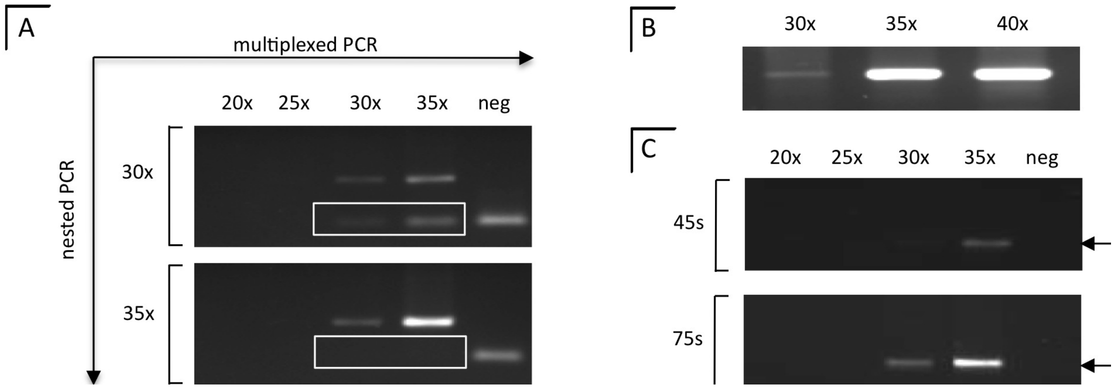


Figure 4.2: PCR optimization for multiplexed and nested steps. Ten microliter aliquots of a representative nested PCR product (114 bp) and negative control (neg) were visualized on 2 % agarose gels. Number of PCR cycles is indicated by x. Neg indicates the negative control following 35 PCR cycles. (A) When multiplexed PCR was performed with < 30 cycles, no visible products were identified following nested PCR. (B) No increase of band intensity was seen after > 35 cycles in nested PCR reaction. (C) PCR products were visible following 30 PCR cycles using an annealing time of 1 min 15s in the nested PCR step.

A range of primer concentrations was tested in multiplexed and nested PCR steps. Higher concentrations of primers in multiplexed PCR improved product formation after nested PCR (Figure 4.3). A total primer concentration of 720 nM (20 nM each) in multiplexed PCR was sufficient to generate PCR products for all targets in the NF1 project. Nested PCR reactions were performed using a final concentration of 200 nM primers (100 nM each). Increasing primer concentration to 1 $\mu$M in multiplexed PCR reactions did not improve the amount of nested PCR products.

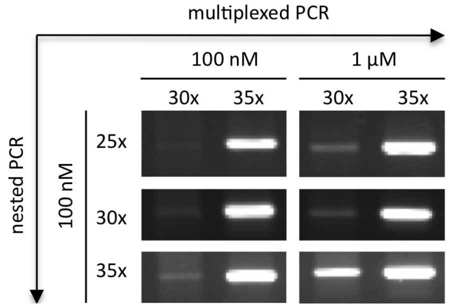


Figure 4.3: Establishing primer concentrations for multiplexed and nested PCR steps in the NF1 project. Ten microliter aliquots of nested PCR products of i2209 were visualized on 2 % agarose gels. Cycle number (x) and concentration for each primer is indicated above or beside.

### 4.3.2 PCR optimization for the miRNA gene project

To generate distinct nested PCR products for the miRNA gene project, it was necessary to increase the total primer concentration to 3400 nM (100 nM each) in multiplexed PCR and to increase the number of PCR amplification cycles to 38 cycles. We diluted the products 1:20 to reduce the concentration of multiplexed primers.

PCR conditions for nested PCR were identical to those in the NF1 project. Increasing total primer concentration of barcoded primers in excess of 200 nM did not result in more intense product bands or products appearing at earlier cycle numbers. No improvements were achieved either by lowering the degree of multiplexing from 18-plex to 2 x 9-plex reactions or following transfer of more template from multiplexed PCR to subsequent nested PCR (data not shown). The amplification of representative miRNA targets 7, 9, 15a and 16 before and after optimization are shown in Figure 4.4.

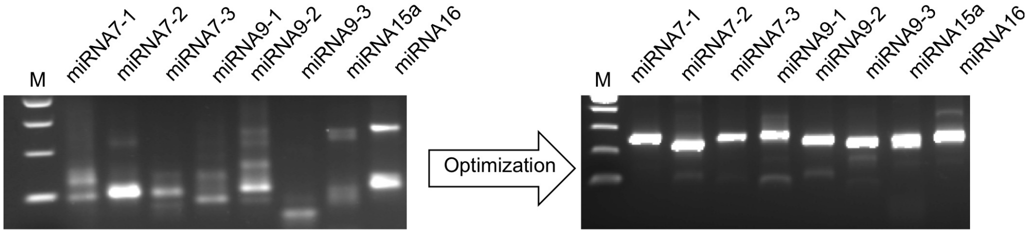


Figure 4.4: Optimization of nested multiplexed PCR for the miRNA gene project. Amplicons representing miRNA targtes 7, 9, 15a and 16 before (left) and after optimization (right).

## 4.4 Workflow and experimental enhancements for increased sample throughput

Multiplexed and nested PCR reactions in 96- and 384-well formats, were assembled using an EpMotion 5075® Automated Pipetting Station. For each project, 4 programs for pipetting, dilution, and pooling were designed and optimized using appropriate pipetting tools, equipment layout, and workflows. DNA samples of the NF1 project were aliquoted manually due to incompatibility between used pipetting tools and the DNA sample tubes. We controlled all steps of the workflow including mixing conditions; 3 mixing cycles of one half of total volume at 7.0 mm/s speed were optimal for subsequent PCR reactions.

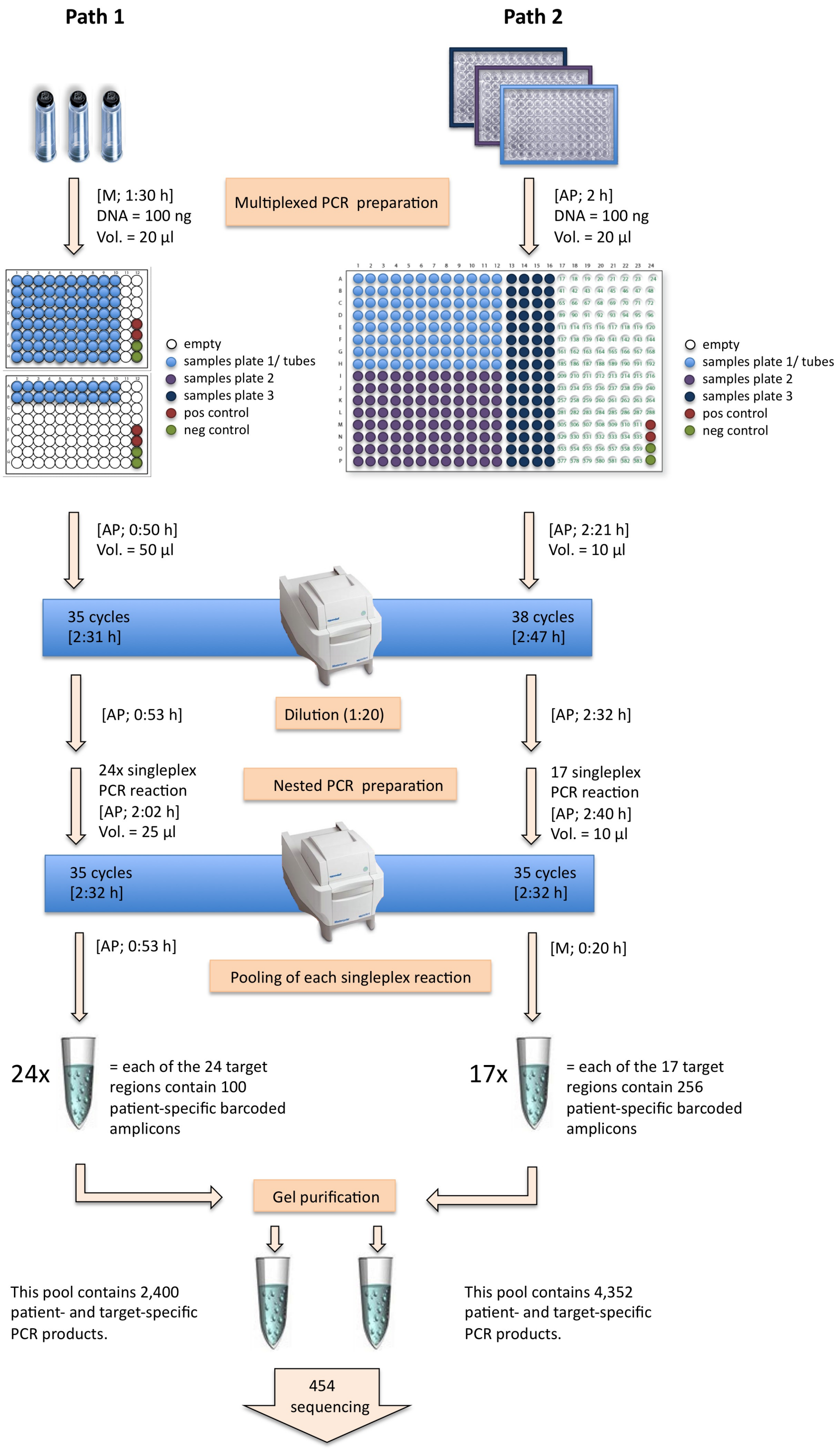


Figure 4.5: Workflow scheme for nested multiplexed PCR assay. Path 1 shows the worklfow for the NF1 project using a 96-well plate format. Path 2 shows the workflow for miRNA gene project using a 384-well plate format. Manual (M) and automated pipetting (AP) steps, followed by duration, are indicated at each step in square brackets.

## 4.5 NF1 project

We designed a nested multiplexed PCR assay to assess sequence variations in 15 FMRP target regions and 9 miRNA target sites (miR103/107, miR137, miR10a/b, miR128a/b, miR27a/b, miR490, miR 30a-5p/b/c/d/e-5p, and miR153) of the NF1 gene in 400 persons with and without autism. The locations of the target sites were identified through PAR-CLIP and Target Scan analyses (Ascano and Renwick, personal communication).

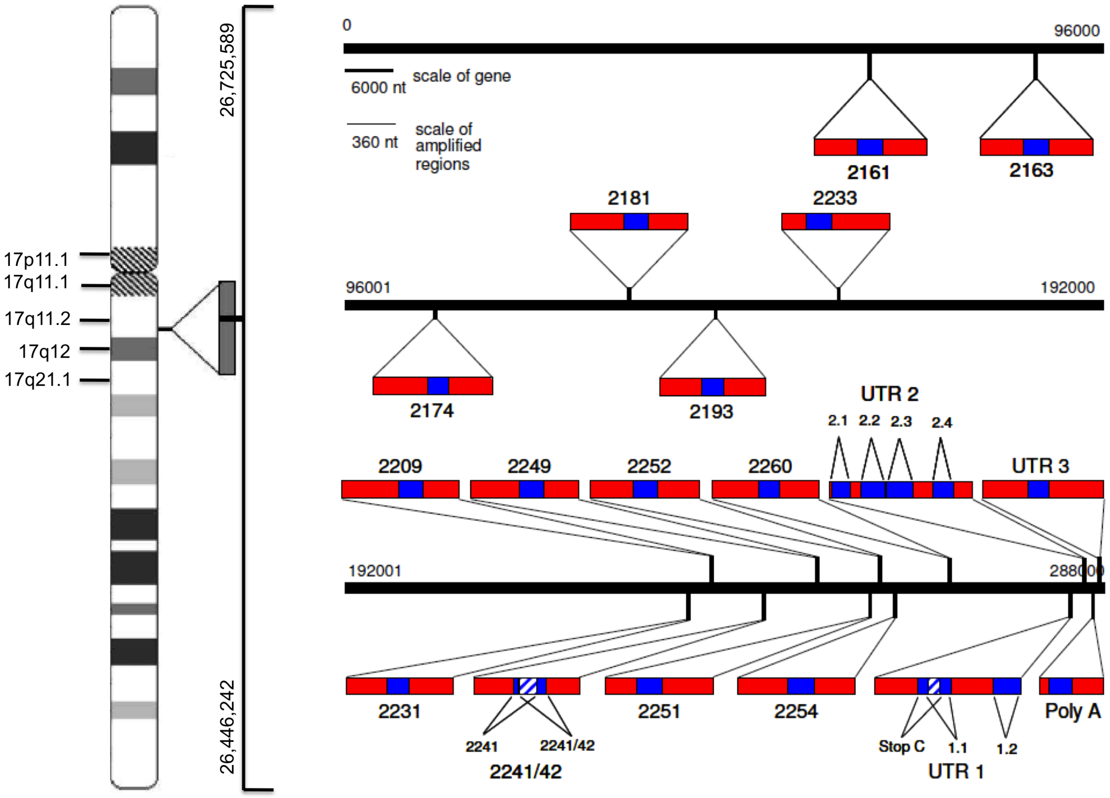


Figure 4.6: Map indicating regions for PCR amplification in the NF1 gene: 18 multiplexed PCR products (red) and 24 nested PCR products (blue) are depicted. Fifteen nested PCR products are located in the CDS, each containing 1 FMRP target site (except i2241/42 contains two FMRP target sites). Nine nested PCR regions are located in the 3'UTR, 1 region containing the stop codon and 8 regions containing the following miRNA target sites in brackets: iUTR1.1 (miR103/107), UTR1.2 (miR137), iUTR2.1 (miR10a/b), iUTR2.2 (miR128a/b and miR27a/b), iUTR2.3 (miR490), iUTR2.4 (miR30a-5p/b/c/d/e-5p), iPoly A (Poly-A-signal), and iUTR3.1 (miR153). Regions where PCR amplicons generate overlapping sequences are indicated by blue diagonal stripes.

### 4.5.1 Study population demographics

For the NF1 project, we studied a total of 400 persons, consisting 134 persons with autism and 266 without autism. One hundred and sixty seven (42 %) were female and 233 (68 %) were male. Most participants were born between 1935 and 1999. The average age of autism onset was 17 months, with a range of 1-3 years. The average age at assessment was 9.3 years, with a range from 0 to 27 years. For 124 persons (31 %), demographic data was either partially complete or not available (N/A). Demographic data are summarized in Table 4.1.

Table 4.1: Demographic characteristics of the study population. Data were unavailable for subsets of participants including no or inconclusive disease (†) and missing information about age (*). Average values are symbolized by ∅.

| | study participants (400) | | | |
|---|---|---|---|---|
| | cases (134) | | controls (266) | |
| Diagnosis | | | | |
| Autism | 134 | [33.5 %] | | |
| Without disorder | | | 160 | [40.0 %] |
| Unknown [†] | | | 106 | [26.5 %] |
| Gender | | | | |
| Female | 20 | [149 %] | 147 | [55.3 %] |
| Male | 114 | [85.1 %] | 119 | [44.7 %] |
| Age | | | | |
| < 11 | | | | |
| 11 - 20 | 68 | [50.7 %] | 10 | [3.8 %] |
| 21 - 30 | 56 | [41.8 %] | 6 | [2.3 %] |
| 31 - 40 | 9 | [6.7 %] | 4 | [1.5 %] |
| 41 - 50 | 1 | [0.8 %] | 36 | [13.5 %] |
| > 50 | | | 117 | [44.0 %] |
| N/A [*] | | | 93 | [35.0 %] |
| at assessment | ∅ 7.2 | | ∅ 9.3 | |
| Twin status | | | | |
| monozygotic | 1 | [0.7 %] | | |
| dizygotic | 9 | [6.7 %] | | |

All participants belonged to a family; 149 families were studied in total. Approximately two thirds of the families had more than three members; 184 (46 %) had three members, 36 (9 %) had four people, 20 (5 %) had five members, 8 (2 %) had six members, and 4 (1 %) had seven persons. The remaining 148 (37 %) participants have been in 'families' with only two members (84; 21 %) or alone (64; 16 %).

Nested multiplexed PCR was performed on DNA samples from these participants in 4 groups of one hundred persons; an overview of the study population in each group is listed in Table 4.2.

Table 4.2: Diagnostic status of the study population. The number of participants with and with unknown status is indicated for each patient group.

| | group 1 | group 2 | group 3 | group 4 |
|---|---|---|---|---|
| Autism | 96 | 4 | 31 | 0 |
| Without disorder | 0 | 3 | 58 | 99 |
| Unknown | 4 | 93 | 11 | 1 |

### 4.5.2 Performance of nested multiplexed PCR assay

During multiplexed PCR, 18 target regions in the NF1 gene were amplified. These products were diluted 20-fold to reduce multiplexed primer concentrations and served as templates for subsequent nested PCR. In total, 9,600 PCR products were generated, representing 24 target regions from 400 patients. To check successful nested PCR amplification, representative PCR reactions for each target were run on 2 % agarose gels (Figure 4.7). Nested PCR reactions were pooled by target region and run on a 3 % Nusieve low-melting point agarose gel. PCR products of expected size were cut out and gel purified.

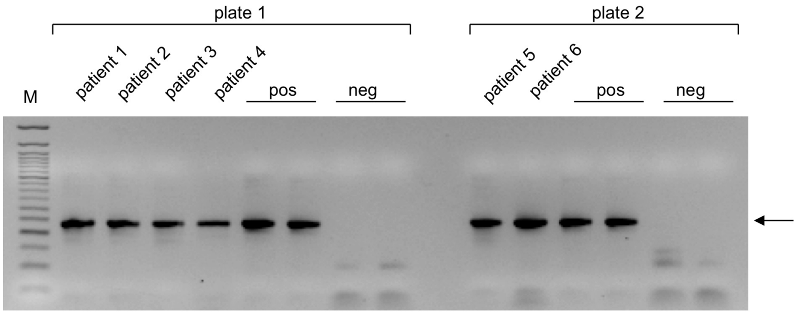


Figure 4.7: Assessment of PCR reactions. For each nested PCR, 5 $\mu$l aliquots of six randomly chosen patient-specific amplicons were assessed by agarose gel electrophoresis. Positive (pos) and negative (neg) controls are indicated. Amplicon sizes were assessed using a 50 bp DNA ladder (M). A representative region, i2209 of patient group 2, is shown.

### 4.5.3 Gel purification of pooled PCR products

All purified PCR pools were run on an 1.5 % agorose gel and compared to the pools prior purification (Figure 4.8). The concentrations of PCR pools were estimated using a self-made marker. Most pooled products for each target and each patient group, were clearly visible following agarose gel electrophoresis. The pooled UTR3-1 target region in patient group 2 was not visualized (arrow), however the reaction was successfully repeated (data not shown).

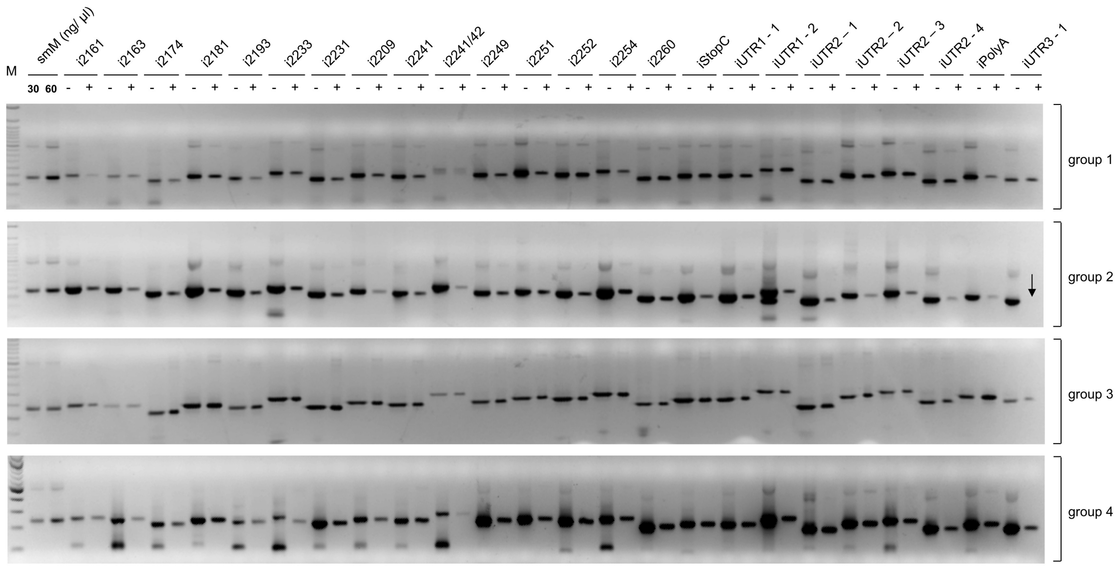


Figure 4.8: Gel purification of pooled PCR products. Five microliter aliquots of pooled nested PCR products for all 24 target regions before and after gel purification for all 400 patients (4 groups of 100 persons) were visualized on a 2 % agarose gel. To assess the size of each band a 25 bp ladder (M) was used. Quantitative size markers (smM) indicate 30 ng and 60 ng dsDNA. Some purified products were less visible than others following purification (tagged by arrow).

### 4.5.4 Quantitation of pooled and purified PCR products

The concentrations of purified pooled products was estimated visually using a self-made marker representing 10-60 ng dsDNA. Band intensities were also compared against a commercially prepared 100 bp DNA ladder. For each group of 100 patients, pooled and purified products were visualized and quantitated using these size markers; a representative example of the gel purification step is shown in Figure 4.9.

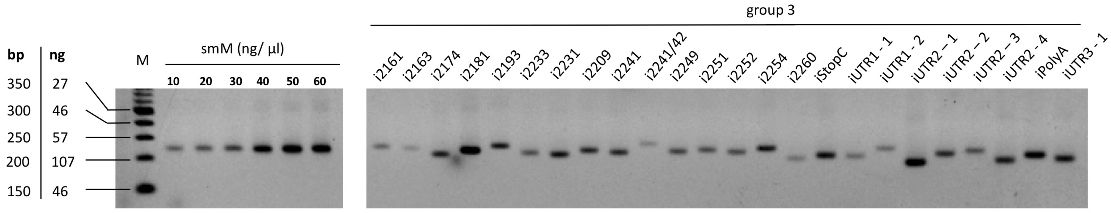


Figure 4.9: Quantitation of pooled and purified PCR products. Pooled and purified products for each target region were quantitated visually using a self-made marker (smM) and a commercial 100 bp DNA ladder.

Concentrations of pooled and purified PCR products for group 1 averaged 8.9 ng/$\mu$l and ranged from < 2 to 10 ng/$\mu$l. Concentrations of pooled and purified PCR products for group 2 averaged 7.3 ng/$\mu$l and ranged from 4 to 12 ng/$\mu$l. Concentrations of pooled and purified PCR products for group 3 averaged 5.3 ng$\mu$l and ranged from 2 to 12 ng/$\mu$l. Concentrations of pooled and purified PCR products for group 4 averaged 9.5 ng$\mu$l and ranged from 2 to 12 ng/$\mu$l (see Table 4.3).

Table 4.3: Concentrations of purified pooled products for all four patient groups.

| | **group 1 [ng/$\mu$l]** | **group 2 [ng/$\mu$l]** | **group 3 [ng/$\mu$l]** | **group 4 [ng/$\mu$l]** |
|---|---|---|---|---|
| i2162 | 10 | 12 | 2 | 4 |
| i2163 | < 2 | 6 | 2 | 4 |
| i2174 | < 2 | 6 | 6 | 6 |
| i2181 | 10 | 12 | 12 | 12 |
| i2193 | 10 | 4 | 6 | 2 |
| i2209 | 10 | 10 | 4 | 2 |
| i2231 | 10 | 10 | 6 | 12 |
| i2233 | 10 | 4 | 6 | 6 |
| i2241 | 10 | 4 | 6 | 12 |
| i2241/42 | 2 | 4 | 2 | 2 |
| i2249 | 10 | 6 | 4 | 12 |
| i2251 | 10 | 8 | 4 | 12 |
| i2252 | 10 | 12 | 4 | 12 |
| i2254 | 10 | 12 | 6 | 12 |
| i2260 | 10 | 10 | 2 | 12 |
| iStopC | 10 | 10 | 6 | 12 |
| iUTR1.1 | 10 | 6 | 2 | 12 |
| iUTR1.2 | 10 | 6 | 2 | 12 |
| iUTR2.1 | 10 | 6 | 12 | 12 |
| iUTR2.2 | 10 | 4 | 8 | 12 |
| iUTR2.3 | 10 | 4 | 4 | 12 |
| iUTR2.4 | 10 | 4 | 6 | 10 |
| iPolyA | 10 | 8 | 8 | 12 |
| iUTR3.1 | 10 | 8 | 6 | 12 |

To ensure optimal sequencing, equimolar concentrations of each target region were pooled for each group of 100 participants. Fifty microliters of each ”masterpool”, at a concentration of approximately 60 ng per $\mu$l, were submitted for 454 sequencing (MSKCC, New York).

## 4.6 miRNA gene project

In this project, we studied sequence variations in 17 miRNA genes, representing highly expressed, multicopy and/ or central nervous system (CNS). To incease sample throughput, we scaled up our assay to a 384-well format by increasing the number of barcoded primers in the nested PCR, and modifiying reaction and cycle conditions, and workflows (Section 4.4).

Through these modifications, 768 patient samples, processed in three 384-well plates (2/4-RU1; 6/7-RU1/2; and AUTp/n-Pa-RU2) were analyzed. A total of 13,056 PCR products, representing 17 targets regions for 768 patients, was generated. Each pooled target region for each patient group, was assessed by agarose gel electrophoresis and pooled as described for the NF1 project (see Figure 4.10).

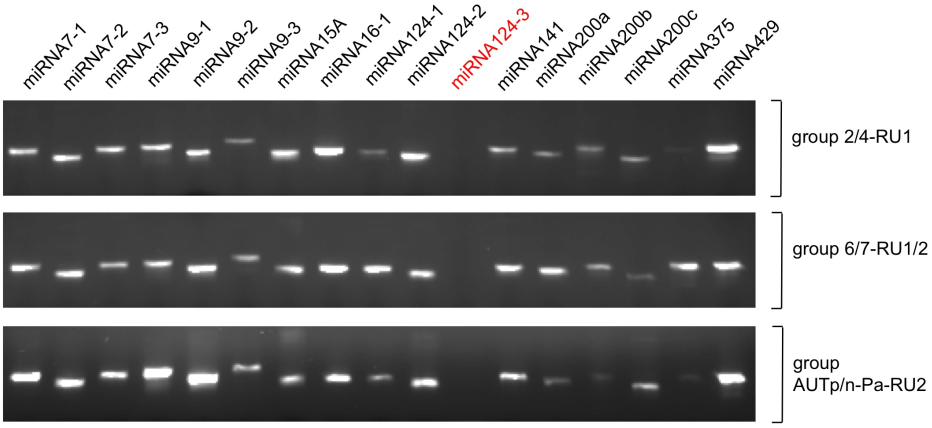


Figure 4.10: Quantitation of pooled and purified PCR products in the miRNA gene project. Pooled and purified products for each target region were quantitated visually using a self-made marker (smM) and a commercial 100 bp DNA ladder.

The concentrations of the isolated and purified PCR products were estimated using a quantitative size marker. Concentrations for all three groups ranged from 2 to 12 ng/$\mu$l; the majority (37 of 48) had a concentration above 6 ng/$\mu$l. Fifty microliters of each ”masterpool” at a final concentration of approx. 60 ng/$\mu$l, were submitted for 454 sequencing (MSKCC, New York).

## 4.7 Bioinformatic analyses and biocomputational approaches

Pooled PCR products from 4 groups of 100 persons and 3 groups of 256 persons were respectively sequenced for the NF1 and miRNA gene project using 454 sequencing. For each group in the NF1 project, sequence read files in .SFF format were generated; sequencing data for the miRNA project were not available due to the long turnaround time for next generation sequencing services.

We developed a bioinformatic approach to identify target-specific primers, targets, and barcode sequences in the sequencing data. We used these data sets to evaluate the performance and accuracy of our assay. Individuals were identified by patient-specifc barcode combinations; analyses were performed separately for each group due to the use of identical barcodes.

### 4.7.1 Sequence analysis using GS Reference Mapper

All .SFF read files were assessed using the GS application "Reference Mapper"; the *Homo sapiens* neurofibromin 1 gene (NG_009018.1; GI:213385299; 01-NOV-2009) was used as reference sequence. A total of 1,366,498 sequence reads was generated for the NF1 project; including 286,880 reads for group 1; 369,212 reads for group 2; 477,253 reads for group 3; and 251,153 sequencing reads for group 4. An total of 1,279,042 (93.6 %) of the reads were mapped against the reference sequence with an inferred read error rate of 0.04 to 0.12 percent, including 265,421 (92.52 %) mapped reads in group 1; 359,723 (97,43 %) mapped reads in group 2; 427,237 (89.52 %) mapped reads for group 3; and 238,495 (94.96 %) mapped reads in group 4. An overview of the Reference Mapper output is presented in Table 4.4.

Table 4.4: Sequence read mapping for the NF1 project using the Reference Mapper (GS, Roche).

| | **group 1** | **group 2** | **group 3** | **group 4** |
|---|---|---|---|---|
| reads | 286,880 | 369,212 | 477,253 | 251,153 |
| bases | 54.8 mio | 70.5 mio | 85.2 mio | 49.6 mio |
| mapped reads | 92.5 % | 97.4 % | 89.5 % | 94.9 % |
| total of mapped reads | 265,421 | 359,723 | 427,237 | 238,495 |
| inferred read errors | 0.04 % | 0.05 % | 0.05 % | 0.12 % |

All 24 target regions were found in each group; target regions i2241 and i2241/42, iStopC and iUTR1.1, and iUTR2.2 and iUTR2.3 were combined as contigs due to their overlapping target sequences (Figure 4.6). An average of 15,227 sequence reads for each of the 21 contigs (18 unique and 3 combined) for all 4 patient groups was generated .

### 4.7.2 Barcode detection, read trimming, and length distribution of sequence reads

Sequence flowgram files (.SFF) were transformed into .FASTA and .FASTA.qual formats using PyroBayes, and restructured using the perl scripts 1_filter.pl and 2_filter_qual.pl. Barcode sequence identification and read trimming were performed using Oligomap; primer sequences, summarized in an additional .FASTA file, served as the query for sequence searches. Barcode detection was performed using 0 mismatches for both strands of the target. A total of 271,413 (19.8 %) of 1,366,498 sequence reads were excluded due to the absence of primer sequences: 46,235 (16.1 % ) were excluded from group 1; 56,537 (15.3 %) were excluded from group 2; 97,979 (20.5 %) were excluded from group 3; and 70,662 (28.1 %) were excluded from group 4.

Filtered data were consolidated using the 4_alignments.pl and 4b_alignments.pl perl script, generating sequence statistics. The quality of trimmed data was assessed using a frequency histogram of sequence length for all four patient groups. Two populations of sequence reads were identified using an arbitrary cut-off value set at 120 nt in length; the majority (approx. 93 %) of sequence reads were longer and the minority (about 7 %) were shorter than this cut-off value (Figure 4.11). The expected sizes of our NF1 targets without fusion 454-adapter sequences ranged between 161 nt (for iUTR2.1) and 234 nt (for i2241/42), corresponding to the size range of the majority population.

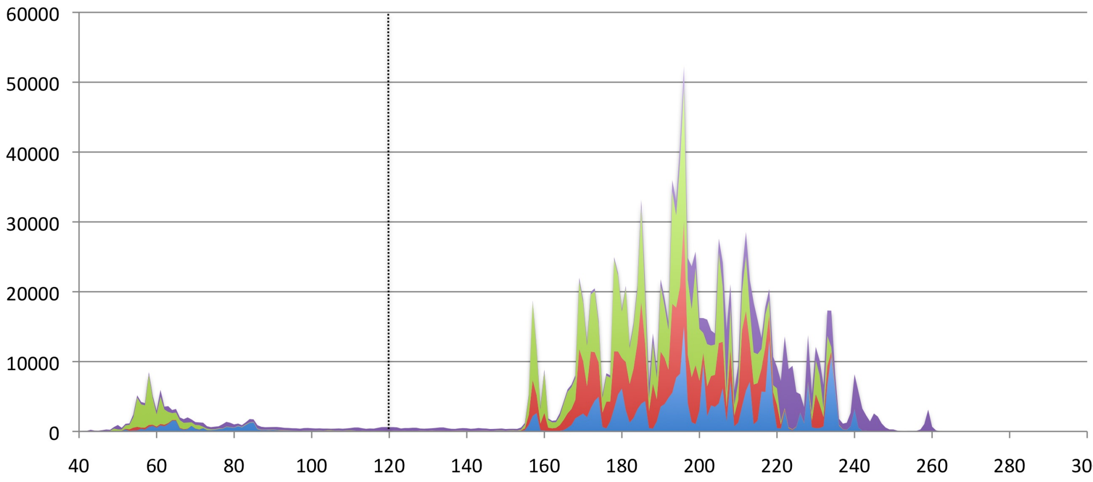


Figure 4.11: Frequency histogram of filtered sequence reads for each patient group. Depicted are the absolute number of sequence reads (y-axis) and read length (x-axis) for group 1 (blue), group 2 (red), group 3 (green), and group 4 (purple).

### 4.7.3 Proportions of target-specific primers in filtered sequence set

Sequence reads with target-specific primer sequences at both ends were identified by the presence of complementary and reverse complementary primer sequences. All 24 target regions were identified for each group in similar proportions (approximately 3-5 %). To assess the performance of our nested multiplexed PCR assay, all trimmed group files were concatenated; the proportions of each target-specifc primer set is summarized in Figure 4.12.

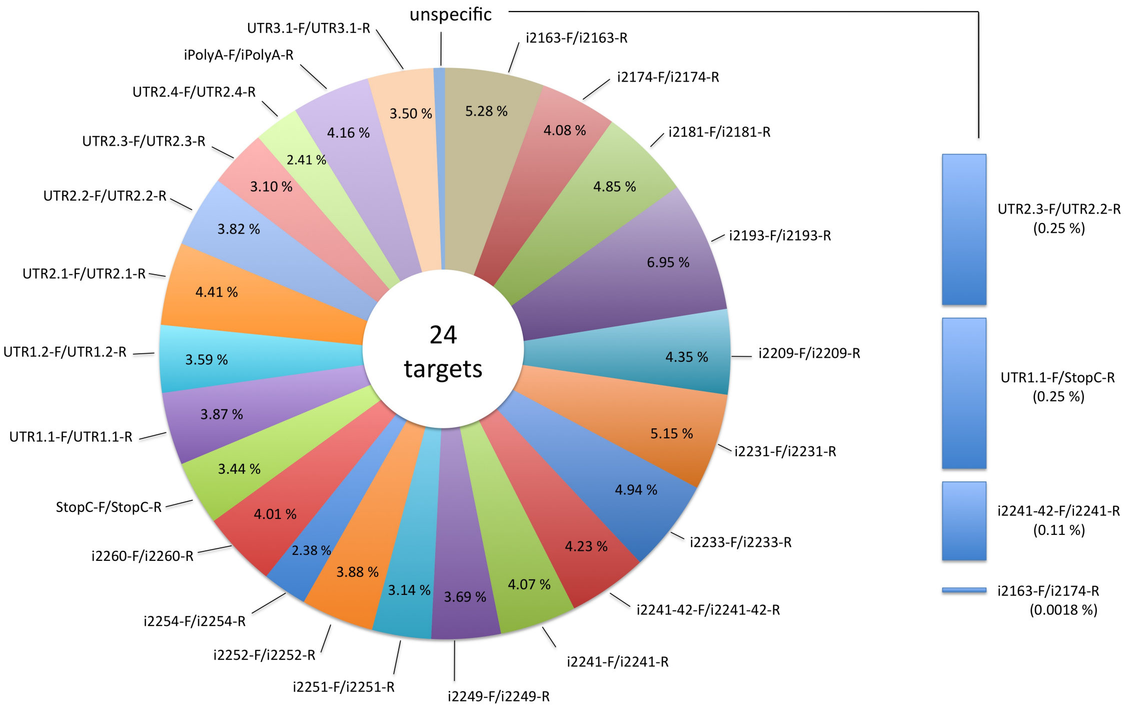


Figure 4.12: Proportions of target-specific primers in the filtered sequence read set. Relative proportions of sequence reads containing target-specific (left) and unspecific (right) primer combinations for all groups in the NF1 project

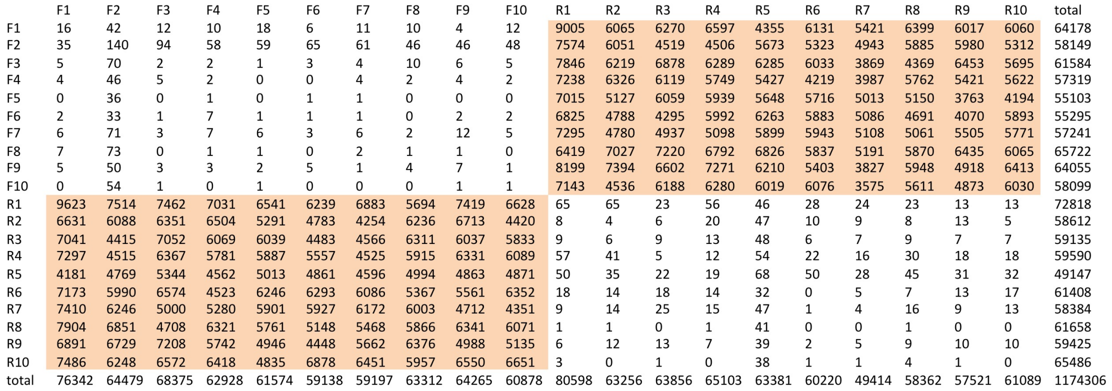
Unspecific sequence reads were found for the following products: i2241 and i2241/42 (0.11 %); iStopC and iUTR1.1 (0.25 %); iUTR2.2 and iUTR2.3 (0.25 %); i2163 and i2174 (0.0018 %). These products were generated when the amplicons of two different nested PCR reactions overlapped (see Figure 4.6).

### 4.7.4 Detection of barcode sequences and combinations

One hundred (patient-specific) barcode combinations were used per patient group. We identified 1,174,306 barcoded sequence reads, consisting of 1,171,143 (99.7 %) with expected, and 3,163 (0.3 %) with unexpected barcode combinations (Figure 4.13). Sequences with expected barcode combinations were present in similar amounts.

Approximately 11,700 sequence reads per barcode combination were expected for all 4 groups, based on the total number of expected sequence reads (1,171,143) divided by the number of barcode combinations (100). The highest number of barcoded sequence reads (18,628) was seen with the barcode combination F1/R1, whereas the lowest number number (7,926) was seen for the barcode combination F10/R7.

The minority of barcoded sequence reads (0.3 %) contained combinations of two forward or two reverse barcode sequences. Barcoded primers F1, F2, R1, R4 and R5 showed the highest number of unspecific combinations, particularly with themselves (Figure 4.13).



| | F1 | F2 | F3 | F4 | F5 | F6 | F7 | F8 | F9 | F10 | R1 | R2 | R3 | R4 | R5 | R6 | R7 | R8 | R9 | R10 | total |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| F1 | 16 | 42 | 12 | 10 | 18 | 6 | 11 | 10 | 4 | 12 | 9005 | 6065 | 6270 | 6597 | 4355 | 6131 | 5421 | 6399 | 6017 | 6060 | 64178 |
| F2 | 35 | 140 | 94 | 58 | 59 | 65 | 61 | 46 | 46 | 48 | 7574 | 6051 | 4519 | 4506 | 5673 | 5323 | 4943 | 5885 | 5980 | 5312 | 58149 |
| F3 | 5 | 70 | 2 | 2 | 1 | 3 | 4 | 10 | 6 | 5 | 7846 | 6219 | 6878 | 6289 | 6285 | 6033 | 3869 | 4369 | 6453 | 5695 | 61584 |
| F4 | 4 | 46 | 5 | 2 | 0 | 0 | 4 | 2 | 4 | 2 | 7238 | 6326 | 6119 | 5749 | 5427 | 4219 | 3987 | 5762 | 5421 | 5622 | 57319 |
| F5 | 0 | 36 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 7015 | 5127 | 6059 | 5939 | 5648 | 5716 | 5013 | 5150 | 3763 | 4194 | 55103 |
| F6 | 2 | 33 | 1 | 7 | 1 | 1 | 1 | 0 | 2 | 2 | 6825 | 4788 | 4295 | 5992 | 6263 | 5883 | 5086 | 4691 | 4070 | 5893 | 55295 |
| F7 | 6 | 71 | 3 | 7 | 6 | 3 | 6 | 2 | 12 | 5 | 7295 | 4780 | 4937 | 5098 | 5899 | 5943 | 5108 | 5061 | 5505 | 5771 | 57241 |
| F8 | 7 | 73 | 0 | 1 | 1 | 0 | 2 | 1 | 1 | 0 | 6419 | 7027 | 7220 | 6792 | 6826 | 5837 | 5191 | 5870 | 6435 | 6065 | 65722 |
| F9 | 5 | 50 | 3 | 3 | 2 | 5 | 1 | 4 | 7 | 1 | 8199 | 7394 | 6602 | 7271 | 6210 | 5403 | 3827 | 5948 | 4918 | 6413 | 64055 |
| F10 | 0 | 54 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 1 | 7143 | 4536 | 6188 | 6280 | 6019 | 6076 | 3575 | 5611 | 4873 | 6030 | 58099 |
| R1 | 9623 | 7514 | 7462 | 7031 | 6541 | 6239 | 6883 | 5694 | 7419 | 6628 | 65 | 65 | 23 | 56 | 46 | 28 | 24 | 23 | 13 | 13 | 72818 |
| R2 | 6631 | 6088 | 6351 | 6504 | 5291 | 4783 | 4254 | 6236 | 6713 | 4420 | 8 | 4 | 6 | 20 | 47 | 10 | 9 | 8 | 13 | 5 | 58612 |
| R3 | 7041 | 4415 | 7052 | 6069 | 6039 | 4483 | 4566 | 6311 | 6037 | 5833 | 9 | 6 | 9 | 13 | 48 | 6 | 7 | 9 | 7 | 7 | 59135 |
| R4 | 7297 | 4515 | 6367 | 5781 | 5887 | 5557 | 4525 | 5915 | 6331 | 6089 | 57 | 41 | 5 | 12 | 54 | 22 | 16 | 30 | 18 | 18 | 59590 |
| R5 | 4181 | 4769 | 5344 | 4562 | 5013 | 4861 | 4596 | 4994 | 4863 | 4871 | 50 | 35 | 22 | 19 | 68 | 50 | 28 | 45 | 31 | 32 | 49147 |
| R6 | 7173 | 5990 | 6574 | 4523 | 6246 | 6293 | 6086 | 5367 | 5561 | 6352 | 18 | 14 | 18 | 14 | 32 | 0 | 5 | 7 | 13 | 17 | 61408 |
| R7 | 7410 | 6246 | 5000 | 5280 | 5901 | 5927 | 6172 | 6003 | 4712 | 4351 | 9 | 14 | 25 | 15 | 47 | 1 | 4 | 16 | 9 | 13 | 58384 |
| R8 | 7904 | 6851 | 4708 | 6321 | 5761 | 5148 | 5468 | 5866 | 6341 | 6071 | 1 | 1 | 0 | 1 | 41 | 0 | 0 | 1 | 0 | 0 | 61658 |
| R9 | 6891 | 6729 | 7208 | 5742 | 4946 | 4448 | 5662 | 6376 | 4988 | 5135 | 6 | 12 | 13 | 7 | 39 | 2 | 5 | 9 | 10 | 10 | 59425 |
| R10 | 7486 | 6248 | 6572 | 6418 | 4835 | 6878 | 6451 | 5957 | 6550 | 6651 | 3 | 0 | 1 | 0 | 38 | 1 | 1 | 4 | 1 | 0 | 65486 |
| total | 76342 | 64479 | 68375 | 62928 | 61574 | 59138 | 59197 | 63312 | 64265 | 60878 | 80598 | 63256 | 63856 | 65103 | 63381 | 60220 | 49414 | 58362 | 57521 | 61089 | 1174306 |

Figure 4.13: Absolute number of sequence reads per barcode combination for all four groups. Expected barcode combinations are highlighted.

Sixty-six barcode combinations were over-represented, and 34 barcoded combinations were under-represented; based on the expected number of reads per barcode combination (approximately 11,100). The absolute number of observed sequence reads per barcode combination is shown in Figure 4.14.

Single read files for each specific barcode combination were generated using the 4b_alignments.pl script; the resulting files were combined and used for further alignment steps through the BWA and MOSAIK algorithms.

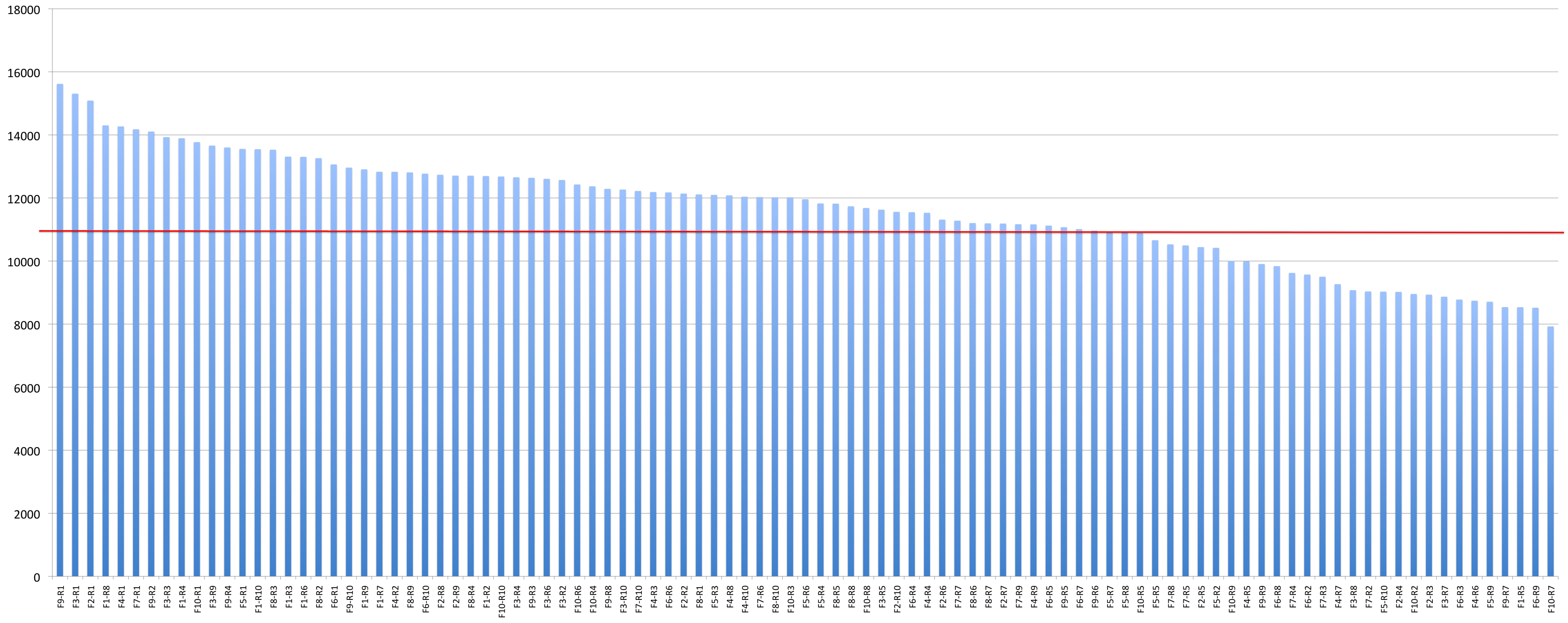


Figure 4.14: Absolute number of sequence reads (y-axis) per barcode combination (x-axis). The expected value per barcode combination is illustrated by the red line. Unspecific combinations (e.g. F-F or R-R) are not shown.

]

### 4.7.5 Alignment of 454 sequencing data

We applied and compared our sequencing data using three different sequence alignment algorithms; (i) the MOSAIK pipeline (Version 1.0.1388), (ii) the Burrows-Wheeler Aligner (BWA), and (iii) the GS Roche Genome Sequencer. In each case we used the oligomap-trimmed group sets as query files, and used an identical reference sequence file (NG_009018.1).

#### 4.7.5.1 MOSAIK alignment

Reference-guided assemblies with gapped alignments were generated using the MOSAIK pipeline, including the new Smith-Waterman algorithm. Trimmed read files for each group, in .FASTA and .QUAL formats, and the reference file, respectively, were parsed, compressed, and rewritten using the MosaikBuild application in '454' mode.

Sequence alignment was performed by the MosaikAligner using all 31 processors of the RNA-world server in the Tuschl laboratory. To find unique reads, we aligned all positions with a maximum mismatch threshold of ten, using a homo-polymer gap open penalty of four. Following exclusion of 8,092 (0.7 %) sequence reads, we aligned 240,297 sequence reads from group 1; 309,941 sequence reads from group 2; 374,779 sequence reads from group 3; and 179,976 sequence reads from group 4. A total of 39 (0.0035 %) non-unique reads was exported to a separate file. Alignment statistics for MOSAIK are listed in Table 4.5.

Table 4.5: Alignment statistics for trimmed reads after MosaikAlignment.

| | **group 1** | **group 2** | **group 3** | **group 4** |
|---|---|---|---|---|
| total reads | 240,645 | 312,675 | 379,274 | 180,491 |
| filtered out | 348 (0.1 %) | 2734 (0.9 %) | 4495 (1.2 %) | 515 (0.3 %) |
| total aligned | 240,297 (99.9 %) | 309,941 (99.1 %) | 374,779 (98.8 %) | 179,976 ( 99.7 %) |
| unique | 240,295 ( 99.9 %) | 309,930 (99.1 %) | 374,754 (98.8 %) | 179,975 (99.7 %) |
| non-unique | 2 (0.0 %) | 11 (0.0 %) | 25 (0.0 %) | 1 (0.0 %) |

The sorting of the sequence reads was done by MosaikSort, approximately 0.006 % of the reads were excluded. The processing of the reference sequence, including inserting of gaps, creating ungapped to gapped conversion tables, writing of assembly header, and appending read index to read data, was done by the MosaikAssembler.

#### 4.7.5.2 BWA alignment

BWA, based on Burrows-Wheeler Transformation, was used to evaluate the MOSAIK alignment. We transformed our reference file (NG_009018.1) using the BWA indexer, and aligned each group with an error rate of less than 3 %. The resulting alignment files were subsequently transformed into .SAM format, using the Samse application of the BWA package, and integrated in the SAM-tools pipeline for further SNP and INDEL analyses.

### 4.7.6 Preliminary SNP and short-INDEL analyses

We used (i) GigaBayes, (ii) SAMtools, and (iii) the GS Amplicon Analyzer for single basepair polymorphism (SNP and short-INDEL) discovery.

The GigaBayes application is part of the MOSAIK pipeline and uses previously aligned and assembled MOSAIK output files (in .GIG format). We used gigaBuild to build the required binary assembly reference (NG_009018.1) file (.GFF). From GigaBayes, imported files were rewritten and transposed using the perlscripts 8_parsemosaik.pl and 9_transpose.pl, respectively. We excluded 218 (46.8 %) of 466 total sequence variations by discarding all variation occurring at a frequency less than 10 %.

Using SAMtools, all BWA output files were transformed into binary .BAM format. We used Pileup to perform SNP/INDEL analysis. Output files were transformed into .VCF format using the sam2vcf.pl perlscript. A total of 6,650 variations were identified; from which 6,463 (97.2 %) with a quality score[1] of less the 1,000 were excluded.

We used the GS Amplicon Analyzer to design an amplicon library. After generating a multiplexer matrix, an association between each patient sample and the patient-specific barcode combination was specified. We excluded 1,190 (89.7 %) of 1,327 total variations by discarding all variation occurring at a frequency less than 10 %, and compared the results with those from MOSAIK/GigaBayes and BWA/SAMtools.

A total of 8,443 sequence variations was detected by GigaBayes, SAMtools, and GS Amplicon Analyzer; 572 (6.8 %) remained after filtering. The remaining variations were combined in a mastertable; 289 duplicate variations were identified and excluded. The remainig 281 (49.1 %) variations were unique; 122 (43.4 %) were identified only by the MOSAIK/GigaBayes pipeline; 91 (32.4 %) were identified only by BWA/SAMtools, and 44 (15.7 %) were detected only by GS Amplicon Analyzer.

[1] High quality scores indicate high confidence calls; based on an assessment logarithm; http://1000genomes.org/wiki/doku.php?id=1000_genomes:analysis:vcfv3.2; available at June 23, 2010

Twenty-four (8.5 %) sequence variations were identified by two or more bioinformatic pipelines; 16 (5.7 %) of 24 were identified by MOSAIK/GigaBayes and the GS Amplicon Analyzer, and 7 (2.5 %) of 24 were deteced by MOSAIK/GigaBayes and BWA/SAMtools. Only 1 (0.36 %) of 24 sequence variations was identified through each of the three pipelines.

### 4.7.7 Distribution of sequence variations in patient groups and relation to FMRP and miRNA target sites

For sequence analyses we excluded 91 (32.4 %) of 281 sequence variations because patient-specific barcodes were missing from the BWA/ SAMtools pipepline. We sorted 190 remaining variations by patient group and located their position relative to FMRP and miRNA binding sites in all 24 target regions.

Sequence variations were found in all 4 groups; 36 variations were identified in group 1; 82 variations were found in group 2; 126 variations were identified in group 3; and 100 variations were identified in group 4 (see Table 4.7).

Table 4.6: Type and frequency of sequence variations in patient groups. The number of single nucleotide polymorphisms (SNP), deletions (DEL), and insertions (INS) are indicated for each of the four patient groups.

| | **SNP** | **INS** | **DEL** | **Total in group** |
|---|---|---|---|---|
| group 1 | 12 | 1 | 23 | 36 |
| group 2 | 34 | 28 | 20 | 82 |
| group 3 | 56 | 44 | 26 | 126 |
| group 4 | 33 | 41 | 26 | 100 |

One hundred three (54.2 %) of 190 sequence variations were detected in only one patient group; 31(16.3 %) variation were found in two patient groups; 45 (23.7 %) in three groups, and 11 (5.8 %) in all patient groups.

We matched each variation with its specific target region. We concatenated iStopC and iUTR1.1, as well as, i2241 and i2241/42 into single contigs, because of their overlapping character. Each target region contained a minimum of one variation. Most sequence variations (37 of 190) were found in region i2174, followed by region i2181 with 27 sequence variations, and region i2193 with 18 sequence variations. Only one match was found in i2252, iUTR1.2 and iUTR2.1. An overview of the distribution of sequence variations in each target region is given in Table 4.7.

Table 4.7: Distribution of sequence variations in each target site. Shown are the numbers of single nucleotide polymorphisms (SNP), deletions (DEL), and insertions (INS) for each target site.

| | **SNP** | **INS** | **DEL** | **Total** |
|---|---|---|---|---|
| i2162 | 3 | 4 | 2 | 9 |
| i2163 | 0 | 7 | 0 | 7 |
| i2174 | 10 | 10 | 17 | 37 |
| i2181 | 18 | 7 | 2 | 27 |
| i2193 | 11 | 4 | 3 | 18 |
| i2209 | 1 | 2 | 1 | 4 |
| i2231 | 2 | 7 | 1 | 10 |
| i2233 | 4 | 6 | 1 | 11 |
| i2241/42 | 3 | 1 | 0 | 4 |
| i2249 | 2 | 4 | 2 | 8 |
| i2251 | 0 | 4 | 1 | 5 |
| i2252 | 0 | 1 | 0 | 1 |
| i2254 | 4 | 1 | 4 | 9 |
| i2260 | 1 | 1 | 0 | 2 |
| iStopC/iUTR1.1 | 1 | 3 | 2 | 6 |
| iUTR1.2 | 0 | 0 | 1 | 1 |
| iUTR2.1 | 0 | 1 | 0 | 1 |
| iUTR2.2 | 0 | 4 | 1 | 5 |
| iUTR2.3 | 2 | 4 | 0 | 6 |
| iUTR2.4 | 0 | 1 | 2 | 3 |
| iPolyA | 3 | 6 | 2 | 11 |
| iUTR3.1 | 0 | 2 | 3 | 5 |

### 4.7.8 Detection of sequence variations in FMRP and miRNA target sites

Sequence variations in 15 FMRP target regions and 9 miRNA target sites of the NF1 gene were identified (Ascano and Renwick, personal communication). We defined the position of each target site using the consensus sequence after aligning the PAR-CLIP and Target Scan data with the NG_009018.1 reference sequence. An example for the i2161 FMRP target site is shown in Figure 4.15.

```
                          i2161 (73343 – 73375)
73331  CTGCCTCTGGGGTTTTATTTTCTCTCAGCTGCAACAACTTCAATGCAGTCTT 73382
       -------------------TTCTCTCAGCTGCAACAACTTCAATG-------
       ------------------TTTCTCTCAGCTGCAACAACTTCAA---------
       ------------------TTTCTCTCAGCTGCAACAACC-------------
       ------------TTTTATTTTCTCTCAGCTGCAACAACTTCAA---------
       -------------TTTATTCTCTCTCAGCTGCAACAACTTCAATG-------
       ---CCTCTGGGGTTTTATTTTCTCTCAGCTGCAAC-----------------
```

Figure 4.15: Location of the i2161 FMRP target site (red) using PAR-CLIP data and the NF1 reference (NG_009018.1) sequence (bold).

In this study 179 of 190 sequence variations were detected in FMRP and miRNA target sites in the NF1 gene. Graphical representation of sequence variations in relation to NF1 sequences is depicted in Figure 4.16.

We identified 40 variations in FMRP target sites, and 114 variations in regions flanking the 15 FMRP target sites (i2161, i2163, i2174, i2181, i2193, i2209, i2231, i2233, i2241/42, i2249, i2251, i2252, i2254, and i2260). Higher numbers of sequence variations were found in specific target regions; 37 (24.0 %) in target site i2174, 27 (17.5 %) in target site i2181, and 19 (12.3 %) in target site i2193. All sequence variations in FMRP target sites are presented in Section 4.7.8.1.

We identified 4 sequence variations in a miRNA target site, and 21 sequence variations in regions flanking the 9 miRNA target sites (miR103/107, miR137, miR10a/b, miR128a/b, miR27a/b, miR490, miR 30a-5p/b/c/d/e-5p, and miR153). All variations are presented in Section 4.7.8.2.

One hundred twenty-four sequence variations (69 %) were located in exonic regions; the majority (24; 13.4%) were found in exon58; followed by 15 (8.4 %) variations in exon17, and 9 (5 %) variations in exon4.

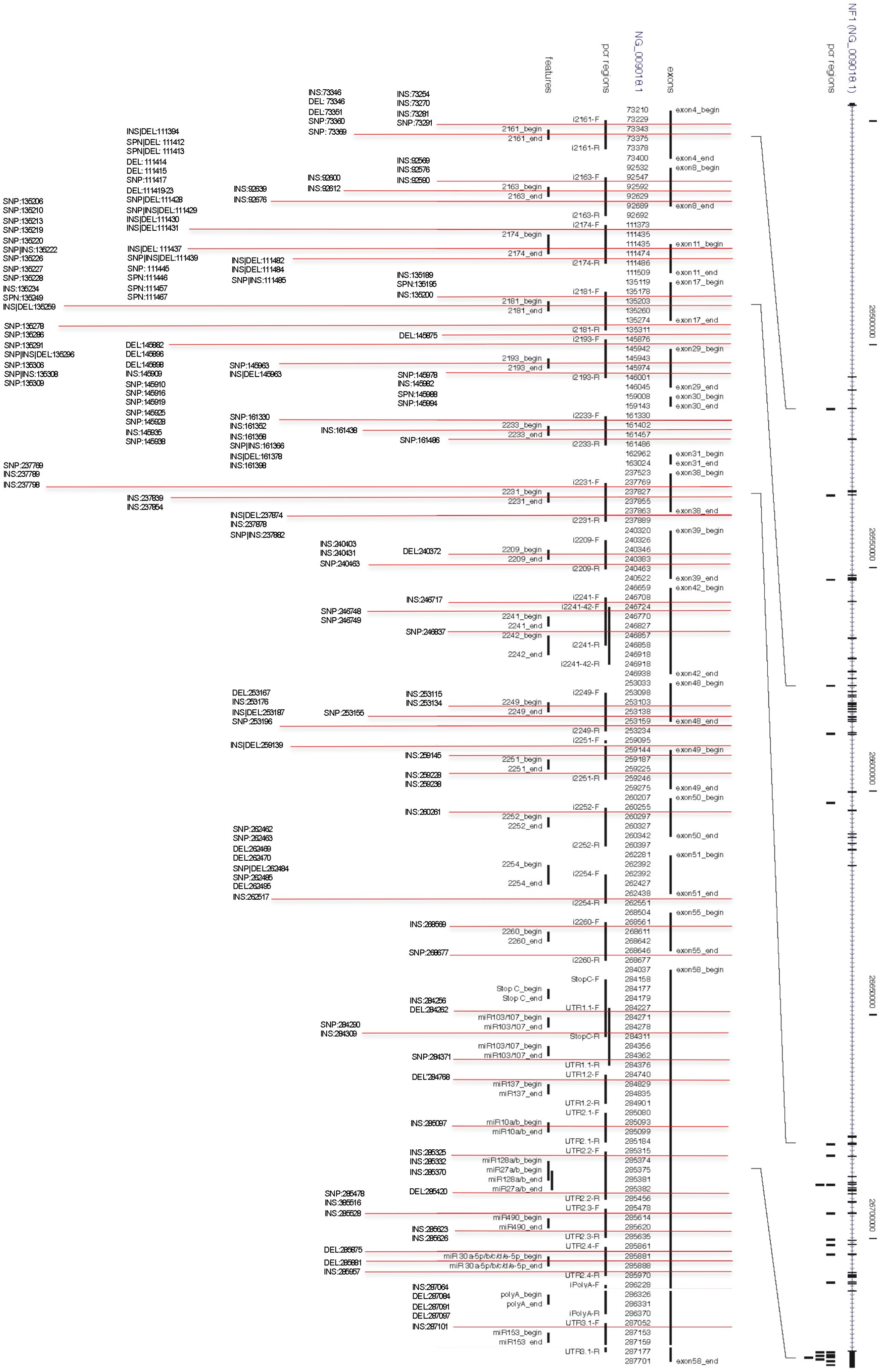


Figure 4.16: Overview of all sequence variations in defined regulatory regions of the NF1 gene (scale is not exact).

#### 4.7.8.1 Sequence variations in FMRP target regions

We identified 154 sequence variations in FMRP target sites (Table 4.8).

Table 4.8: Sequence variations in or flanking FMRP target regions. Listed are the FMRP target regions, the location of the variation in relation to the FMRP target site, the variation, the NF1 gene locus, the bioinformatic pipeline (MG=MOSIAK/GigaBayes; GS=GS Amplicon Analyzer), patient groups (1-4), and the number of identified patients per group. The location of each variation in relation to the FMRP target site is specified as followed: nucleotide upstream of the actual target site (>), in the target site (=), and nucleotides downstream of the target site (<).

| FMRP | Location | Var. | NF1 locus | Pipeline/Group(s) | # patients/ group |
|---|---|---|---|---|---|
| i2161 | 89> | -/A | 73254 | MG/4 | 37 |
| | 73> | -/A | 73270 | MG/4 | 52 |
| | 62> | -/T | 73281 | MG/4 | 73 |
| | 52> | T/C | 73291 | GS/3 | 10 |
| | = | -/T | 73346 | MG/4 | 36 |
| | = | T/- | 73346 | MG/4 | 34 |
| | = | T/- | 73351 | MG/4 | 44 |
| | = | T/C | 73360 | GS/3 | 10 |
| | = | T/C | 73369 | GS/3 | 10 |
| i2163 | 23> | -/T | 92569 | MG/4 | 56 |
| | 16> | -/A | 92576 | MG/4 | 32 |
| | 2> | -/A | 92590 | MG/4 | 34 |
| | = | -/T | 92600 | MG/2,3,4 | 17,44,77 |
| | = | -/A | 92612 | MG/4 | 14 |
| | <10 | -/A | 92639 | MG/2,4 | 11,14 |
| | <47 | -/A | 92676 | MG/4 | 18 |
| i2174 | 41> | -/T | 111394 | MG/4 | 20 |
| | 41> | T/- | 111394 | MG/2,3,4 | 29,57,83 |
| | 23> | C/- | 111412 | MG/4 | 80 |
| | 23> | C/T | 111412 | GS/2,3,4 | 30,16,57 |
| | 22> | T/- | 111413 | MG/4 | 79 |
| | 22> | T/C | 111413 | GS/2,3,4 | 30,16,57 |
| | 21> | T/- | 111414 | MG/4 | 76 |
| | 20> | T/- | 111415 | MG/4 | 43 |
| | 18> | T/C | 111417 | MG/4 | 36 |

Table 4.8: – continued on next page

Table 4.8: – continued from previous page

| FMRP | Location | Var. | NF1 locus | Pipeline/Group(s) | # patients/ group |
|---|---|---|---|---|---|
| | (16-12)> | TTTT/—- | 111419 | GS/2,3,4 \| MG/2,3,4 | 76,97,78 \| 37,12,60 |
| | 15> | T/- | 111420 | MG/1,2,3,4 | 11,97,99,98 |
| | 14> | T/- | 111421 | MG/1,2,3,4 | 62,94,100,100 |
| | 13> | T/- | 111422 | MG/1,2,3,4 | 13,78,84,100 |
| | 7> | T/- | 111428 | MG/1,2,3,4 | 54,94,100,98 |
| | 7> | T/C | 111428 | GS/2,3,4 | 38,61,24 |
| | 6> | -/C | 111429 | MG/2,3,4 | 96,100,89 |
| | 6> | C/- | 111429 | MG/2,3,4 | 94,99,69 |
| | 6> | C/T | 111429 | GS/2,3,4 | 38,61,24 |
| | 5> | -/T | 111430 | MG/4 | 12 |
| | 5> | T/- | 111430 | MG/4 | 11 |
| | 4> | -/A | 111431 | MG/2,3,4 | 55,70,47 |
| | 4> | A/- | 111431 | MG/2,3,4 | 55,74,48 |
| | = | -/C | 111437 | MG/4 | 13 |
| | = | C/- | 111437 | MG/4 | 10 |
| | = | -/G | 111439 | MG/4 | 31 |
| | = | G/- | 111439 | MG/4 | 31 |
| | = | G/T | 111439 | MG/2,3 \| GS/2,3,4 | 14,73 \| 59,96,29 |
| | = | C/T | 111445 | MG/2,3 \| GS/2,3,4 | 60,99 \| 75,98,46 |
| | = | T/G | 111446 | MG/2,3 \| GS/2,3,4 | 51,99 \| 75,98,46 |
| | = | C/A | 111457 | MG/2,3 \| GS/2,3,4 | 57,99 \| 75,98,46 |
| | = | -/T | 111467 | MG/4 | 20 |
| | <8 | -/A | 111482 | MG/2,3,4 | 74,19,31 |
| | <8 | A/- | 111482 | MG/2,3,4 | 66,16,30 |
| | <10 | -/A | 111484 | MG/4 | 54 |
| | <10 | A/- | 111484 | MG/4 | 35 |
| | <11 | -/T | 111485 | MG/4 | 11 |
| | <11 | T/A | 111485 | MG/2,3,4 | 28,95,53 |
| i2181 | 14> | -/A | 135189 | MG/2,3 | 17,74 |
| | 8> | G/A | 135195 | GS/3 | 81 |

Table 4.8: – continued on next page

Table 4.8: – continued from previous page

| FMRP | Location | Var. | NF1 locus | Pipeline/Group(s) | # patients/ group |
|---|---|---|---|---|---|
| | 3> | -/A | 135200 | MG/2,3,4 | 53,61,58 |
| | = | A/G | 135206 | GS/3 | 15 |
| | = | A/G | 135210 | GS/3 | 85 |
| | = | A/T | 135213 | GS/3 | 19 |
| | = | A/T | 135219 | GS/3 | 84 |
| | = | A/G | 135220 | GS/3 | 86 |
| | = | -/T | 135222 | MG/4 | 10 |
| | = | T/C | 135222 | GS/3 | 84 |
| | = | A/G | 135226 | GS/3 | 20 |
| | = | C/T | 135227 | GS/3 | 73 |
| | = | G/A | 135228 | GS/3 | 21 |
| | = | -/C | 135234 | MG/4 | 16 |
| | = | G/A | 135249 | GS/3 | 84 |
| | = | -/A | 135259 | MG/4 | 45 |
| | = | A/- | 135259 | MG/4 | 14 |
| | <18 | A/T | 135278 | GS/3 | 37 |
| | <26 | C/T | 135286 | GS/3 | 80 |
| | <31 | C/A | 135291 | GS/3 | 82 |
| | <36 | -/T | 135296 | MG/2,4 | 16,55 |
| | <36 | T/- | 135296 | MG/1,2,4 | 49,12,35 |
| | <36 | T/C | 135296 | GS/3 | 80 |
| | <46 | T/C | 135306 | GS/3 | 83 |
| | <48 | -/T | 135308 | MG/2,4 | 12,81 |
| | <48 | T/C | 135308 | GS/3 | 83 |
| | <49 | A/G | 135309 | GS/3 | 83 |
| i2193 | 68> | A/- | 145875 | MG/2,3 | QUAL |
| | 61> | T/- | 145882 | MG/4 | 11 |
| | 47> | A/- | 145896 | MG/4 | 28 |
| | 45> | A/G | 145898 | MG/3 \| GS/2,3,4 | 51 \| 40,92,52 |
| | 34> | -/G | 145909 | MG/4 | 10 |
| | 33> | T/C | 145910 | MG/3 \| GS/2,3,4 | 61 \| 40,92,54 |

Table 4.8: – continued on next page

Table 4.8: – continued from previous page

| FMRP | Location | Var. | NF1 locus | Pipeline/Group(s) | # patients/ group |
|---|---|---|---|---|---|
| | 27> | A/G | 145916 | MG/3 \| GS/2,3,4 | 44 \| 41,92,54 |
| | 24> | C/G | 145919 | MG/3 \| GS/2,3,4 | 42 \| 41,92,54 |
| | 18> | T/C | 145925 | MG/3 \| GS/2,3,4 | 46 \| 43,91,54 |
| | 15> | G/A | 145928 | MG/3 \| GS/2,3,4 | 49 \| 41,92,54 |
| | 8> | -/T | 145935 | MG/2,3 | 11,44 |
| | 5> | G/A | 145938 | MG/3 \| GS/2,3,4 | 40 \| 42,92,53 |
| | = | C/T | 145963 | MG/3 \| GS/2,3,4 | 52 \| 42,9151 |
| | = | -/A | 145968 | MG/4 | 90 |
| | = | A/- | 145968 | MG/4 | 35 |
| | <4 | C/G | 145978 | MG/3 \| GS/2,3,4 | 39 \| 39,92,52 |
| | <8 | -/T | 145982 | MG/2,3,4 | 25,74,30 |
| | <14 | G/A | 145988 | GS/3 | 23 |
| | <20 | T/A | 145994 | MG/3 \| GS/2,3,4 | 41 \| 40,92,53 |
| i2233 | 72> | T/C | 161330 | GS/1,2,3,4 | 89,71,91,72 |
| | 72> | T/G | 161330 | GS/1 | 12 |
| | 50> | -/A | 161352 | MG/4 | 20 |
| | 44> | -/C | 161358 | MG/4 | 32 |
| | 36> | -/T | 161366 | MG/2,3,4 | 29,19,39 |
| | 36> | T/C | 161366 | MG/1,2,3,4 | 35,41,45,37 |
| | | | | GS/1,2,3,4 | 46,44,59,43 |
| | 24> | -/T | 161378 | MG/2,3,4 | 45,42,71 |
| | 24> | T/- | 161378 | MG/3,4 | 41,11 |
| | 4> | -/T | 161398 | MG/2,3,4 | 13,13,67 |
| | = | -/T | 161438 | MG/4 | 67 |
| | >29 | G/A | 161486 | GS/1,2,4 | 15,10,10 |
| i2231 | 58> | A/G | 237769 | GS/1,2,3 | 11,19,13 |
| | 38> | -/T | 237789 | MG/4 | 53 |
| | 29> | -/A | 237798 | MG/4 | 55 |
| | = | -/T | 237839 | MG/3 | 14 |
| | = | -/C | 237854 | MG/4 | 39 |
| | <19 | -/T | 237874 | MG/2,4 | 32,39 |

Table 4.8: – continued on next page

Table 4.8: – continued from previous page

| FMRP | Location | Var. | NF1 locus | Pipeline/Group(s) | # patients/ group |
|---|---|---|---|---|---|
| | <19 | T/- | 237874 | MG/1,2,3,4 | 91,90,68,63 |
| | <13 | -/A | 237878 | MG/4 | 44 |
| | <27 | -/A | 237882 | MG/1 | 41 |
| | <27 | T/A | 237882 | MG/1,2,3 | 43,47,30 |
| | | | | GS/1,2,3,4 | 74,81,52,82 |
| i2209 | = | A/- | 240372 | MG/2,3 | 42,82 |
| | <20 | -/T | 240403 | MG/2,3 | 13,43 |
| | <48 | -/T | 240431 | MG/3 | 30 |
| | <80 | A/G | 240463 | GS/1,2,3,4 | 84,92,98,100 |
| i2241/42 | 53> | -/T | 246717 | MG/2,3 | 73,65 |
| | 22> | T/G | 246748 | GS/2,3,4 | 30,51,29 |
| | 21> | A/G | 246749 | GS/2,3 | 10,13 |
| | = | T/C | 246837 | GS/1,2,3,4 | 74,61,91,80 |
| i2249 | = | -/T | 253115 | MG/2,3 | 14,41 |
| | = | -/T | 253134 | MG/2 | 11 |
| | <17 | A/T | 253155 | GS/4 | 18 |
| | <29 | A/- | 253167 | MG/1,2,3 | 94,90,99 |
| | <38 | -/T | 253176 | MG/2 | 10 |
| | <49 | -/T | 253187 | MG/2,3 | 15,10 |
| | <49 | T/- | 253187 | MG/1 | 13 |
| | <58 | C/G | 253196 | MG/1,2,3 | 20,14,25 |
| | | | | GS/1,2,3,4 | 93,90,99,94 |
| i2251 | 48> | -/T | 259139 | MG/2,3 | 56,50 |
| | 48> | T/- | 259139 | MG/1,2,3 | 87,83,100 |
| | 42> | -/G | 259145 | MG/2,3 | 79,77 |
| | <3 | -/A | 259228 | MG/2,3 | 22,12 |
| | <13 | -/T | 259238 | MG/2,3 | 49,21 |
| i2252 | 36> | -/A | 260261 | MG/2,3 | 58,46 |
| i2254 | <35 | A/T | 262462 | GS/4 | 24 |
| | <36 | T/A | 262463 | GS/4 | 22 |
| | <42 | T/- | 262469 | MG/2 | 26 |

Table 4.8: – continued on next page

Table 4.8: – continued from previous page

| FMRP | Location | Var. | NF1 locus | Pipeline/Group(s) | # patients/ group |
|---|---|---|---|---|---|
| | <43 | T/- | 262470 | MG/1,2,3 | 88,98,100 |
| | <57 | A/- | 262484 | MG/1,2,3 | 90,96,97 |
| | <57 | A/T | 262484 | GS/4 | 13 |
| | <58 | T/A | 262485 | GS/4 | 13 |
| | <68 | T/- | 262495 | MG/1,2,3 | 35,33,29 |
| | <90 | -/A | 262517 | MG/2,3 | 47,90 |
| i2260 | 42> | -/A | 268569 | MG/2,3 | 51,34 |
| | <35 | G/A | 268677 | GS/3 | 10 |

#### 4.7.8.2 Sequence variations in miRNA target sites

We identified 25 sequence variations in miRNA target sites (Table 4.9).

Table 4.9: Sequence variations in or flanking miRNA target regions. Listed are the miRNA target regions, the location of the variation in relation to the miRNA target site, the variation, the NF1 gene locus, the bioinformatic pipeline (MG=MOSIAK/GigaBayes; GS=GS Amplicon Analyzer), patient groups (1-4), and the number of identified patients per group. The location (Loc) of each variation (Var) in relation to the miRNA target site is specified as followed: nucleotide upstream of the actual target site (>), in the target site (=), and nucleotides downstream of the target site (<).

| miRNA | Loc. | Var. | Locus | Pipeline(s)/ Group(s) | hits | non-affected | affected | unknown |
|---|---|---|---|---|---|---|---|---|
| miR103/107 | 15> | -/T | 284256 | MG/3 | 16 | 94 % | 6% | |
| | 9> | C/- | 284262 | MG/1,3 | 187 | 36 % | 64 % | |
| | = | A/G | 284290 | GS/2,3 | 46 | 46 % | 20 % | 34 % |
| | = | -/T | 284309 | MG/3 | 91 | 66 % | 25 % | 9 % |
| | <9 | -/T | 284371 | MG/3 | 43 | 81 % | 19 % | |
| miR137 | 61> | G/- | 284768 | MG/1,3 | 188 | 80 % | 13 % | 7 % |
| miR10a/b | = | -/G | 285097 | MG/3 | 51 | 57 % | 35 % | 8 % |
| miR128a/b | 49> | -/A | 285325 | MG/3 | 59 | 71 % | 24 % | 5 % |
| | 42> | -/T | 285332 | MG/3 | 26 | 73 % | 27 % | |
| | 4> | -/T | 285370 | MG/3 | 30 | 71 % | 29 % | |

Table 4.9: – continued on next page

Table 4.9: – continued from previous page

| miRNA | Loc. | Var. | Locus | Pipeline(s)/<br>Group(s) | hits | non-<br>affected | affected | unknown |
|---|---|---|---|---|---|---|---|---|
| miR27a/b | <45 | G/- | 285420 | MG/1,3 | 191 | 35 % | 61 % | 4 % |
| | <96 | A/C | 285478 | GS/2 | 13 | 8 % | | 92 % |
| | <96 | A/G | 285478 | GS/1,2,3 | 103 | 19 % | 37 % | 44 % |
| miR490 | 98> | -/T | 285516 | GS/3 | 11 | 72 % | 28 % | |
| | 86> | -/T | 285528 | GS/3 | 20 | 60 % | 25 % | 15 % |
| | <3 | -/C | 285623 | GS/3 | 28 | 39 % | 29 % | 32 % |
| | <6 | -/T | 285626 | GS/3 | 49 | 67 % | 27 % | 6 % |
| miR 30a... | 6> | T/- | 285875 | GS/1,3 | 192 | 35 % | 60 % | 5 % |
| | = | T/- | 285881 | GS/1 | 23 | | 100 % | |
| | <69 | -/T | 285957 | GS/3 | 29 | 76 % | 21 % | 3 % |
| miR153 | 89> | -/T | 287064 | MG/3 | 24 | 67 % | 29 % | 4 % |
| | 69> | T/- | 287084 | MG/1,3 | 192 | 35 % | 60 % | 4 % |
| | 62> | T/- | 287091 | MG/1,3 | 191 | 35 % | 60 % | 4 % |
| | 56> | T/- | 287097 | MG/1,3 | 98 | 22 % | 73 % | 5 % |
| | 52> | -/T | 287101 | MG/3 | 27 | 70 % | 30 % | |

# 5 Discussion

We developed a nested, multiplexed PCR method to assess sequence variation in multiple genomic sites in hundreds of individuals. We developed two assays to study FMRP and miRNA target regions in: (i) the NF1 gene in persons with and without autism, and (ii) miRNA genes from patients with chronic lymphocytic leukemias, Type 2 diabetes mellitus, and autism. High sample throughput was achieved using barcoded PCR primers, an automated pipetting station for PCR preparation and pooling of amplified products, and massive parallel (454) sequencing followed by bioinformatic analyses. Our results indicate that reliable sequencing data were generated for the majority of participants and that sequence variations and genetic mutations can be identified. Our method holds enormous promise for rapid, accurate and simultaneous interrogation of sequence variation in multiple genomic sites in hundreds of individuals.

## 5.1 Selection of target sites and genes

We evaluated FMRP and miRNA target site variation in the NF1 gene for the following reasons: (i) there is a striking clinical overlap between NF1, FXS, and Autism. (ii) NF1 is directly targeted by FMRP as determined by PAR-CLIP analysis and confirmed by Electrophoretic Mobility Shift Assay (EMSA) (Ascano, personal communication) and (iii) the total number of 24 targets sites, corresponding to 15 FMRP and 9 miRNA target sites, in the NF1 gene represents a challenging but manageable objective for this thesis project.

We evaluated miRNA genetic variations in 17 miRNA genes for following reasons; (i) miRNA 15a and 16 are also deleted in more than half of chronic lymphocytic leukemias (CLL) cases, (ii) miR-375 may a target for the treatment of diabetes, and (iii) multicopy miRNA genes (e.g. miR-7, miR-9, miR-124, and miR-200) may contain variations since they are under less selective pressure.

## 5.2 Method development

In recent years, several massive parallel sequencing technologies have been developed and have revolutionized the study of genetics due to their costs, effectiveness, accuracy, and high-throughput nature (Consortium, 2004; Church, 2006; Hall, 2007). These technologies have brought new opportunities, such as whole-genome sequencing, as well as challenges in the field of bioinformatics, especially in primary analysis of NGS data. Efficient mapping of short reads (< 200 bp) for instance is a considerable arithmetical problem (Mardis, 2008; Voelkerding et al., 2009). To improve workflows from data generation through to analysis and publication, streamlined bioinformatic pipelines must be developed; including adapter alignment algorithms, mapping tools, and the conversion of sequence data into an uniform format (Moore et al., 2010; Koboldt, 2010).

### 5.2.1 Assay strategy and PCR optimization

We designed and optimized a nested, multiplexed PCR assay based on PCR conditions that were previously established in the laboratory by Stefanie Großwendt for a project on miRNA gene sequence variations. We used a multiplexed PCR format to amplify multiple target regions simultaneously, saving time and resources by parallel amplification. We used nested PCR to increase sensitivity and specificity of target amplification.

Three critical modifications to the original approach were made: (i) adjustment of total primer concentration in the multiplexed PCR reaction, (ii) enhancement of the number of PCR cycles from 30 to 35 in multiplexed PCR, and (iii) elongation of the annealing time from 45 s to 1:15 min for multiplexed and nested PCR steps.

Lowering total primer concentration in the multiplexed reaction decreased the amount of primers that were carried forward to nested PCR reactions, reducing the possibility for non-specific amplification and primer dimer formation, which can consume dNTPs and compete with desired amplification products. In general a primer:dNTP ratio of 1:1,000 is recommended. We used a primer:dNTP ratio of 1:2,083 (720 nM primers related to 0.3 mM dNTPs) for NF1 targets. For the multiplexed amplification of miRNA genes we amplified distinct PCR products using a 5-fold higher total primer concentration; which used a primer:dNTP ratio of 1:176 (3.4 $\mu$M primers : 0.3 mM dNTPs).

The number of PCR cycles for multiplexed and nested PCRs were determined by the generation of visible PCR products prior to subsequent gel purification; PCR cycles were limited to minimize amplification errors. We used the AccuPrime$^{TM}$ DNA polymerase for multiplexed PCR because this enzyme provides (i) comparatively high 3' to 5' exonuclease proofreading activity, (ii) enhanced specific primer-template hybridization using anti-DNA polymerase antibodies, (iii) and high processivity for chain extension. Thirty-five amplification cycles were selected for the NF1 project because it increased product yield in some nested PCR reactions. Thirty-eight amplification cycles were selected for the multiplexed PCR for the miRNA gene project to generate visible amplification products following nested PCR.

For nested PCR, we used self-prepared *Taq* polymerase, because it is reliable for amplifying short PCR products (< 500 bp) and inexpensive. All nested PCR reactions were performed with 35 PCR amplification cycles and a final concentration of 200 nM primers (except for miRNA 124-3) and 0.2 mM dNTP, meeting the recommended primer:dNTP ratio.

Lengthening the annealing time in multiplexed and nested PCR reactions increased amplification efficiency. However it remains to be clarified whether elongation is required for both multiplexed and nested PCR steps. These experiments were performed using target-specific primers and we anticipated that an increase in annealing time would be beneficial for fusion primers of approximately 50 nt more in length.

### 5.2.2 Barcoded fusion primers

We introduced patient-specific barcoded primers in the nested PCR step to facilitate rapid matching of sequence reads with patient identifiers during bioinformatic analyses. Each patient was identified by a combination of forward and reverse barcode sequences, enabling pooling of PCR products. Using barcode combinations, instead of a single barcode, per patient reduced the total number of primers required for both projects from 2,400 to 480 barcoded primers for the NF1 project, and from 4,352 to 544 barcoded miRNAs primers for the miRNA gene project.

### 5.2.3 Automated sample preparation and pooling

We used the EpMotion 5075 Automated Pipetting Station to prepare PCR reactions and to pool amplification products. Benefits of this system included (i) accuracy and reduction of pipetting mistakes in comparison to manual pipetting, (ii) flexibility to change robotic commands with ease, (iii) opportunity to perform and observe ”dry run” experiments to ensure that sample and reagent transfer steps, tip changes, and volumes were pipetted appropriately, and (iv) control of technology for optimizing specific conditions to ensure sample protection (temperature etc.). PCR contaminations are unlikely in this system (Stefan Roth, Eppendorf Instrumente, Germany; personal communication).

PCR reactions were monitored by visualizing positive and negative controls, and an aliquot of pooled nested PCR regions following agarose gel electrophoresis. PCR amplification products were pooled in two steps. First, PCR products form one patient group were pooled by target region and purified. Prior to 454 sequencing, the concentrations of target regions were estimated visually on an agarose gel using a set of quantitative size markers, and an equimolar mixture of all target regions was prepared.

### 5.2.4 Massive parallel (454) sequencing

Several massive parallel sequencing technologies and platforms are available; including the Roche 454 Genome Sequencer (GS)[1], the Applied Biosystems (ABI) SOLiD Sequencer[2], and the Solexa (Illumina) Genome Analyzer[3]. Emerging technologies, like the Single Molecule Real Time (SMRT$^{TM}$) DNA Sequencing (Pacific Biosciences, USA) or Nanopore Sequencing (Oxford Nanopore Technologies®, UK) hold incredible promise for studying whole genomes (Tucker et al., 2009; Flusberg et al., 2010).

We chose massive parallel (454 GS) sequencing over Solexa or SOLiD sequencing because it provides sufficient read length of approximately 250 nt in comparison to other platforms with a read length of approximately 100 nt. The GS-FLX System generates about 400-600 megabases (Mb) of sequence data per 10 hr run at a cost of approx. $84 per Mb (Mardis, 2008). This system also has a high accuracy of 99.5 % per sequence read until the second-hundred base and reaches fidelity of 99.5 % with an appropriate sequencing depth.

### 5.2.5 Sequence read analysis pipeline

Typically analysis pipelines consist of: (i) conversion of data into a generic file format, (ii) sequence alignment, (iii) sequence assembly, and (iv) data analysis using several bioinformatic tools. We developed a pipeline to analyze our barcoded sequence data sets. The necessary computing power for each step was achieved by using the RNAworld server.

To assess the performance and accuracy of our nested multiplexed PCR assay, all trimmed group files were concatenated into one large group. The distribution of target-specifc primers and barcode sequences was assessed using Oligomap and application-oriented perl scripts. Oligomap was chosen because of its simplicity, accuracy and speed. Perl scripts were written to connect single pipeline steps and to generate a well-structured output file with all relevant information.

Sequence reads with target-specific primer sequences at both ends were identified for all 24 target regions of the NF1 gene project in similar proportions. One hundred patient-specific barcoded primers were detected with almost similar abundance per combination; only 0.3 % showed unexpected barcode combinations. Non-specific sequence reads were generated when the amplicons of two different nested PCR reactions overlapped.

[1] http://www.454.com/; available at March 3, 2010

[2] http://www.appliedbiosystems.com/absite/us/en/home/applications-technologies/solid-next-generation-sequencing.html; available at March 4, 2010

[3] http://www.illumina.com/; available at March 3, 2010

### 5.2.6 Sequence reads provide significant patient-specific information

When patient-specific combinations of barcodes and target-specific information is combined with a requirement of minimum insert length, inferences can be made concerning the lack or absence of sequence coverage for particular patients.

Using a matrix format, simple comparison of the number of reads with other regions for the same patient or other patients indicates whether the lack of coverage is due to (i) general amplification or sequencing problems of the region, (ii) quality problems of the patient sample or a suboptimal multiplexed PCR for this patient sample, or (iii) a longer deletion or suboptimal conditions of the corresponding nested PCR reaction.

A small subset of sequence reads were excluded for the following reasons; unexpected combinations of target-specific primers were found where two nested PCR regions overlapped. When one of those regions was not synthesized or sequenced to its end, the read got misinterpreted by the program which was designed to detect the first and the last target-specific sequence read. Other reads with a combination of two barcode sequences that are both only parts of forward primers or both only parts of reverse primers, may be explained as mistakes in barcoded sequencing.

Although barcodes are designed to prevent misinterpretation after single-nucleotide exchanges the risk cannot be excluded; the small number of these events shows that the detection of illegitimate barcode combination happens rarely and does not impact analysis. Barcodes F1, F2, R1, R4 and R5 show a noticeable accumulation of unusual combinations compared to other barcodes, potentially indicating that mistakes in sequencing turns the barcode sequence easily in to an alternative sequence.

### 5.2.7 Variations in FMRP and miRNA target sites

We used the Burrows-Wheeler and Smith-Waterman algorithms included in the BWA/Samtools and MOSAIK application, respectively for sequence read alignment, because of their ability to map longer reads and use multiple CPUs simultanoeusly. Additional alignment and mapping tools for massive parallel sequencing data for the 454 platform are Newbler[4], SSAHA2[5], and the GS software package (Roche). Other alignment and mapping tools include the Bowtie[6] and SOAP[7] aligner, mainly used for the Illumina platform, and SHRiMP[8] primarily used for SOLiD data.

[4] http://www.454.com/; available at April 3, 2010

[5] http://www.sanger.ac.uk/; available at April 6, 2010

[6] http://bowtie-bio.sourceforge.net/; available at April 6, 2010

[7] http://soap.genomics.org.ch/; available at April 7, 2010

[8] http://compbio.cs.toronto.edu/; available at April 14, 2010

Following alignment, single base pair polymorphism (variation and short-INDEL) discovery was performed using GigaBayes, Pileup (SAMtools), and GS Amplicon Variant Analyzer. GigaBayes was used based on its ability to (i) analyze large numbers of next-generation sequence reads, (ii) align sequence reads to chromosomes of model organisms or mammalian genomes, and (iii) generate binary output files. We used SAMtools and .SAM formatted files because of their flexibility to store alignment information from alignment formats generated through other alignment programs. SAMtools also provides utilities for manipulating alignments including sorting, merging, and indexing. The GS Amplicon Analyzer was used because sequence data were provided in .SFF format and the multiplexer matrix function enabled rapid matching of patient-specific barcodes to sequence reads.

We excluded all variations identified through GigaBayes and GS Amplicon Analyzer by discarding all variations occurring at a frequency less than 10 % per patient group. For Pileup analyses, we excluded sequence variations with a high quality score of < 1,000. We chose these boundaries to delimit the amount of sequence variations, and to exclude mismatches with a frequency of 1 %.

In this study,190 potential sequence variations were detected in the NF1 gene by GigaBayes and GS Amplicon. We excluded data from the BWA/SAMtools pipepline for variation analyses because of the missing ability to generate a multiplexer matrix. Sixteen of the remaining 179 variations were simultaneously identified by MOSAIK/GigaBayes and the GS Amplicon Analyzer. Because of the proprietory nature of the GS package, it was not possible to adapt the matching parameters to those from GigaBayes.

We identified 40 of 179 sequence variations in FMRP target sites, and 114 of 179 variations in regions flanking FMRP target sites at an average distance of 30 nt from the PAR-CLIP defined sequence of the FMRP target site. Higher numbers of sequence variations were found in target sites i2174, i2181 and i2193. Variations in these areas may result in transcript dysregulation.

We identified 4 of 179 sequence variations in a miRNA target site, and 21 of 179 variations in regions flanking miRNA target sites at an average distance of 48 nt from the seed sequence of the miRNA target site. Higher numbers of sequence variations were found in target sites miRNA103/104 and miRNA153. Sequence variations in these regions may also strengthen or weaken miRNA:target interactions, and perturb post-transcriptional regulatory networks.

## 5.3 Directions

Sequence variation is likely an underestimated source of regulatory and pathogenetic change in the human genome.

Our immediate plans are to (i) finish the SNP analysis and look for other forms of genetic variation in our data sets, (ii) examine relevant pedigrees for identified sequence variations, and (iii) confirm variations using conventional Sanger sequencing.

In the intermediate-term, we will consider a bioinformatic approach to select the most interesting target genes or target sites within a gene for sequencing. With bioinformatic ranking, criteria such as the relative proximity of target sites, information on strength derived from the length of the complementary seed sequence, and the expression of a gene in the organ of interest can be used to evaluate genes and target sites for disease relevance.

At present large population sequencing studies and analysis are still too expensive and predominantly limited to the study of model organisms (Futschik and Schlötterer, 2010); personalized whole-genome sequencing seems currently unthinkable. With impending reductions in sequence costs ($1000 per genome) these technologies will rapidly expand our knowledge of cancer genetics, epigenetics and genetically heterogenous disorders and disease.

# 6 Summary

We designed a nested multiplexed PCR assay followed by barcoded 454 pyrosequencing to assess multiple genomic regions in hundreds of individual patients. We first applied this method to study sequence variations in FMRP and miRNA target sites in the NF1 gene of 400 persons (4 patient groups) with and without autism. We modified the assay for increased sample throughput, and analyzed highly expressed, multicopy miRNA genes in multiple patients with chronic lymphocytic leukemias, Type 2 diabetes mellitus, and autism.

We developed a bioinformatic pipeline for data management and bioinformatic analyses. We used Oligomap for barcode sequence identification and read trimming. We identified all target regions in similar proportions for each patient group, and found similar abundances of barcoded primer combinations.

For sequence alignment, we compared three different sequence alignment algorithms; the MOSAIK package, the Burrows-Wheeler Aligner, and the Roche Genome Sequencer package. Single base pair polymorphism (variations and short-insertions and delutions) discovery was performed using GigaBayes, Pileup (SAMtools), and GS Amplicon. We excluded data from the BWA/ SAMtools pipepline because it was not possible to generate patient-specific data.

In our study, 179 sequence variations were detected; 40 variations were found in FMRP target regions (i2161, i2163, i2174, i2181, i2193, i2209, i2231, i2233, i2241/42, and i2249) and 4 variations were identified in miRNA target sites (miRNA103/107, miRNA10a/b, and miR 30a-5p/b/c/d/e-5p). We detected 114 variations in regions flanking an FMRP target region, and 21 variations in regions flanking a miRNA target site.

Sequence variation is likely an underestimated source of regulatory and pathogenetic changes in the human genome.

# References

Allen-Brady, K., Miller, J., Matsunami, N., Stevens, J., Block, H., Farley, M., Krasny, L., Pingree, C., Lainhart, J., Leppert, M., McMahon, W. M., and Coon, H. (2009). A high-density SNP genome-wide linkage scan in a large autism extended pedigree. Mol Psychiatry, 14(6):590–600.

Antar, L. N., Li, C., Zhang, H., Carroll, R. C., and Bassell, G. J. (2006). Local functions for FMRP in axon growth cone motility and activity-dependent regulation of filopodia and spine synapses. Mol Cell Neurosci, 32(1-2):37–48.

Araujo, F. (2009). Real-time PCR assays for high-throughput blood group genotyping. Methods Mol Biol, 496:25–37.

Ashley, C. T., Wilkinson, K. D., Reines, D., and Warren, S. T. (1993). FMR1 protein: conserved RNP family domains and selective RNA binding. Science, 262(5133):563–566.

Backe, P. H., Messias, A. C., Ravelli, R. B. G., Sattler, M., and Cusack, S. (2005). X-ray crystallographic and NMR studies of the third KH domain of hnRNP K in complex with single-stranded nucleic acids. Structure, 13(7):1055–1067.

Bartel, D. P. (2004). MicroRNAs: genomics, biogenesis, mechanism, and function. Cell, 116(2):281–297.

Bartel, D. P. (2009). MicroRNAs: target recognition and regulatory functions. Cell, 136(2):215–233.

Bassell, G. J. and Warren, S. T. (2008). Fragile X syndrome: loss of local mRNA regulation alters synaptic development and function. Neuron, 60(2):201–214.

Bear, M. F., Huber, K. M., and Warren, S. T. (2004). The mGluR theory of fragile X mental retardation. Trends Neurosci, 27(7):370–377.

Berninger, P., Gaidatzis, D., van Nimwegen, E., and Zavolan, M. (2008). Computational analysis of small RNA cloning data. Methods, 44(1):13–21.

Berry-Kravis, E. (2002). Epilepsy in fragile X syndrome. Dev Med Child Neurol, 44(11):724–728.

Boulle, K. D., Verkerk, A. J., Reyniers, E., Vits, L., Hendrickx, J., Roy, B. V., den Bos, F. V., de Graaff, E., Oostra, B. A., and Willems, P. J. (1993). A point mutation in the FMR-1 gene associated with fragile X mental retardation. Nat Genet, 3(1):31–35.

Brennecke, J., Stark, A., Russell, R. B., and Cohen, S. M. (2005). Principles of microRNA-target recognition. PLoS Biol, 3(3):e85.

Breslauer, K. J., Frank, R., Blöcker, H., and Marky, L. A. (1986). Predicting DNA duplex stability from the base sequence. Proc Natl Acad Sci U S A, 83(11):3746–3750.

Brown, V., Small, K., Lakkis, L., Feng, Y., Gunter, C., Wilkinson, K. D., and Warren, S. T. (1998). Purified recombinant FMRP exhibits selective RNA binding as an intrinsic property of the fragile X mental retardation protein. J Biol Chem, 273(25):15521–15527.

Budimirovic, D. B., Bukelis, I., Cox, C., Gray, R. M., Tierney, E., and Kaufmann, W. E. (2006). Autism spectrum disorder in Fragile X syndrome: differential contribution of adaptive socialization and social withdrawal. Am J Med Genet A, 140A(17):1814–1826.

Butler, M. G., Dasouki, M. J., Zhou, X.-P., Talebizadeh, Z., Brown, M., Takahashi, T. N., Miles, J. H., Wang, C. H., Stratton, R., Pilarski, R., and Eng, C. (2005). Subset of individuals with autism spectrum disorders and extreme macrocephaly associated with germline pten tumour suppressor gene mutations. J Med Genet, 42(4):318–321.

Buxbaum, J. D., Cai, G., Chaste, P., Nygren, G., Goldsmith, J., Reichert, J., Anckarsäter, H., Rastam, M., Smith, C. J., Silverman, J. M., Hollander, E., Leboyer, M., Gillberg, C., Verloes, A., and Betancur, C. (2007). Mutation screening of the PTEN gene in patients with autism spectrum disorders and macrocephaly. Am J Med Genet B Neuropsychiatr Genet, 144B(4):484–491.

Calin, G. A., Cimmino, A., Fabbri, M., Ferracin, M., Wojcik, S. E., Shimizu, M., Taccioli, C., Zanesi, N., Garzon, R., Aqeilan, R. I., Alder, H., Volinia, S., Rassenti, L., Liu, X., Liu, C.-G., Kipps, T. J., Negrini, M., and Croce, C. M. (2008). MiR-15a and miR-16-1 cluster functions in human leukemia. Proc Natl Acad Sci U S A, 105(13):5166–5171.

Carninci, P., Kasukawa, T., Katayama, S., Gough, J., Frith, M. C., Maeda, N., Oyama, R., Ravasi, T., Lenhard, B., Wells, C., Kodzius, R., Shimokawa, K., Bajic, V. B., Brenner, S. E., Batalov, S., Forrest, A. R. R., Zavolan, M., Davis, M. J., Wilming, L. G., Aidinis, V., Allen, J. E., Ambesi-Impiombato, A., Apweiler, R., Aturaliya, R. N., Bailey, T. L., Bansal, M., Baxter, L., Beisel, K. W., Bersano, T., Bono, H., Chalk, A. M., Chiu, K. P., Choudhary, V., Christoffels, A., Clutterbuck, D. R., Crowe, M. L., Dalla, E., Dalrymple, B. P., de Bono, B., Gatta, G. D., di Bernardo, D., Down, T., Engstrom, P., Fagiolini, M., Faulkner, G., Fletcher, C. F., Fukushima, T., Furuno, M., Futaki, S., Gariboldi, M., Georgii-Hemming, P., Gingeras, T. R., Gojobori, T., Green, R. E., Gustincich, S., Harbers, M., Hayashi, Y., Hensch, T. K., Hirokawa, N., Hill, D., Huminiecki, L., Iacono, M., Ikeo, K., Iwama, A., Ishikawa, T., Jakt, M., Kanapin, A., Katoh, M., Kawasawa, Y., Kelso, J., Kitamura, H., Kitano, H., Kollias, G., Krishnan, S. P. T., Kruger, A., Kummerfeld,

S. K., Kurochkin, I. V., Lareau, L. F., Lazarevic, D., Lipovich, L., Liu, J., Liuni, S., McWilliam, S., Babu, M. M., Madera, M., Marchionni, L., Matsuda, H., Matsuzawa, S., Miki, H., Mignone, F., Miyake, S., Morris, K., Mottagui-Tabar, S., Mulder, N., Nakano, N., Nakauchi, H., Ng, P., Nilsson, R., Nishiguchi, S., Nishikawa, S., Nori, F., Ohara, O., Okazaki, Y., Orlando, V., Pang, K. C., Pavan, W. J., Pavesi, G., Pesole, G., Petrovsky, N., Piazza, S., Reed, J., Reid, J. F., Ring, B. Z., Ringwald, M., Rost, B., Ruan, Y., Salzberg, S. L., Sandelin, A., Schneider, C., Schönbach, C., Sekiguchi, K., Semple, C. A. M., Seno, S., Sessa, L., Sheng, Y., Shibata, Y., Shimada, H., Shimada, K., Silva, D., Sinclair, B., Sperling, S., Stupka, E., Sugiura, K., Sultana, R., Takenaka, Y., Taki, K., Tammoja, K., Tan, S. L., Tang, S., Taylor, M. S., Tegner, J., Teichmann, S. A., Ueda, H. R., van Nimwegen, E., Verardo, R., Wei, C. L., Yagi, K., Yamanishi, H., Zabarovsky, E., Zhu, S., Zimmer, A., Hide, W., Bult, C., Grimmond, S. M., Teasdale, R. D., Liu, E. T., Brusic, V., Quackenbush, J., Wahlestedt, C., Mattick, J. S., Hume, D. A., Kai, C., Sasaki, D., Tomaru, Y., Fukuda, S., Kanamori-Katayama, M., Suzuki, M., Aoki, J., Arakawa, T., Iida, J., Imamura, K., Itoh, M., Kato, T., Kawaji, H., Kawagashira, N., Kawashima, T., Kojima, M., Kondo, S., Konno, H., Nakano, K., Ninomiya, N., Nishio, T., Okada, M., Plessy, C., Shibata, K., Shiraki, T., Suzuki, S., Tagami, M., Waki, K., Watahiki, A., Okamura-Oho, Y., Suzuki, H., Kawai, J., Hayashizaki, Y., Consortium, F. A. N. T. O. M., Group, R. I. K. E. N. G. E. R., and Group), G. S. G. G. N. P. C. (2005). The transcriptional landscape of the mammalian genome. Science, 309(5740):1559–1563.

Castillo, P. E., Francesconi, A., and Carroll, R. C. (2008). The ups and downs of translation-dependent plasticity. Neuron, 59(1):1–3.

Cavallaro, G., Basile, U., Polistena, A., Giustini, S., Arena, R., Scorsi, A., Zinnamosca, L., Letizia, C., Calvieri, S., and Toma, G. D. (2010). Surgical management of abdominal manifestations of type 1 neurofibromatosis: experience of a single center. Am Surg, 76(4):389–396.

Chen, L., Yun, S. W., Seto, J., Liu, W., and Toth, M. (2003). The fragile X mental retardation protein binds and regulates a novel class of mRNAs containing U rich target sequences. Neuroscience, 120(4):1005–1017.

Church, G. M. (2006). Genomes for all. Sci Am, 294(1):46–54.

Cohen, R. and Shuper, A. (2010). Developmental manifestation in children with neurofibromatosis type 1. Harefuah, 149(1):49–52, 61.

Consortium, I. H. G. S. (2004). Finishing the euchromatic sequence of the human genome. Nature, 431(7011):931–945.

Corona, G. and Toffoli, G. (2004). High throughput screening of genetic polymorphisms by matrix-assisted laser desorption ionization time-of-flight mass spectrometry. Comb Chem High Throughput Screen, 7(8):707–725.

Coutinho, A. M., Oliveira, G., Morgadinho, T., Fesel, C., Macedo, T. R., Bento, C., Marques, C., Ataíde, A., Miguel, T., Borges, L., and Vicente, A. M. (2004). Variants of the serotonin transporter gene (SLC6A4) significantly contribute to hyperserotonemia in autism. Mol Psychiatry, 9(3):264–271.

Darnell, J. C., Fraser, C. E., Mostovetsky, O., Stefani, G., Jones, T. A., Eddy, S. R., and Darnell, R. B. (2005). Kissing complex RNAs mediate interaction between the Fragile-X mental retardation protein KH2 domain and brain polyribosomes. Genes Dev, 19(8):903–918.

Dettman, J. R., Anderson, J. B., and Kohn, L. M. (2010). Genome-wide investigation of reproductive isolation in experimental lineages and natural species of neurospora: identifying candidate regions by microarray-based genotyping and mapping. Evolution, 64(3):694–709.

Dictenberg, J. B., Swanger, S. A., Antar, L. N., Singer, R. H., and Bassell, G. J. (2008). A direct role for FMRP in activity-dependent dendritic mRNA transport links filopodial-spine morphogenesis to fragile X syndrome. Dev Cell, 14(6):926–939.

Farh, K. K.-H., Grimson, A., Jan, C., Lewis, B. P., Johnston, W. K., Lim, L. P., Burge, C. B., and Bartel, D. P. (2005). The widespread impact of mammalian MicroRNAs on mRNA repression and evolution. Science, 310(5755):1817–1821.

Fiore, R., Siegel, G., and Schratt, G. (2008). MicroRNA function in neuronal development, plasticity and disease. Biochim Biophys Acta, 1779(8):471–478.

Flusberg, B. A., Webster, D. R., Lee, J. H., Travers, K. J., Olivares, E. C., Clark, T. A., Korlach, J., and Turner, S. W. (2010). Direct detection of DNA methylation during single-molecule, real-time sequencing. Nat Methods, 7(6):461–465.

Fombonne, E., Bolton, P., Prior, J., Jordan, H., and Rutter, M. (1997). A family study of autism: cognitive patterns and levels in parents and siblings. J Child Psychol Psychiatry, 38(6):667–683.

Friedman, L. M. and Avraham, K. B. (2009). MicroRNAs and epigenetic regulation in the mammalian inner ear: implications for deafness. Mamm Genome, 20(9-10):581–603.

Friedman, R. C., Farh, K. K.-H., Burge, C. B., and Bartel, D. P. (2009). Most mammalian mRNAs are conserved targets of microRNAs. Genome Res, 19(1):92–105.

Futschik, A. and Schlötterer, C. (2010). Massively Parallel Sequencing of Pooled DNA Samples–The Next Generation of Molecular Markers. Genetics.

Gabhane, S. K., Kotwal, M. N., and Bobhate, S. K. (2010). Segmental neurofibromatosis: a report of 3 cases. Indian J Dermatol, 55(1):105–108.

Gama-Carvalho, M., Barbosa-Morais, N. L., Brodsky, A. S., Silver, P. A., and Carmo-Fonseca, M. (2006). Genome-wide identification of functionally distinct subsets of cellular mRNAs associated with two nucleocytoplasmic-shuttling mammalian splicing factors. Genome Biol, 7(11):R113.

Germer, S. and Higuchi, R. (1999). Single-tube genotyping without oligonucleotide probes. Genome Res, 9(1):72–78.

Geschwind, D. H. (2008). Autism: many genes, common pathways? Cell, 135(3):391–395.

Gillberg, C. and Forsell, C. (1984). Childhood psychosis and neurofibromatosis–more than a coincidence? J Autism Dev Disord, 14(1):1–8.

Griffiths-Jones, S. (2004). The microRNA Registry. Nucleic Acids Res, 32(Database issue):D109–D111.

Griffiths-Jones, S., Grocock, R. J., van Dongen, S., Bateman, A., and Enright, A. J. (2006). mirbase: microrna sequences, targets and gene nomenclature. Nucleic Acids Res, 34(Database issue):D140–D144.

Griffiths-Jones, S., Saini, H. K., van Dongen, S., and Enright, A. J. (2008). miRBase: tools for microRNA genomics. Nucleic Acids Res, 36(Database issue):D154–D158.

Grimson, A., Farh, K. K.-H., Johnston, W. K., Garrett-Engele, P., Lim, L. P., and Bartel, D. P. (2007). Microrna targeting specificity in mammals: determinants beyond seed pairing. Mol Cell, 27(1):91–105.

Hafner, M., Landthaler, M., Burger, L., Khorshid, M., Hausser, J., Berninger, P., Rothballer, A., Ascano, M., Jungkamp, A.-C., Munschauer, M., Ulrich, A., Wardle, G. S., Dewell, S., Zavolan, M., and Tuschl, T. (2010). Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. Cell, 141(1):129–141.

Hall, N. (2007). Advanced sequencing technologies and their wider impact in microbiology. J Exp Biol, 210(Pt 9):1518–1525.

Han, J., Lee, Y., Yeom, K.-H., Kim, Y.-K., Jin, H., and Kim, V. N. (2004). The Drosha-DGCR8 complex in primary microRNA processing. Genes Dev, 18(24):3016–3027.

Hillier, L. W., Marth, G. T., Quinlan, A. R., Dooling, D., Fewell, G., Barnett, D., Fox, P., Glasscock, J. I., Hickenbotham, M., Huang, W., Magrini, V. J., Richt, R. J., Sander, S. N., Stewart, D. A., Stromberg, M., Tsung, E. F., Wylie, T., Schedl, T., Wilson, R. K., and Mardis, E. R. (2008). Whole-genome sequencing and variant discovery in c. elegans. Nat Methods, 5(2):183–188.

Jamain, S., Quach, H., Betancur, C., Råstam, M., Colineaux, C., Gillberg, I. C., Soderstrom, H., Giros, B., Leboyer, M., Gillberg, C., Bourgeron, T., and Study, P. A. R. I. S. (2003). Mutations of the X-linked genes encoding neuroligins NLGN3 and NLGN4 are associated with autism. Nat Genet, 34(1):27–29.

Jenkins, S. and Gibson, N. (2002). High-throughput SNP genotyping. Comp Funct Genomics, 3(1):57–66.

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

Keene, J. D. (2007). Rna regulons: coordination of post-transcriptional events. Nat Rev Genet, 8(7):533–543.

Keene, J. D., Komisarow, J. M., and Friedersdorf, M. B. (2006). RIP-Chip: the isolation and identification of mRNAs, microRNAs and protein components of ribonucleoprotein complexes from cell extracts. Nat Protoc, 1(1):302–307.

Keene, J. D. and Tenenbaum, S. A. (2002). Eukaryotic mRNPs may represent posttranscriptional operons. Mol Cell, 9(6):1161–1167.

Kelleher, R. J. and Bear, M. F. (2008). The autistic neuron: troubled translation? Cell, 135(3):401–406.

Kidd, J. M., Cooper, G. M., Donahue, W. F., Hayden, H. S., Sampas, N., Graves, T., Hansen, N., Teague, B., Alkan, C., Antonacci, F., Haugen, E., Zerr, T., Yamada, N. A., Tsang, P., Newman, T. L., Tüzün, E., Cheng, Z., Ebling, H. M., Tusneem, N., David, R., Gillett, W., Phelps, K. A., Weaver, M., Saranga, D., Brand, A., Tao, W., Gustafson, E., McKernan, K., Chen, L., Malig, M., Smith, J. D., Korn, J. M., McCarroll, S. A., Altshuler, D. A., Peiffer, D. A., Dorschner, M., Stamatoyannopoulos, J., Schwartz, D., Nickerson, D. A., Mullikin, J. C., Wilson, R. K., Bruhn, L., Olson, M. V., Kaul, R., Smith, D. R., and Eichler, E. E. (2008). Mapping and sequencing of structural variation from eight human genomes. Nature, 453(7191):56–64.

Koboldt, D. C. (2010). Challenges of sequencing human genomes. Brief Bioinform.

Kwiatkowski, T. J., Bosco, D. A., Leclerc, A. L., Tamrazian, E., Vanderburg, C. R., Russ, C., Davis, A., Gilchrist, J., Kasarskis, E. J., Munsat, T., Valdmanis, P., Rouleau, G. A., Hosler, B. A., Cortelli, P., de Jong, P. J., Yoshinaga, Y., Haines, J. L., Pericak-Vance, M. A., Yan, J., Ticozzi, N., Siddique, T., McKenna-Yasek, D., Sapp, P. C., Horvitz, H. R., Landers, J. E., and Brown, R. H. (2009). Mutations in the FUS/TLS gene on chromosome 16 cause familial amyotrophic lateral sclerosis. Science, 323(5918):1205–1208.

Kwok, P. Y. (2000). High-throughput genotyping assay approaches. Pharmacogenomics, 1(1):95–100.

Laggerbauer, B., Ostareck, D., Keidel, E. M., Ostareck-Lederer, A., and Fischer, U. (2001). Evidence that fragile x mental retardation protein is a negative regulator of translation. Hum Mol Genet, 10(4):329–338.

Landa, R. (2007). Early communication development and intervention for children with autism. Ment Retard Dev Disabil Res Rev, 13(1):16–25.

Landgraf, P., Rusu, M., Sheridan, R., Sewer, A., Iovino, N., Aravin, A., Pfeffer, S., Rice, A., Kamphorst, A. O., Landthaler, M., Lin, C., Socci, N. D., Hermida, L., Fulci, V., Chiaretti, S., Foà, R., Schliwka, J., Fuchs, U., Novosel, A., Müller, R.-U., Schermer, B., Bissels, U., Inman, J., Phan, Q., Chien, M., Weir, D. B., Choksi, R., Vita, G. D., Frezzetti, D., Trompeter, H.-I., Hornung, V., Teng, G., Hartmann, G., Palkovits, M., Lauro, R. D., Wernet, P., Macino, G., Rogler, C. E., Nagle, J. W., Ju, J., Papavasiliou, F. N., Benzing, T., Lichter, P., Tam, W., Brownstein, M. J., Bosio, A., Borkhardt, A., Russo, J. J., Sander, C., Zavolan, M., and Tuschl, T. (2007). A mammalian microRNA expression atlas based on small RNA library sequencing. Cell, 129(7):1401–1414.

Laxova, R. (1994). Fragile x syndrome. Adv Pediatr, 41:305–342.

Lee, Y., Ahn, C., Han, J., Choi, H., Kim, J., Yim, J., Lee, J., Provost, P., Rådmark, O., Kim, S., and Kim, V. N. (2003). The nuclear RNase III Drosha initiates microRNA processing. Nature, 425(6956):415–419.

Levy, S. E., Mandell, D. S., and Schultz, R. T. (2009). Autism. Lancet, 374(9701):1627–1638.

Lewis, B. P., Burge, C. B., and 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.

Lewis, H. A., Musunuru, K., Jensen, K. B., Edo, C., Chen, H., Darnell, R. B., and Burley, S. K. (2000). Sequence-specific RNA binding by a Nova KH domain: implications for paraneoplastic disease and the fragile X syndrome. Cell, 100(3):323–332.

Li, H. and Durbin, R. (2009). Fast and accurate short read alignment with burrows-wheeler transform. Bioinformatics, 25(14):1754–1760.

Li, H. and Durbin, R. (2010). Fast and accurate long-read alignment with burrows-wheeler transform. Bioinformatics, 26(5):589–595.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and Subgroup, . G. P. D. P. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics, 25(16):2078–2079.

Li, H., Wang, H.-Y., Cui, X., Luo, M., Hu, G., Greenawalt, D. M., Tereshchenko, I. V., Li, J. Y., Chu, Y., and Gao, R. (2007). High-throughput genotyping of single nucleotide polymorphisms with high sensitivity. Methods Mol Biol, 396:281–294.

Li, L., Meng, T., Jia, Z., Zhu, G., and Shi, B. (2010). Single nucleotide polymorphism associated with nonsyndromic cleft palate influences the processing of mir-140. Am J Med Genet A, 152A(4):856–862.

Licatalosi, D. D. and Darnell, R. B. (2006). Splicing regulation in neurologic disease. Neuron, 52(1):93–101.

Losh, M., Childress, D., Lam, K., and Piven, J. (2008a). Defining key features of the broad autism phenotype: a comparison across parents of multiple- and single-incidence autism families. Am J Med Genet B Neuropsychiatr Genet, 147B(4):424–433.

Losh, M., Sullivan, P. F., Trembath, D., and Piven, J. (2008b). Current developments in the genetics of autism: from phenome to genome. J Neuropathol Exp Neurol, 67(9):829–837.

Lunde, B. M., Moore, C., and Varani, G. (2007). Rna-binding proteins: modular design for efficient function. Nat Rev Mol Cell Biol, 8(6):479–490.

Mardis, E. R. (2008). The impact of next-generation sequencing technology on genetics. Trends Genet, 24(3):133–141.

Maris, C., Dominguez, C., and Allain, F. H.-T. (2005). The RNA recognition motif, a plastic RNA-binding platform to regulate post-transcriptional gene expression. FEBS J, 272(9):2118–2131.

Marth, G. T., Korf, I., Yandell, M. D., Yeh, R. T., Gu, Z., Zakeri, H., Stitziel, N. O., Hillier, L., Kwok, P. Y., and Gish, W. R. (1999). A general approach to single-nucleotide polymorphism discovery. Nat Genet, 23(4):452–456.

Martin, K. C. and Ephrussi, A. (2009). mrna localization: gene expression in the spatial dimension. Cell, 136(4):719–730.

Mayes, L., Volkmar, F., Hooks, M., and Cicchetti, D. (1993). Differentiating pervasive developmental disorder not otherwise specified from autism and language disorders. J Autism Dev Disord, 23(1):79–90.

McConkie-Rosell, A., Finucane, B., Cronister, A., Abrams, L., Bennett, R. L., and Pettersen, B. J. (2005). Genetic counseling for fragile x syndrome: updated recommendations of the national society of genetic counselors. J Genet Couns, 14(4):249–270.

McKee, A. E. and Silver, P. A. (2007). Systems perspectives on mRNA processing. Cell Res, 17(7):581–590.

Meister, G. and Tuschl, T. (2004). Mechanisms of gene silencing by double-stranded rna. Nature, 431(7006):343–349.

Mencía, A., Modamio-Høybjør, S., Redshaw, N., Morín, M., Mayo-Merino, F., Olavarrieta, L., Aguirre, L. A., del Castillo, I., Steel, K. P., Dalmay, T., Moreno, F., and Moreno-Pelayo, M. A. (2009). Mutations in the seed region of human mir-96 are responsible for nonsyndromic progressive hearing loss. Nat Genet, 41(5):609–613.

Meyer, K., Fredriksen, A., and Ueland, P. M. (2009). MALDI-TOF MS genotyping of polymorphisms related to 1-carbon metabolism using common and mass-modified terminators. Clin Chem, 55(1):139–149.

Moore, J. H., Asselbergs, F. W., and Williams, S. M. (2010). Bioinformatics challenges for genome-wide association studies. Bioinformatics, 26(4):445–455.

Moore, M. J. (2005). From birth to death: the complex lives of eukaryotic mRNAs. Science, 309(5740):1514–1518.

Mouridsen, S. E., Andersen, L. B., Sörensen, S. A., Rich, B., and Isager, T. (1992). Neurofibromatosis in infantile autism and other types of childhood psychoses. Acta Paedopsychiatr, 55(1):15–18.

Muddashetty, R. S., Kelic, S., Gross, C., Xu, M., and Bassell, G. J. (2007). Dysregulated metabotropic glutamate receptor-dependent translation of ampa receptor and postsynaptic density-95 mrnas at synapses in a mouse model of fragile x syndrome. J Neurosci, 27(20):5338–5348.

Muhle, R., Trentacoste, S. V., and Rapin, I. (2004). The genetics of autism. Pediatrics, 113(5):e472–e486.

Myers, S. M., Johnson, C. P., and of Pediatrics Council on Children With Disabilities, A. A. (2007). Management of children with autism spectrum disorders. Pediatrics, 120(5):1162–1182.

Napoli, I., Mercaldo, V., Boyl, P. P., Eleuteri, B., Zalfa, F., Rubeis, S. D., Marino, D. D., Mohr, E., Massimi, M., Falconi, M., Witke, W., Costa-Mattioli, M., Sonenberg, N., Achsel, T., and Bagni, C. (2008). The fragile x syndrome protein represses activity-dependent translation through cyfip1, a new 4e-bp. Cell, 134(6):1042–1054.

Nicoloso, M. S., Sun, H., Spizzo, R., Kim, H., Wickramasinghe, P., Shimizu, M., Wojcik, S. E., Ferdin, J., Kunej, T., Xiao, L., Manoukian, S., Secreto, G., Ravagnani, F., Wang, X., Radice, P., Croce, C. M., Davuluri, R. V., and Calin, G. A. (2010). Single-nucleotide polymorphisms inside microRNA target sites influence tumor susceptibility. Cancer Res, 70(7):2789–2798.

Nielsen, C. B., Shomron, N., Sandberg, R., Hornstein, E., Kitzman, J., and Burge, C. B. (2007). Determinants of targeting by endogenous and exogenous micrornas and sirnas. RNA, 13(11):1894–1910.

Noens, I., van Berckelaer-Onnes, I., Verpoorten, R., and van Duijn, G. (2006). The comfor: an instrument for the indication of augmentative communication in people with autism and intellectual disability. J Intellect Disabil Res, 50(Pt 9):621–632.

North, K. N., Riccardi, V., Samango-Sprouse, C., Ferner, R., Moore, B., Legius, E., Ratner, N., and Denckla, M. B. (1997). Cognitive function and academic performance in neurofibromatosis. 1: consensus statement from the nf1 cognitive disorders task force. Neurology, 48(4):1121–1127.

Pearson, W. R. and Lipman, D. J. (1988). Improved tools for biological sequence comparison. Proc Natl Acad Sci U S A, 85(8):2444–2448.

Persico, A. M. and Bourgeron, T. (2006). Searching for ways out of the autism maze: genetic, epigenetic and environmental clues. Trends Neurosci, 29(7):349–358.

Piao, X., Zhang, X., Wu, L., and Belasco, J. G. (2010). CCR4-NOT deadenylates mRNA associated with RNA-induced silencing complexes in human cells. Mol Cell Biol, 30(6):1486–1494.

Pieretti, M., Zhang, F. P., Fu, Y. H., Warren, S. T., Oostra, B. A., Caskey, C. T., and Nelson, D. L. (1991). Absence of expression of the FMR-1 gene in fragile X syndrome. Cell, 66(4):817–822.

Poy, M. N., Eliasson, L., Krutzfeldt, J., Kuwajima, S., Ma, X., Macdonald, P. E., Pfeffer, S., Tuschl, T., Rajewsky, N., Rorsman, P., and Stoffel, M. (2004). A pancreatic islet-specific microrna regulates insulin secretion. Nature, 432(7014):226–230.

Prasad, T. S. K., Kandasamy, K., and Pandey, A. (2009). Human Protein Reference Database and Human Proteinpedia as discovery tools for systems biology. Methods Mol Biol, 577:67–79.

Query, C. C., Bentley, R. C., and Keene, J. D. (1989). A common RNA recognition motif identified within a defined U1 RNA binding domain of the 70K U1 snRNP protein. Cell, 57(1):89–101.

Quinlan, A. R., Stewart, D. A., Strömberg, M. P., and Marth, G. T. (2008). Pyrobayes: an improved base caller for SNP discovery in pyrosequences. Nat Methods, 5(2):179–181.

Rahman, O. A., Sasvari-Szekely, M., Szekely, A., Faludi, G., Guttman, A., and Nemoda, Z. (2010). Analysis of a polymorphic microrna target site in the purinergic receptor p2rx7 gene. Electrophoresis.

Ritchie, D. M. and Thompson, K. (1983). The unix time-sharing system. Communications of the ACM, Volume 26 , Issue 1:84 – 89.

Rogers, S. J. (2009). What are infant siblings teaching us about autism in infancy? Autism Res, 2(3):125–137.

Rozen, S. and Skaletsky, H. (2000). Primer3 on the WWW for general users and for biologist programmers. Methods Mol Biol, 132:365–386.

Ruby, J. G., Stark, A., Johnston, W. K., Kellis, M., Bartel, D. P., and Lai, E. C. (2007). Evolution, biogenesis, expression, and target predictions of a substantially expanded set of drosophila micrornas. Genome Res, 17(12):1850–1864.

Saemundsen, E., Ludvigsson, P., and Rafnsson, V. (2007). Autism spectrum disorders in children with a history of infantile spasms: a population-based study. J Child Neurol, 22(9):1102–1107.

Saetrom, P., Heale, B. S. E., Snøve, O., Aagaard, L., Alluin, J., and Rossi, J. J. (2007). Distance constraints between microRNA target sites dictate efficacy and cooperativity. Nucleic Acids Res, 35(7):2333–2342.

Schaeffer, C., Bardoni, B., Mandel, J. L., Ehresmann, B., Ehresmann, C., and Moine, H. (2001). The fragile X mental retardation protein binds specifically to its mRNA via a purine quartet motif. EMBO J, 20(17):4803–4813.

Sebat, J., Lakshmi, B., Malhotra, D., Troge, J., Lese-Martin, C., Walsh, T., Yamrom, B., Yoon, S., Krasnitz, A., Kendall, J., Leotta, A., Pai, D., Zhang, R., Lee, Y.-H., Hicks, J., Spence, S. J., Lee, A. T., Puura, K., Lehtimäki, T., Ledbetter, D., Gregersen, P. K., Bregman, J., Sutcliffe, J. S., Jobanputra, V., Chung, W., Warburton, D., King, M.-C., Skuse, D., Geschwind, D. H., Gilliam,

T. C., Ye, K., and Wigler, M. (2007). Strong association of de novo copy number mutations with autism. Science, 316(5823):445–449.

Shen, M. H., Harper, P. S., and Upadhyaya, M. (1996). Molecular genetics of neurofibromatosis type 1 (NF1). J Med Genet, 33(1):2–17.

Shumaker, J. M., Metspalu, A., and Caskey, C. T. (1996). Mutation detection by solid phase primer extension. Hum Mutat, 7(4):346–354.

Sonenberg, N. and Hinnebusch, A. G. (2009). Regulation of translation initiation in eukaryotes: mechanisms and biological targets. Cell, 136(4):731–745.

Sreedharan, J., Blair, I. P., Tripathi, V. B., Hu, X., Vance, C., Rogelj, B., Ackerley, S., Durnall, J. C., Williams, K. L., Buratti, E., Baralle, F., de Belleroche, J., Mitchell, J. D., Leigh, P. N., Al-Chalabi, A., Miller, C. C., Nicholson, G., and Shaw, C. E. (2008). TDP-43 mutations in familial and sporadic amyotrophic lateral sclerosis. Science, 319(5870):1668–1672.

Tavazoie, S. F., Alvarez, V. A., Ridenour, D. A., Kwiatkowski, D. J., and Sabatini, B. L. (2005). Regulation of neuronal morphology and function by the tumor suppressors tsc1 and tsc2. Nat Neurosci, 8(12):1727–1734.

Tonsgard, J. H., Yelavarthi, K. K., Cushner, S., Short, M. P., and Lindgren, V. (1997). Do nf1 gene deletions result in a characteristic phenotype? Am J Med Genet, 73(1):80–86.

Tsuchihashi, Z. and Dracopoli, N. C. (2002). Progress in high throughput SNP genotyping methods. Pharmacogenomics J, 2(2):103–110.

Tucker, T., Marra, M., and Friedman, J. M. (2009). Massively parallel sequencing: the next big thing in genetic medicine. Am J Hum Genet, 85(2):142–154.

Turk, J. and Cornish, K. (1998). Face recognition and emotion perception in boys with fragile-x syndrome. J Intellect Disabil Res, 42 ( Pt 6):490–499.

Ule, J., Jensen, K., Mele, A., and Darnell, R. B. (2005). CLIP: a method for identifying protein-RNA interaction sites in living cells. Methods, 37(4):376–386.

van den Berg, A., Mols, J., and Han, J. (2008). Risc-target interaction: cleavage and translational suppression. Biochim Biophys Acta, 1779(11):668–677.

Venter, J. C. (2003). A part of the human genome sequence. Science, 299(5610):1183–1184.

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., Gocayne, J. D., Amanatides, P., Ballew, R. M., Huson, D. H., Wortman, J. R., Zhang, Q., Kodira, C. D., Zheng, X. H., Chen, L., Skupski, M., Subramanian, G., Thomas, P. D., Zhang, J., Miklos, G. L. G., Nelson, C., Broder, S., Clark, A. G., Nadeau, J., McKusick, V. A., Zinder, N., Levine, A. J., Roberts, R. J., Simon, M., Slayman, C., Hunkapiller, M., Bolanos, R., Delcher, A., Dew, I., Fasulo, D., Flanigan, M., Florea, L., Halpern, A., Hannenhalli, S., Kravitz, S., Levy, S., Mobarry, C., Reinert, K., Remington, K., Abu-Threideh, J., Beasley, E., Biddick, K., Bonazzi, V., Brandon, R., Cargill, M., Chandramouliswaran, I., Charlab, R., Chaturvedi, K., Deng, Z., Francesco, V. D., Dunn, P., Eilbeck, K., Evangelista, C., Gabrielian, A. E., Gan, W., Ge, W., Gong, F., Gu, Z., Guan, P., Heiman, T. J., Higgins, M. E., Ji, R. R., Ke, Z., Ketchum, K. A., Lai, Z., Lei, Y., Li, Z., Li, J., Liang, Y., Lin, X., Lu, F., Merkulov, G. V., Milshina, N., Moore, H. M., Naik, A. K., Narayan, V. A., Neelam, B., Nusskern, D., Rusch, D. B., Salzberg, S., Shao, W., Shue, B., Sun, J., Wang, Z., Wang, A., Wang, X., Wang, J., Wei, M., Wides, R., Xiao, C., Yan, C., Yao, A., Ye, J., Zhan, M., Zhang, W., Zhang, H., Zhao, Q., Zheng, L., Zhong, F., Zhong, W., Zhu, S., Zhao, S., Gilbert, D., Baumhueter, S., Spier, G., Carter, C., Cravchik, A., Woodage, T., Ali, F., An, H., Awe, A., Baldwin, D., Baden, H., Barnstead, M., Barrow, I., Beeson, K., Busam, D., Carver, A., Center, A., Cheng, M. L., Curry, L., Danaher, S., Davenport, L., Desilets, R., Dietz, S., Dodson, K., Doup, L., Ferriera, S., Garg, N., Gluecksmann, A., Hart, B., Haynes, J., Haynes, C., Heiner, C., Hladun, S., Hostin, D., Houck, J., Howland, T., Ibegwam, C., Johnson, J., Kalush, F., Kline, L., Koduru, S., Love, A., Mann, F., May, D., McCawley, S., McIntosh, T., McMullen, I., Moy, M., Moy, L., Murphy, B., Nelson, K., Pfannkoch, C., Pratts, E., Puri, V., Qureshi, H., Reardon, M., Rodriguez, R., Rogers, Y. H., Romblad, D., Ruhfel, B., Scott, R., Sitter, C., Smallwood, M., Stewart, E., Strong, R., Suh, E., Thomas, R., Tint, N. N., Tse, S., Vech, C., Wang, G., Wetter, J., Williams, S., Williams, M., Windsor, S., Winn-Deen, E., Wolfe, K., Zaveri, J., Zaveri, K., Abril, J. F., Guigó, R., Campbell, M. J., Sjolander, K. V., Karlak, B., Kejariwal, A., Mi, H., Lazareva, B., Hatton, T., Narechania, A., Diemer, K., Muruganujan, A., Guo, N., Sato, S., Bafna, V., Istrail, S., Lippert, R., Schwartz, R., Walenz, B., Yooseph, S., Allen, D., Basu, A., Baxendale, J., Blick, L., Caminha, M., Carnes-Stine, J., Caulk, P., Chiang, Y. H., Coyne, M., Dahlke, C., Mays, A., Dombroski, M., Donnelly, M., Ely, D., Esparham, S., Fosler, C., Gire, H., Glanowski, S., Glasser, K., Glodek, A., Gorokhov, M., Graham, K., Gropman, B., Harris, M., Heil, J., Henderson, S., Hoover, J., Jennings, D., Jordan, C., Jordan, J., Kasha, J., Kagan, L., Kraft, C., Levitsky, A., Lewis, M., Liu, X., Lopez, J., Ma, D., Majoros, W., McDaniel, J., Murphy, S., Newman, M., Nguyen, T., Nguyen, N., Nodell, M., Pan, S., Peck, J., Peterson, M., Rowe, W., Sanders, R., Scott, J., Simpson, M., Smith, T., Sprague, A., Stockwell, T., Turner, R., Venter, E., Wang, M., Wen, M., Wu, D., Wu, M., Xia, A., Zandieh, A., and Zhu, X. (2001). The sequence of the human genome. Science, 291(5507):1304–1351.

Verheij, C., Bakker, C. E., de Graaff, E., Keulemans, J., Willemsen, R., Verkerk, A. J., Galjaard, H., Reuser, A. J., Hoogeveen, A. T., and Oostra, B. A. (1993). Characterization and localization of the FMR-1 gene product associated with fragile X syndrome. Nature, 363(6431):722–724.

Verkerk, A. J., Pieretti, M., Sutcliffe, J. S., Fu, Y. H., Kuhl, D. P., Pizzuti, A., Reiner, O., Richards, S., Victoria, M. F., and Zhang, F. P. (1991). Identification of a gene (FMR-1) containing a CGG repeat coincident with a breakpoint cluster region exhibiting length variation in fragile X syndrome. Cell, 65(5):905–914.

Voelkerding, K. V., Dames, S. A., and Durtschi, J. D. (2009). Next-generation sequencing: from basic research to diagnostics. Clin Chem, 55(4):641–658.

Wang, W.-P., Ni, K.-Y., and Zhou, G.-H. (2006). Approaches for SNP genotyping. Yi Chuan, 28(1):117–126.

Waung, M. W., Pfeiffer, B. E., Nosyreva, E. D., Ronesi, J. A., and Huber, K. M. (2008). Rapid translation of Arc/Arg3.1 selectively mediates mGluR-dependent LTD through persistent increases in AMPAR endocytosis rate. Neuron, 59(1):84–97.

Yang, E., van Nimwegen, E., Zavolan, M., Rajewsky, N., Schroeder, M., Magnasco, M., and Darnell, J. E. (2003). Decay rates of human mRNAs: correlation with functional characteristics and sequence attributes. Genome Res, 13(8):1863–1872.

Yekta, S., Shih, I.-H., and Bartel, D. P. (2004). MicroRNA-directed cleavage of HOXB8 mRNA. Science, 304(5670):594–596.

# List of Figures

# List of Tables

# List of Abbreviations

| | |
|---|---|
| A | *Adenine* |
| ALS | *Amyotrophic Lateral Sclerosis* |
| ASD | *Autism spectrum disorders* |
| AVA | *Amplicon Variant Analyzer* |
| BAM | *Binary Alignment/Map* |
| BWA | *Burrows-Wheeler* Aligner |
| BWT | *Burrows-Wheeler* Transformation |
| C | *Cytosine* |
| CCD | *Charge Coupled Device* |
| CIGAR | *Compact Idiosyncratic Gapped* Alignment Report |
| CLI | *Commando Line Iinterface* |
| CNS | *Central nervous system* |
| CNV | *Copy number variations* |
| dsRBD | *double-stranded RBD* |
| EMSA | *Electrophoretic Mobility Shift Assay* |
| F/For | *Forward* |
| Fig | *Figure* |
| FMR1 | *Fragile X Mental Retardation 1* |
| FMRP | *Fragile X Mental Retardation Protein* |
| FTLD | *Frontotemporal Lobar Dementia* |
| FUS | *Fused in Sarcoma* |
| FXS | *Fragile X Syndrome* |
| G | *Guanine* |
| GB | *GigaByte* |
| Grep | *Global/ regular expression/ print* |
| GS | *Genome Sequencer* |
| GUI | *Graphical User Iinterface* |
| ISIZE | *Interred Insert Size* |
| KH | *K-homology* |
| lmp | *Low melting point* |
| Loc | *Location* |
| LTD | *Long-term depression* |

| | |
|---|---|
| MAPQ | *Mapping Quality* |
| max | *Maximum* |
| Mb | *megabases* |
| MG | *MOSAIK/GigaBayes* |
| min | *Minimum* |
| miRNA | *micro ribonuclein acid* |
| mRNA | *messenger ribonuclein acid* |
| NCBI | *National Center for Biotechnology Information* |
| NF1 | *Neurofibromatosis type 1* |
| NGS | *Next-generation sequencing* |
| nt | *Nucleotides* |
| OFR | *Open reading frame* |
| OS | *Operating System* |
| PAR-CLIP | *Photoactivatable-ribonucleoside-enhanced cross-linking and immunoprecipitation* |
| PAZ | *Piwi, Argonaute and Zwille/Pinhead* |
| PCR | *Polymerase Chain Reaction* |
| PDD | *Pervasive developmental disorders* |
| pre-(mi)RNA | *precursor-(mi)RNA* |
| pri-miRNA | *primary-miRNA* |
| R/Rev | *Reverse* |
| RBD | *RNA-binding domain* |
| RBP | *RNA-binding protein* |
| RFLP | *Restriction fragment length polymorphism* |
| RISC | *RNA-induced silencing* complex |
| RNA | *Ribonucleic acid* |
| RNP | *Ribonucleinprotein* |
| RRM | *RNA-recognition motif* |
| RT | *Room temperature* |
| SAM | *Sequence Alignment/Map* |
| siRNA | *small interfering RNA* |
| smM | *self-made Marker* |
| SMRT | *Single molecule real time* |
| SNP | *Single nucleotide polymorphism* |
| T | *Thymine* |
| Tab | *Table* |

| | |
|---|---|
| TDP | *TAR DNA binding Protein* |
| Tm | *Melting temperature* |
| Tview | *Text Alignment Viewer* |
| UTR | *Untranslated region* |
| Var | *Variation* |
| w/o | *without* |

# Appendix 1

## Multiplex primers for the NF1 project

Eighteen regions of the NF1 gene were amplified using 18 gene-specific primer pairs (Table 6.1).

Table 6.1: Primer pairs for multiplexed PCR for the NF1 targets; listed are the target region, the specific location on the NF1 gene, seventeen used primer pairs (F and R primer), and the expected product size (nt) of each amplicon (region).

| region | gene location | | gene-specific primer sequence | product size |
|---|---|---|---|---|
| o2161 | 72888 ↦ 73741 | F | 5’ -CTCACAGCAAACTGGGGATT- 3’ | 854 |
| | | R | 5’ -TGCTCACGGCAGATATTTTT- 3’ | |
| o2163 | 92163 ↦ 93020 | F | 5’ -CACAACTGCAAGGCAGAGAA- 3’ | 858 |
| | | R | 5’ -TGAACTCAAGTAGGGCAATCC- 3’ | |
| o2174 | 110940 ↦ 111845 | F | 5’ -TCTTCTGGCAGCTGGATTTT- 3’ | 906 |
| | | R | 5’ -GTTTGAAAAGCTTCGGGATG- 3’ | |
| o2181 | 134748 ↦ 135644 | F | 5’ -GTTATCCCAACATGGCACAG- 3’ | 897 |
| | | R | 5’ -ACCTTTGTTAGCCGGTACGA- 3’ | |
| o2193 | 145534 ↦ 146337 | F | 5’ -TGGTTGTCAACTTTGGGTTT- 3’ | 804 |
| | | R | 5’ -ATGCATGGAAAAAGGACAGG- 3’ | |
| o2209 | 239867 ↦ 240753 | F | 5’ -GGGCCATCTGATATCTGTCC- 3’ | 887 |
| | | R | 5’ -TCTTGCCGCTTTGCTTTTAT- 3’ | |
| o2231 | 237441 ↦ 238250 | F | 5’ -GGTTGGTTTCTGGAGCCTTT- 3’ | 810 |
| | | R | 5’ -TTGAACATTAGGCCTTCTTTTGT- 3’ | |
| o2233 | 161139 ↦ 161954 | F | 5’ -TGAGTCTGCTTCCCAAACCT- 3’ | 816 |
| | | R | 5’ -GGGAATTCTGGACATCCAAC- 3’ | |
| o2241/42 | 246391 ↦ 247194 | F | 5’ -TCATCAAAACCAGTGCAACAG- 3’ | 804 |
| | | R | 5’ -CATGGTCAACTGGCACTGAG- 3’ | |
| o2249 | 252708 ↦ 253527 | F | 5’ -TCTAGTGCCCTTTGGATGCT- 3’ | 820 |
| | | R | 5’ -GTGTGATCACGGCTCACTGT- 3’ | |
| o2251 | 258833 ↦ 259655 | F | 5’ -AATCTCTGTGCTGCTTTAGGG- 3’ | 823 |
| | | R | 5’ -GCTCATCCATTTGGAGTGGT- 3’ | |
| o2252 | 259911 ↦ 260739 | F | 5’ -TGAACCGTATCATCTTGAGCA- 3’ | 829 |
| | | R | 5’ -CAAATGTGTTAGCATGCCTGT- 3’ | |

Table 6.1: – continued on next page

Table 6.1: – continued from previous page

| region | gene location | | gene-specific primer sequence | product size |
|---|---|---|---|---|
| o2254 | 261986 ↦ 262879 | F | 5’ -GAGATTGGACACCCCTGTTG- 3’ | 894 |
| | | R | 5’ -TCAAAGAAAGCACGGCAAAT- 3’ | |
| o2260 | 268190 ↦ 268999 | F | 5’ -GCAGACAGAGCCAACCTTGT- 3’ | 810 |
| | | R | 5’ -CCAAAACCTGCAACAAATCA- 3’ | |
| oUTR1 | 283819 ↦ 284929 | F | 5’ -TGTTGCCTGGGAGAAACAG- 3’ | 1111 |
| | | R | 5’ -AGTTTGGATGACTGAGTTTGGA- 3’ | |
| oUTR2 | 285054 ↦ 286136 | F | 5’ -TTTTTCCCCTCCCCCTCTT- 3’ | 1038 |
| | | R | 5’ -CATCCTTGGCACTGGGTTAC- 3’ | |
| oUTR3 | 286684 ↦ 287635 | F | 5’ -GACCAAGTTGCCCATTTCTG- 3’ | 952 |
| | | R | 5’ -AGGAAAATGGGGTGAGGAAC- 3’ | |
| oPolyA | 286147 ↦ 286631 | F | 5’ -ACCATGGTCTTGGCTCTCC- 3’ | 485 |
| | | R | 5’ -TGCATTTCTCAAGGGACACA- 3’ | |

## Nested primers for the NF1 project

Twenty-four regions of the NF1 gene were amplified using 24 gene-specific primer pairs (Table 6.2).

Table 6.2: Primer pairs for nested PCR for the NF1 targets; listed are the target region, the specific location on the NF1 gene, twenty-four used primer pairs (F and R primer), and the expected product size (nt) of each amplicon (region) with ⊕ and without ⊖ F/ R - adapter and barcode sequence.

| target | gene location | | gene-specific primer sequence | product size |
|---|---|---|---|---|
| | | | | ⊖ / ⊕ |
| i2161 | 73209 ↦ 73401 | F | 5’ -GCAACCAAAGGACACAATGA- 3’ | 193 / 247 |
| | | R | 5’ -CCTGGTAGAAATGCGACTAAAGA- 3’ | |
| i2163 | 92527 ↦ 92713 | F | 5’ -TGCAGAATGTGCAGAAAAGC- 3’ | 187 / 241 |
| | | R | 5’ -TGGAAATAATTTTGCCCTCCT- 3’ | |
| i2174 | 111351 ↦ 111506 | F | 5’ -ATTTGCTGTTCTTTTTGGCTTC- 3’ | 156 / 210 |
| | | R | 5’ -GGTGATGATTCGATGGAGTG- 3’ | |
| i2181 | 135157 ↦ 135334 | F | 5’ -CGGGGTAGGATGTGATATTCC- 3’ | 178 / 231 |
| | | R | 5’ -TCATTCAGAAAACAAACAGAGCA- 3’ | |

Table 6.2: – continued on next page

Table 6.2: – continued from previous page

| target | gene location | | gene-specific primer sequence | product size |
|---|---|---|---|---|
| | | | | ⊖ / ⊕ |
| i2193 | 145856 ↦ 146021 | F | 5’ -TACAGAATGTGCAGGGCTGA- 3’ | 166 / 220 |
| | | R | 5’ -ACATGTTGCCAATCAGAGGA- 3’ | |
| i2209 | 240303 ↦ 240483 | F | 5’ -TTTGTGTTTTCTCCTAGGTCAGC- 3’ | 181 / 235 |
| | | R | 5’ -CTAACGTGAGGTGTGGCTCA- 3’ | |
| i2231 | 237749 ↦ 237909 | F | 5’ -AACTGTCACAGCCCGACTCT- 3’ | 161 / 215 |
| | | R | 5’ -AGCAACAAACCCCAAATCAA- 3’ | |
| i2233 | 161310 ↦ 161508 | F | 5’ -GGGAAATGTTGAAAGGGAAG- 3’ | 199 / 253 |
| | | R | 5’ -TCTGACACATGTTCACAGTTGG- 3’ | |
| i2241 | 246685 ↦ 246857 | F | 5’ -TGACAAGACATGCTTATCTCCAA- 3’ | 173 / 227 |
| | | R | 5’ -GCGGACCTGTGGCTACTAAG- 3’ | |
| i2241/42 | 246725 ↦ 246942 | F | 5’ -CATCTTATGTGGGATGATATTGC- 3’ | 218 / 272 |
| | | R | 5’ -TTACCACTAAAATGAAGCTGTGAA- 3’ | |
| i2249 | 253078 ↦ 253260 | F | 5’ -CACTGCAAGCAAATGGATCA- 3’ | 183 / 237 |
| | | R | 5’ -GGGCTAACTACTTCAATTTATTTCA- 3’ | |
| i2251 | 259075 ↦ 259265 | F | 5’ -CCTCAGCAGATGCTTGTTCA- 3’ | 191 / 245 |
| | | R | 5’ -GGCCACGCTCTGTGTATTC- 3’ | |
| i2252 | 260235 ↦ 260417 | F | 5’ -AAGTCGCTGCAGCCTAAAAC- 3’ | 183 / 237 |
| | | R | 5’ -GCTGCTTGCCTCCATTAGTT- 3’ | |
| i2254 | 262373 ↦ 262572 | F | 5’ -GAGCCAGGAAATCCATGAG- 3’ | 200 / 254 |
| | | R | 5’ -AAATGGCATCAAAAACTTTGC- 3’ | |
| i2260 | 268541 ↦ 268697 | F | 5’ -TCATTGTGCCAAGATCCAAA- 3’ | 157 / 211 |
| | | R | 5’ -GCGCATGTTAGCAAGTTCAT- 3’ | |
| iStopC | 284138 ↦ 284310 | F | 5’ -GCTGGCAGTTTCAAACGTAA- 3’ | 173 / 227 |
| | | R | 5’ -CAAACCGGATGGGTTCATTA- 3’ | |
| iUTR1.1 | 284228 ↦ 284398 | F | 5’ -TAGTGACCCCTTCCCTGTCC- 3’ | 171 / 225 |
| | | R | 5’ -CACGCCAAAAGTAGAAGAAAAA- 3’ | |
| iUTR1.2 | 284720 ↦ 284924 | F | 5’ -GCCTCAGTGACTTGACACCA- 3’ | 204 / 258 |
| | | R | 5’ -GGATGACTGAGTTTGGATAAGGA- 3’ | |
| iUTR2.1 | 285061 ↦ 285205 | F | 5’ -CCTCCCCCTCTTCTTTCCT- 3’ | 145 / 199 |
| | | R | 5’ -TCAGCTACCCTAAATGTCACG- 3’ | |

Table 6.2: – continued on next page

Table 6.2: – continued from previous page

| target | gene location | | gene-specific primer sequence | product size<br>⊖ / ⊕ |
|---|---|---|---|---|
| iUTR2.2 | 285295 ↦ 285476 | F | 5’ -TCTTCCTCCTCCTCTCCAAA- 3’ | 182 / 236 |
| | | R | 5’ -TGAAAATTCCAATGCCATGA- 3’ | |
| iUTR2.3 | 285457 ↦ 285656 | F | 5’ -TCATGGCATTGGAATTTTCAT- 3’ | 200 / 254 |
| | | R | 5’ -CCCCCTCAAAGCCATTATATC- 3’ | |
| iUTR2.4 | 285834 ↦ 285992 | F | 5’-AATGGTTTTGATACTCAGAATAACA-3’ | 159 / 213 |
| | | R | 5’ -GCACCTGTCTTCAGTATTTCCA- 3’ | |
| iUTR3.1 | 287032 ↦ 287199 | F | 5’ -AACCCTCCCAGGTTTGTAGG- 3’ | 168 / 222 |
| | | R | 5’ -TGTGTGGCTGACACTAAGAATG- 3’ | |
| iPolyA | 286208 ↦ 286391 | F | 5’ -CCTGGAATAGCAGGCAGTGT- 3’ | 184 / 238 |
| | | R | 5’ -CCTTGGTAAACCCGTTTATGG- 3’ | |

# Appendix 2

## Multiplex primers for the miRNA gene project

Seventeen regions encoding 17 miRNA genes were amplified using 17 gene-specific primer pairs (Table 6.3).

Table 6.3: Primer pairs for multiplexed PCR for the miRNA genes; listed are the target miRNA, the specific location in the chromosome, seventeen used primer pairs (F and R primer), and the expected product size (nt) of each amplicon (region).

| miRNA | gene locus | | gene-specific primer sequence | product size |
|---|---|---|---|---|
| miRNA7-1 | 9q21.32 | F | 5' -CCGAAGATCGGATCATTACC- 3' | 962 |
| | | R | 5' -GCAGAGGAATGTTGGCTTTT- 3' | |
| miRNA7-2 | 15q26.1 | F | 5' -GATCACTTGAGCCCAGGA- 3' | 952 |
| | | R | 5' -CGGTCTCCAGTGCATACCTC- 3' | |
| miRNA7-3 | 19p13.3 | F | 5' -GGGTAACAGCTGCCTGCTAA- 3' | 959 |
| | | R | 5' -AAGTGGTCCACCTGCCTTG- 3' | |
| miRNA9-1 | 1q22 | F | 5' -CAGAGAAGGGCAGTGGAGAC- 3' | 1009 |
| | | R | 5' -TCTCAGGTGACTTCCTCCAAA- 3' | |
| miRNA9-2 | 5q14.3 | F | 5' -TTAACCTGCGCTGGAAATTG- 3' | 966 |
| | | R | 5' -CAAATTGACATACAGCCCAAA- 3' | |
| miRNA9-3 | 15q26.1 | F | 5' -AAACTCTGATGGTCCGCAGT- 3' | 923 |
| | | R | 5' -GGACCAGGAAAGAGGAGGAC- 3' | |
| miRNA15A | 13q14.3 | F | 5' -TTTGAAAGGTGTACTGCAAGGA- 3' | 993 |
| | | R | 5' -GTAATTTCAAAACAAAGGGAAA- 3' | |
| miRNA16-1 | 13q14.3 | F | 5' -AAAGGTGCAGGcCATATTGT- 3' | 409 |
| | | R | 5'-TGCAATTACAGTATTTTAAGAGATGA- 3' | |
| miRNA124-1 | 8p23.1 | F | 5' -AAAAGCCTGGATGCGAAAG- 3' | 697 |
| | | R | 5' -GCCCAGAGAAAAATCTGCAC- 3' | |
| miRNA124-2 | 8q12.3 | F | 5' -ACATGCAATAGCGTGGTCCT- 3' | 1000 |
| | | R | 5' -GCGGCAAGTGTTCTTCAGAT- 3' | |
| miRNA124-3 | 20q13.33 | F | 5' -GCTCGCTGGGTTGTAAAAAG- 3' | 937 |
| | | R | 5' -GTGGACCCTCCCTTTGTCTC- 3' | |

Table 6.3: – continued on next page

Table 6.3: – continued from previous page

| miRNA | gene locus | | gene-specific primer sequence | product size |
|---|---|---|---|---|
| miRNA141 | 12p13.31 | F | 5' -TGTCCCTGTGTCAGCAACAT- 3' | 866 |
| | | R | 5' -ACTGCCAAATGCAGATAGGG- 3' | |
| miRNA200a | 1p36.33 | F | 5' -GGATTAGGACGCTCAGGTGT- 3' | 922 |
| | | R | 5' -AGCCAGGTCACCCTAAAACC- 3' | |
| miRNA200b | 1p36.33 | F | 5' -TACTGAGCTTCCCAGCGAGT- 3' | 900 |
| | | R | 5' -GGGTCACCTTTGAACATCGT- 3' | |
| miRNA200c | 12p13.31 | F | 5' -AGGGGTGAGACTAGGCAGGT- 3' | 981 |
| | | R | 5' -GTCGACTGTGGGTTCTGGAT- 3' | |
| miRNA375 | 2q35 | F | 5' -CCTCTCGGTTCCTCCTAACC- 3' | 868 |
| | | R | 5' -CCCGTATTACGACGCAGAAT- 3' | |
| miRNA429 | 1p36.33 | F | 5' -CTTCCCCCAGTCAACAAGAA- 3' | 936 |
| | | R | 5' -CTGTGACATTGTCCCTGCTG- 3' | |

## Nested primers for the miRNA gene project

Seventeen regions encoding 17 miRNA genes were amplified using 17 gene-specific primer pairs (Table 6.4).

Table 6.4: Primer pairs for nested PCR for the miRNA targets; listed are the target region, the specific location, 17 used primer pairs (F and R primer), and the expected product size (nt) of each amplicon (region) with ⊕ and without ⊖ F/ R - adapter and barcode sequence.

| miRNA | gene locus | | gene-specific primer sequence | product size |
|---|---|---|---|---|
| | | | | ⊖ / ⊕ |
| miRNA7-1 | 9q21.32 | F | 5' -GGTGAAAACTGCTGCCAAA- 3' | 207 / 261 |
| | | R | 5' -GCTGCATTTTACAGCACCAA- 3' | |
| miRNA7-2 | 15q26.1 | F | 5' -AACGCCTTGCAGAACTGG- 3' | 179 / 233 |
| | | R | 5' -CGTGGAAGGATAGCCAAAAA- 3' | |
| miRNA7-3 | 19p13.3 | F | 5' -GTCTCAGACATGGGGCAGA- 3' | 209 / 263 |
| | | R | 5' -CCGAGTGGAAGCGATTCTT- 3' | |

Table 6.4: – continued on next page

Table 6.4: – continued from previous page

| miRNA | gene locus | | gene-specific primer sequence | product size |
|---|---|---|---|---|
| | | | | ⊖ / ⊕ |
| miRNA9-1 | 1q22 | F | 5’ -AAGAGGCGGCGACAGCAG- 3’ | 217 / 271 |
| | | R | 5’ -TATCCTCTGGTGCTGGTCA- 3’ | |
| miRNA9-2 | 5q14.3 | F | 5’ -CTTCGGTACTGCCAGAAAGG- 3’ | 195 / 249 |
| | | R | 5’ -CTCTCGGCTGTAGTCTTTCATT- 3’ | |
| miRNA9-3 | 15q26.1 | F | 5’ -TGTGTCTGTCCATCCCCTCT- 3’ | 235 / 289 |
| | | R | 5’ -GCTCGCACGCAGAAGTTG- 3’ | |
| miRNA15A | 13q14.3 | F | 5’ -ATTCTTTAGGCGCGAATGTG- 3’ | 191 / 245 |
| | | R | 5’ -AGGCACTGCTGACATTGCT- 3’ | |
| miRNA16-1 | 13q14.3 | F | 5’ -TGCTGCCTCAAAAATACAAGG- 3’ | 197 / 251 |
| | | R | 5’ -ATATACATTAAAACACAACTGTAGA- 3’ | |
| miRNA124-1 | 8p23.1 | F | 5’ -CCACCCCTCTTCCTTTCTT- 3’ | 195 / 249 |
| | | R | 5’ -TCGGTCGCTCCTTCCTTG- 3’ | |
| miRNA124-2 | 8q12.3 | F | 5’ -GTGGTAATCGCAGTGGGTCT- 3’ | 180 / 234 |
| | | R | 5’ -GCTTTTTCCCCTTTCTGACC- 3’ | |
| miRNA124-3 | 20q13.33 | F | 5’ -GAAGACGCCTGAGCGTTC- 3’ | 209 / 263 |
| | | R | 5’ -GCCATTTCCATGAGAAAGGA- 3’ | |
| miRNA141 | 12p13.31 | F | 5’ -CCCTGTAGCAACTGGTGAGC- 3’ | 208 / 262 |
| | | R | 5’ -AGGGGTGAAGGTCAGAGGTT- 3’ | |
| miRNA200a | 1p36.33 | F | 5’ -GGTTCTTCCCTGGGCTTC- 3’ | 193 / 247 |
| | | R | 5’ -GAGTAGGAGCTCCGGATGTG- 3’ | |
| miRNA200b | 1p36.33 | F | 5’ -TACTGAGCTTCCCAGCGAGT- 3’ | 217 / 271 |
| | | R | 5’ -CTGTGTGGGAGGGGAGTGT- 3’ | |
| miRNA200c | 12p13.31 | F | 5’ -GATGGGAGCCAGGGATCT- 3’ | 180 / 234 |
| | | R | 5’ -ATGGATGTTGCTGACACAGG- 3’ | |
| miRNA375 | 2q35 | F | 5’ -GGAAGACCAGGACCAGGAG- 3’ | 219 / 273 |
| | | R | 5’ -ACCCGTACGGTTGAGATGG- 3’ | |
| miRNA429 | 1p36.33 | F | 5’ -CAGACCCAGGAGGGATCAG- 3’ | 220 / 274 |
| | | R | 5’ -GCTCTCTCCTCTAACGGTGAT- 3’ | |