Optimizing Molecular Inversion Probe Assays to Characterize Human-Specific Neurodevelopmental Genes

Author: Taylor Real

Supervisor: Dr. Sofie Salama
Dr. David Haussler

April 5th, 2020
Abstract

Copy number variation (CNV) of genes in the 2.3 Mb 1q21.1 distal duplication/deletion syndrome region of the human genome is highly correlated with neurodevelopmental disorders including microcephaly, macrocephaly, autism, and schizophrenia. NOTCH2NL is a human-specific gene with two paralogs within this region and we theorize that they contribute to the phenotypes of 1q21.1 duplication/deletion syndrome. NOTCH2NL is thought to regulate neural stem cell proliferation and thereby regulate the amount of neurogenesis in the developing brain. Population sequencing of NOTCH2NL genes revealed not only that paralogs A and B are breakpoints for the syndrome, but that there are small sequence variants (SSV) within different paralogs segregating in the population that we predict alter protein features, resulting in different activity levels. Our goal is to assess the total copy number of NOTCH2NL paralogs, as well as significant SSV in both neurotypical individuals and those with pathogenic 1q21.1 CNV events. To accomplish this we must first optimize Molecular Inversion Probes (MIPs), a PCR-based capture method. The MIPs protocol prepares products suitable for pooling a large number of samples for low cost, high-throughput sequencing. Molecular barcodes within the sequenced products are used to create ratios based on the numbers of reads corresponding to each allele detected within each MIP. Preliminary MIP experiments using gDNA with known genotypes have successfully captured NOTCH2NL variants. Characterizing NOTCH2NL allele profiles of those with typical and atypical neurodevelopment will help us understand the role of NOTCH2NL genes in the complex process of human brain development.
Acknowledgments

First and foremost, I want to thank my principal investigators, Dr. David Haussler and Dr. Sofie Salama, for giving me a research position in the Haussler Wet Lab. To date, this opportunity has been my most life changing experience. I want to especially thank Sofie for being a tremendous role model of the type of scientist I want to be.

I’d like to thank my graduate student mentor, Gary Mantalas, who has taught me nearly everything I know about molecular biology. I’m certain he’s one of the smartest people I’ll ever know, and getting to learn from him has been a privilege. I appreciate his endless patience with me and that he almost always laughs at my jokes.

I’d like to thank my other graduate student mentor, Colleen Bosworth, for holding my hand through all my computational firsts. I never would’ve been able to start the MIPs project without her help and she’s another example of the brilliant people I feel so lucky to have worked with the past two years.

I’d like to thank Nick Heyer for all his project contributions, specifically to our data analysis program. Without the major revisions he made to MipAnalysis, I would probably still be debugging the code as we speak.

I’d like to thank the RMI writing cohort for helping me through the laborious task of writing a thesis. I’d especially like to thank Zia Isola for the emotional, academic, and financial support she has provided as my program director for the last two years. Zia has always pushed me to be my best and go the extra mile for the things that I want. I know that I’m prepared to keep pushing myself in graduate school and beyond because of her.

I’d also like to acknowledge and thank my sources of funding and affiliations: the National Human Genome Research Institute’s (NHGRI) Research Mentoring Internship (RMI) program at the UC Santa Cruz Genomics institute, Simons Foundation Autism Research Initiative, and the Howard Hughes Medical Institute.

Finally I’d like to thank all my friends and family in Santa Cruz, Simi Valley, and all over. I love you all very much and not just because you allow me to complain to you endlessly. Most of all, I want to thank my mother Erin, who is always, always there for me.
## Contents

### Introduction
- A genomic perspective on human brain evolution ................................................. 4
- The role of NOTCH signalling in typical development ........................................ 6
- Brain development & NOTCH2NL variation ......................................................... 6
- A high throughput method for counting NOTCH2NL alleles .............................. 7

### Background
- NOTCH2NL diversity might be important for neural function ............................ 10

### Methods
- Molecular inversion probe design ................................................................. 13
- Molecular inversion probe phosphorylation ..................................................... 13
- Target capture .................................................................................................. 13
- Exonuclease treatment ..................................................................................... 14
- Library amplification ....................................................................................... 14
- Library clean-up .............................................................................................. 14
- Quality control ................................................................................................. 14
- Sequencing ....................................................................................................... 15
- Sequencing data analysis ................................................................................ 15

### Results & Discussion
- MIPs capture NOTCH2NL paratypes ............................................................... 16
- Optimizing the target capture efficiency of MIPs ............................................. 16
- Optimizing molecular counting of unique barcodes ........................................ 20
- Scaling up MIP assays for larger reactions ....................................................... 21

### Conclusion

### Appendix
- Oligo sequences ............................................................................................... 26
- Materials .......................................................................................................... 28
Introduction

A genomic perspective on human brain evolution

Since the sequencing and draft assembly of the first human genome in 2003, there have been tremendous advances in the study of genomics and biomedical research. These advances have been enriched by the growing field of comparative genomics, which characterizes the genomes of two or more organisms based on their similarities and differences. In 2005, the Chimpanzee Sequencing and Analysis Consortium published the first draft genome sequence of the chimpanzee, *Pan troglodytes*. This was a particularly impactful advancement for scientists seeking to use comparative genomics to understand human health and evolution. The genome of the chimpanzee, our closest living relative, gives us a genetic backdrop to study the evolutionary history of humans and has informed our understanding of the unique adaptive and developmental features of our genome since diverging from a common ancestor.

One key fact we have learned from studying the two genomes is that humans share nearly 99% DNA similarity with chimpanzees in protein coding regions (Waterson et al. 2005). Considering how much genetic material we share, one very significant phenotypic difference comes to mind—the brain. The human brain is 3x the size of the chimpanzee brain and has a more complex architecture of neural connections (Fiddes et al. 2018). We investigate uniquely human features of the genome in order to understand what is uniquely human about our brains.

A promising area of divergence to study are human-specific segmental duplications (Figure 1A). Segmental duplications are duplicated gene sequences that are substrates for genome evolution (Samonte and Eichler 2002, Marques-Bonet, Girirajan, and Eichler 2009). The mechanism of origin for segmental duplications is not well characterized, but it can result in “broken” sequences, or pseudogenes, as well as novel genes with functional importance (Samonte and Eichler 2002, Marques-Bonet, Girirajan, and Eichler 2009). A possible candidate for expanding the size of the brain is a group of novel human-specific genes, referred to as NOTCH2NL (Fiddes et al. 2018). NOTCH2NL genes are segmental duplications derived from the N-terminus of NOTCH2, a receptor in the NOTCH signalling pathway (Fiddes et al. 2018, Suzuki et al. 2018). The process of NOTCH2NL evolution began around eight million years ago in the Great Ape lineage (Figure 1C)(Fiddes et al. 2018). Due to earlier
Figure 1: (A) Segmental duplications arise when regions of DNA are donated and transposed to an acceptor region of the genome. The product of this can be both complete duplication of the sequences or a partial duplication of the sequences at the acceptor locus (Adapted from Samonte and Eichler 2002). (B) The maternal chromosome (green) and paternal chromosome (blue) contain three related segmental duplications (A, B, & C) that share very similar sequences. Typical crossing over between the chromosomes during meiosis produces gametes with genetic material from both parents. Misalignment of loci because of high sequence homology can lead to aberrant crossing over events that are often a cause for copy number variation. (C) A phylogenetic tree showing the evolutionary relationship between speciation within the great ape lineage and NOTCH2NL duplication events (From Fiddes et al. 2018).
Figure 2: \textit{NOTCH2NL} stimulates the \textit{NOTCH} signalling pathway, which regulates neural stem cell proliferation and differentiation. The self renewal of neural stem cells is promoted and the differentiation of neural stem cells is inhibited.

duplication events, NOTCH2NL-like pseudogenes exist in other primates like chimpanzees and gorillas, however the only functional copies exist in humans (Fiddes et al. 2018).

**The role of NOTCH signalling in typical development**

The idea that NOTCH2NL is helping to expand the size of the brain is supported by its functional role in the NOTCH signalling pathway. Activation of the NOTCH signalling pathway regulates neural stem cell proliferation (Zhou et al. 2010). NOTCH2NL is believed to activate the NOTCH signalling pathway, inhibiting neuronal differentiation and promoting neural stem cell renewal, therefore increasing the number of cells in the developing cerebral cortex (Figure 2) (Fiddes et al. 2018, Suzuki et al. 2018). NOTCH2NL is highly expressed in neural stem cells of the developing cerebral cortex (Fiddes et al. 2018). It is hypothesized that the function of NOTCH2NL is essential to the typical growth and development of the human brain (Fiddes et al. 2018).

**Brain development & NOTCH2NL variation**

Studying the sequence of NOTCH2NL further suggests that the genes play a role in neurodevelopment. Paralogs are related genes within an organism that descend from an
ancestral gene. There are five total NOTCH2NL paralogs: A, B, C, D, and the ancestral NOTCH2 (Fiddes et al. 2018). Three functional copies of NOTCH2NL—A, B, and C—exist in the human genome within the region 1q21.1. Two of those NOTCH2NL loci, A and B, are considered breakpoint sites for genomic copy number variations (CNV) resulting in 1q21.1 distal duplication/deletion syndrome (Figure 3). 1q21.1 distal duplication/deletion syndrome is a genomic syndrome that can display a large array of cognitive and physical phenotypes when there is CNV of genes in that chromosomal region (Mefford et al. 2008). Some of these phenotypes include microcephaly, macrocephaly, schizophrenia, and autism (Mefford et al. 2008). Neural tissue organoids with deletions of NOTCH2NL genes produce organoids smaller in size that express genes more associated with later time points in development than control organoids of the same age. (Fiddes et al. 2018).

Individuals who do not possess phenotypes associated with 1q21.1 distal duplication/deletion syndrome typically have a total of 4 NOTCH2NL A/B paralogs. (Fiddes et al. 2018). However, duplication/deletion events in NOTCH2NL and the 1q21 locus display incomplete penetrance (Mefford et al. 2008). Duplication and deletion events in NOTCH2NL are sometimes inherited from a “healthy parent” who does not display symptoms of 1q21.1 distal duplication/deletion syndrome (Fiddes et al. 2018), which is referred to as incomplete penetrance. In addition to differences in NOTCH2NL copy number, there are small sequence variants (SSV) that appear to alter the protein sequence and consequently the protein’s activity level (Fiddes et al. 2018). It is not definitively known how these SSV affect human brain development.

Figure 3: Illustration of chromosome 1, the 1q21.1 region, and the genes found in that region associated with 1q21.1 distal/duplication syndrome. NOTCH2NL paralogs shown in red (From Fiddes et al. 2018).
A high throughput method for counting NOTCH2NL alleles

To more comprehensively characterize the developmental effects of NOTCH2NL variation we need a greater volume of sequence data than currently available. Unfortunately, due to sequence identity greater than 99.3% over 200 kb, NOTCH2NL paralogs are difficult to sequence (Fiddes et al. 2018). Because of this we must develop an alternate way to measure those variant regions in a given individual. Here we use molecular inversion probes (MIPs) (Cantsilieris et al. 2017) as a method for genotyping NOTCH2NL. MIPs are single oligo assays with a strategy for identifying unique molecular sequences (Figure 4). An MIP is composed of a universal backbone sequence and Nmer molecular barcodes flanked by primer sequence arms oriented so that a circle is formed blocking the template strand and allowing for single molecule counting. Circular molecules go through PCR library amplification, which converts the circular molecules into linear molecules, adding Illumina sequencing adaptors and a sample ID barcode for pooling. These linear molecules are sequenced, separated by ID, and selected for unique molecular barcodes. An allele frequency can be called using the ratios of unique barcodes to NOTCH2NL alleles. While this method is not suitable for sequence and assembly of whole NOTCH2NL paralogs, MIP assays are advantageous for generating cost efficient genotype data of specific variants in a high-throughput manner.
Figure 4: (A) A single MIP oligo hybridizes to template gDNA using inverted targeting arms to prime the region of interest. The target extension arm is extended by the DNA polymerase Hemo Klentaq and ligated to the target ligation arm with ampligase. The unique circular molecules are amplified as linear molecules, which are then sequenced using next generation Illumina sequencing (Adapted from Cantsilieris et al. 2017). (B) The final linear MIP product consists of the captured region of interest flanked by two molecular barcodes and universal backbone sequence. There is a sample barcode and Illumina adaptor sequence (Adapted from Cantsilieris et al. 2017).
Background

NOTCH2NL diversity might be important for neural function

In the human genome there are four NOTCH2NL paralogs on chromosome 1 excluding the ancestral NOTCH2 (Fiddes et al. 2018). NOTCH2NLA and NOTCH2NLB have the highest sequence similarity among the paralogs, with evidence of gene conversion between them (Fiddes et al. 2018). Because of this, NOTCH2NLA and NOTCH2NLB behave like a singular gene with four possible copies (Fiddes et al. 2018). Previous Singly-Unique Nucleotides (SUNs) data of genomic sequences from the 1000 Genomes Project (Auton et al. 2015) and the Simon’s Diversity Project (Mallick et al. 2016, Nothaft 2017) has shown that the copy number of A/B within the population is stable (Fiddes et al. 2018). The position of paralogs A and B in the 1q21.1 distal duplication/deletion syndrome region, their low rate of copy number variation (CNV) in the population, and their association with atypical neurodevelopmental phenotypes when there is a CNV event suggest functional importance of A/B (Fiddes et al. 2018).

NOTCH2NLC lies next to paralogs A/B but outside the 1q21.1 distal duplication/deletion region, exhibits a greater CNV among the population, and consequently does not appear to be as essential as A/B (Fiddes et al. 2018). NOTCH2NLR is on the p arm of chromosome 1, near NOTCH2 (Fiddes et al. 2018). Unlike the paralogs A, B, and C, R lacks a 4 bp deletion in exon 5 necessary for protein expression (Fiddes et al. 2018). This 4 bp deletion is also absent in NOTCH2 (Fiddes et al. 2018). Sequence data from the Simon’s Diversity Project suggests that 14% of the population have 0 copies of NOTCH2NLR (Mallick et al. 2016, Fiddes et al. 2018). These characteristics suggest NOTCH2NLR is likely a pseudogene (Fiddes et al. 2018).

In addition to multiple unique paralogs, there are several small sequence variants (SSV) that help us distinguish different NOTCH2NL alleles (Fiddes et al. 2018). Known SSV alter the structure of the NOTCH2NL protein, and likely the protein’s activity level (Fiddes et al. 2018). The most NOTCH2-like allele is NOTCH2NL-L, or Long, and may be referred to as the wild-type paratype (Fiddes et al. 2018). It encodes a secretory signal, six EGF-like domains, and a fucosylation site in the fifth domain (Fiddes et al. 2018). The more
well characterized SSV are: an A/B specific G → A substitution at the canonical initiator methionine (NOTCH2NL-Sh or Short) and the C specific 2 bp deletion just downstream of the start site (NOTCH2NL-Sh2ntdel) which both interrupt the secretory signal and truncate exon 1; the A/B specific C → T substitution (NOTCH2NL-T197I) that removes the threonine fucosylation site and the G → A substitution (NOTCH2NL-A154T) that revives the threonine fucosylation site; and the A/B specific C → T substitution (NOTCH2NL-R113*) in exon 3 that creates a premature stop codon and a truncated protein that are presumably removed through nonsense-mediated decay (Fiddes et al. 2018). More recently we’ve identified other variant allele candidates including C19W, R113Q, P152L, and others. We believe we will continue to find more NOTCH2NL variants (Fiddes et al. 2018). Alleles distinguished by SSV are referred to as paratypes (paralog-haplotypes) because of their haplotype-like relationship yet varying paralogous locations (Fiddes et al. 2018). The typical profile of NOTCH2NL is 6 total paratypes as opposed to a standard gene with 2 alleles (Fiddes et al. 2018).

Because of both CNV among paralogs and the various heritable paratypes there is di-
versity in NOTCH2NL profiles. While we know there is CNV among A/B/C/R paralogs, there is little understanding of the correlation between NOTCH2NL paratypes and typical neurodevelopment. We predict these protein altering paratypes will have an additional effect on neural function (Fiddes et al. 2018).
Methods\textsuperscript{1}

Molecular inversion probe design

Each MIP is a single oligo with a universal backbone sequence, a molecular barcode of Ns, and inverted sequence specific flanking arms designed to hybridize to a template in a circular fashion (Appendix, Table 2 & Table 3). N bases represent a random base which will be incorporated during synthesis, resulting in a unique sequence that will be used for counting molecules. MIPs were designed using the MIPgen pipeline (Appendix, Table 1) (Boyle et al. 2014). NOTCH2NL variant probes were designed using a consensus sequence of all NOTCH2NL paralogs as a reference genome and a bed file containing NOTCH2NL paralog regions. Reference gene probes are designed using hg38 as a reference genome and a bed file containing the reference gene regions.

Molecular inversion probe phosphorylation

Prior to target capture, MIPs were pooled equimolar (0.1n uL of 100 uM oligo, where n = number of MIPs in the reaction) and combined with nuclease free water up to 25 uL. This MIP pool is phosphorylated using T4 PNK(10 U), 10X T4 DNA ligase buffer, nuclease-free water up to 30 uL and incubated at 37°C for 45 minutes and 65°C for 20 minutes.

Target capture

1 uL of phosphorylated MIP is diluted in 100 uL of nuclease-free water. 1 uL of the dilution and was combined with 100 ng (10 ng/uL) template gDNA, .25 mM dNTPs, Ampligase (1 U), 10X Ampligase buffer, Hemo Klentaq (3.2 U), nuclease-free water to a total of 25 uL, and incubated at 95°C for 10 minutes followed by 60°C for 22 hours. Once the MIPs hybridize to the appropriate region of the template strand, gap extension is facilitated with Hemo Klentaq, a non strand displacing DNA polymerase. The extension arm is then ligated

\textsuperscript{1}Detailed list of reagents and materials in Appendix
to the ligation arm using Ampligase, a thermostable ligase. The results of target capture were single stranded circular molecules.

Alternate ligation protocol: Phosphorylated MIPs were combined with template gDNA, dNTPs, Ampligase buffer, Hemo Klentaq, nuclease-free water to the same proportions as above and incubated at 95°C for 10 minutes, 70°C for 10 minutes with 1°C per minute ramp down to 60°C for 2 hours. Water and Ampligase were added to reaction on ice, and again incubated at 60°C for 1 hour.

Exonuclease treatment

Immediately after target capture, an exonuclease treatment was performed to remove all linear DNA with 2 μL of exonucleases I and III (4 U each), 10X Ampligase buffer, and nuclease-free water added to each reaction on a cold block on ice. The reaction was incubated at 37°C for 45 minutes and 95°C for 2 minutes.

Library amplification

5 μL of the exonuclease treated product was combined with 2X iProof Hi-Fidelity Master Mix, a 10 μM universal forward primer, a 10 μM sample-specific barcoded reverse primer (Appendix, Table 4), and nuclease-free water up to 25 μL. Both primers hybridized to the universal backbone sequence and added Illumina adaptor sequences. The PCR reaction was incubated at 98°C for 30s, followed by 35 cycles of 98°C for 10s, 60°C for 30s, and 72°C for 30s, with a final extension of 72°C for 2 minutes. The products of library amplification were double stranded linear molecules.

Library clean-up

5 μL of each 25 μL amplification reaction was pooled, cleaned up, and in earlier experiments size selected using a concentration of 0.9x magnetic SPRIselct beads. A more stringent concentration of 0.7x was adapted to remove small sequences.

After adding SPRIselct beads to a MIP library, the reaction was incubated for 3 minutes at room temperature and an additional 3 minutes room temperature on a magnet. The clear supernatant is removed and the beads are washed twice with a solution of 70% ethanol. All liquid was removed from the reaction and the beads were left to dry for 3-5 minutes. The dry beads were moved off the magnet, resuspended in elution buffer and incubated at room temperature for 3 minutes. The reaction was returned to the magnet and the clear supernatant which contains the product was pipetted to a clean tube.
Quality control

Concentration of DNA product was quantified using the Qubit dsDNA High Sensitivity Assay. Correct size range of DNA products was primarily validated using the Bioanalyzer High Sensitivity DNA kit and chip. Alternatively, correct size range of DNA products was sometimes validated with a 2% agarose gel electrophoresis at 100 volts for 45 minutes and stained post electrophoresis with ethidium bromide.

Sequencing

Products were sequenced at the UCSC Paleogenomics Lab using the MiSeq and NextSeq platform 300 cycle (2x150 bp paired-end reads) and a minimum sequencing depth of 100,000 reads per MIP in the reaction pool.

Sequencing data analysis

Illumina sequencing data will be separated by sample barcode ID and analyzed with the MipAnalysis program (https://github.com/Nheyer/MipAnalysis). The MipAnalysis program is designed to process sequencing statistics and calculate the frequency of NOTCH2NL alleles captured by MIPs, using FASTQ formatted files as an input. The program outputs a report of the sequencing statistics which includes average mapping quality per read, the average base quality per read, average depth of coverage per variant region, average overall coverage, average read size, the total number of molecular barcodes, and three histograms showing the distribution of read mapping qualities, read base qualities, and depth of coverage. The program continues to output the number of barcodes per set variant region, as well as the allele frequencies present in a set variant region.
Results & Discussion

MIPs capture NOTCH2NL paratypes

In my preliminary experiment I used four MIPs designed to capture paratype regions M1I/CC6DEL (NOTCH2NL-Sh & NOTCH2NL-Sh2ntdel), R113*, A154T, and T197I with molecular barcodes containing 5 Ns, which is the default pipeline design (Appendix, Table 2). This preliminary MIP reaction was sequenced using the MiSeq 300 cycle V2 kit and after aligning the reads to the consensus NOTCH2NL reference using the Integrative Genomics Viewer (IGV) (Robinson et al. 2011), showed the successful capture of the major variant regions using cell line gDNA from H9, NA12878, NA19240, and SV72l (Table 1, Figure 6). While all variant regions were successfully captured, not all variants themselves were. Genotype NA19240 has 2 copies of NOTCH2NLC, both NOTCH2NL-Sh2ntdel paratypes (Table 1B). In our sequencing data, I observed reads corresponding to the NOTCH2NL-Sh2ntdel region of genotype NA19240 but not the actual mutant sequence (Figure 6A). The same was true for MIP-R113* (not pictured) which captured sequences alignable to the R113* paratype but did not contain the actual mutation. 1 copy of NOTCH2NL-R113* was expected in H9 and SV72l.

However, I also found that there was a population of reads with universal backbone sequences but no alignable NOTCH2NL sequences. I believed these were MIPs that had ligated to themselves instead of amplifying a template. This observation suggested some inefficiency of the MIPs ability to capture variants. Additionally, the maximum number of possible random molecular barcodes (1024) was reached by the single molecules. Unique molecules cannot be confidently distinguished in a reaction pool with a complete saturation of possible molecular barcodes because of the possibility of doublets.

Optimizing the target capture efficiency of MIPs

The success of a MIPs library prep is largely dependent on the success of the 22 hour long target capture where MIPs are annealing to templates and being synthesized. Through literature research, I found alternate ligation protocols for the target capture step (Lin et al. 2010, Lau et al. 2014) that shortened the reaction time significantly. I adapted a similar
A. Paratype | Paralog | Mutation | Position  
--- | --- | --- | ---  
NOTCH2NL-Sh | A/B | ATG → ATA | 127052  
NOTCH2NL-Sh2ntdel | C | GCCCG → GCCG | 127064  
NOTCH2NL-R113* | A/B | CGA → TGA | 191863  
NOTCH2NL-A154T | A/B | GCA → ACA | 199997  
NOTCH2NL-T197I | A/B | ACC → ATC | 200127  

B. Paratype | H9 | NA12878 | NA19240 | SV721  
--- | --- | --- | --- | ---  
NOTCH2NL-L | 1 | 0 | 0 | 1  
NOTCH2NL-L, T197I | 0 | 1 | 0 | 0  
NOTCH2NL-Sh | 1 | 3 | 1 | 1  
NOTCH2NL-Sh, T197I | 1 | 0 | 2 | 2  
NOTCH2NL-Sh2ntdel | 2 | 2 | 2 | 2  
NOTCH2NL-L, R113* | 1 | 0 | 0 | 1  
NOTCH2NL-L, A154T | 0 | 0 | 1 | 0  
TOTAL | 6 | 6 | 6 | 7  

Table 1: (A) Table of NOTCH2NL paratypes corresponding to MIPs in the preliminary capture reaction. Includes their mutation and position in the consensus NOTCH2NL reference “genome”, which was used to design MIPs and align to IGV. (B) Table of human cell lines H9, NA12878, NA19240, and SV721 and their known copy number of NOTCH2NL paratypes.

alternate ligation protocol in the hopes of reducing the reaction time and as a comparison to the template capture abilities of the original ligation protocol. I conducted a test of the alternate ligation protocol against the original ligation protocol using a single MIP (P152L_A154T, Appendix 3) with a single cell line sample (NA19240) per reaction. Before moving onto sequencing, I needed to verify that I had the correct product size. I used the Agilent 2100 Bioanalyzer instrument to measure the size and concentration of DNA fragments within my MIP reaction pool. The expected size of the primary product from this reaction was 326 bp, or in the general 320 bp range. When the Bioanalyzer chip results were compared side by side to the original ligation step, the alternate protocol produced fewer molecules in the expected size range (ESR) (Figure 7B). The clearly larger peak in the 320 bp size range of the original protocol indicated that there was a greater concentration of DNA of that size (Figure 7A). I concluded based on these traces that the original ligation protocol was more efficient for capturing variants.

I then performed a temperature gradient for the annealing temperature during target capture using the same probe and gDNA genotype. In addition to optimizing annealing temperature of our assays, I wanted to test the range of a single probe. A possible concern
Figure 6: (A) NOTCH2NL-Sh and NOTCH2NL-Sh2ntdel variants are physically close enough to be captured by the same MIP. However, because they are A/B and C specific paratypes respectively, both mutant genotypes cannot be found in the same read. All genotypes have at least one Sh paratype and all have 2 Sh-2ntdel paratypes. These paratypes were captured in all genotypes except for NA12878 NOTCH2NL-Sh2ntdel. (B) NA19240 is the only genotype with a NOTCH2NL-A154T paratype copy. The only appearance of mutant sequences corresponding to this paratype were found in sample NA19240. (C) There is at least one T197I paratype, both Sh and L, in each genotype expected and the mutant T197I was captured in all samples.

of scaling up MIP reactions to include more probes is some probes having different properties that interfere with the reaction.
Figure 7: (A) The sample using the original ligation protocol’s most prominent peak is in the 320 bp size range (red), as expected for a successful MIP reaction. (B) The sample using the alternate ligation protocol does not have a prominent peak in the ESR (red) and implies a failure to capture the correct regions.

The original protocol uses an annealing and extension step at 60°C, and I used a gradient of 58°C, 59°C, 60°C, and 61°C. Based on the Bioanalyzer chip results, there were no apparent differences in molecules of ESR across the temperature gradient (Figure 8). The relatively unchanging 320 bp range peak size across different temperatures suggests that this specific probe is not sensitive to moderate changes in annealing temperature and that it likely has little effect on the capture efficiency. If I see future evidence of some probes’ sensitivity to annealing temperature, a larger scale reaction of this same experiment can be performed.

Figure 8: A temperature gradient of 4 degrees (58-61°C) was done on 8 samples (2 replicates). There was no clear deviation between the different traces, implying all the samples performed similarly during target capture. All traces have prominent peaks in the 320 bp size region.
The next variable I wanted to alter to understand its effect on MIP capture efficiency was the probe concentration. In the original protocol with a single MIP I were using .01 ul of the phosphorylated MIP “pool” (.04 uM). I performed a gradient of .005 ul, .01 ul, .02 ul, and .04 ul of the same probe with the same gDNA genotype. Agarose gel electrophoresis of these samples revealed one large smear with a concentrated band in the ESR. There was not a specific distinction between the different lanes. These samples were sent for Sanger sequencing, and returned sequences of incorrect sizes with low quality bases that do not map to NOTCH2NL regions. This was consistent with no-template MIPs and other non-specific sequences contaminating the primary product.

Our two main goals for optimizing MIP assays were to improve capture efficiency, or decrease the number of no-template MIPs, and be able to count molecules to calculate NOTCH2NL allele frequency. The target capture optimizations I tested had very little effect on the protocol I was using, and the presence of the no-template MIPs remained the same throughout. It became clearer to us that the problem could be too much efficiency during target capture, which was supported by consistently maxing out the possible number of unique molecular barcodes in following experiments. Moving forward I focused on optimizations that would correct our molecular counting abilities and remove no-template MIP contamination during the clean-up step.

Optimizing molecular counting of unique barcodes

We designed 13 probes of the same P152L_A154T design, but different N lengths from 5-0 Ns (5 Ns on arm, O Ns on the other) to 6-6 Ns. The goal of titrating the number of Ns was to find the length with optimal statistical power for barcode/molecule counting. The bioanalyzer traces of these reactions looked relatively similar in ESR peak size. The 6-2 N length molecular barcode sample was sent for Illumina sequencing with the NextSeq 300 cycle Medium Output kit. From these results I could see that the allele frequency was
incorrectly about 5% instead of the expected 10 %(Table 1B). Additionally, the maximum number of possible barcodes was reached like in previous experiments.

I repeated the above experiment with the probes of barcode sizes 6-2 Ns to 6-6 Ns and sent those for sequencing using the MiSeq 300 cycle V2 kit. Sequence analysis of these samples is currently in progress.

Scaling up MIP assays for larger reactions

I designed our most complete set of MIPs which included our original variants, several new ones, and several reference genes (Appendix, Table 3). Some variant sites are close enough to be captured with the same MIP. I performed a MIP reaction with this full set on the single genotype NA19240 (Figure 10). I hoped to see if any particular MIPs inhibited the function of others or performed at different activity levels than others. This experiment was done concurrently with the most recent barcode titration and sequenced in the same MiSeq lane. Sequencing analysis of these samples is currently in progress.

Figure 10: Bioanalyzer trace of MIP reaction pool using our largest set of variant and reference probes. No-template MIPs and other non-specific sequences smaller than 300 bp have been removed (blue) and the largest peak is in close to the range of 320 bp (red).
Conclusion

The NOTCH2NL project aims to understand the developmental role of human-specific NOTCH2NL genes from both an evolutionary and biomedical perspective. While we know about NOTCH2NL’s functional role in regulating neurogenesis and that copy number variation (CNV) may impact neurogenesis, we know much less about how the diversity of NOTCH2NL paratypes impacts the functional role. A possible theory suggests certain paratypes may affect the penetration of 1q21.1 distal duplication/deletion syndrome. To investigate theories like this we need more sequencing data of NOTCH2NL genes than available. Sequencing data from both the “typical” population and those with atypical brain development, specifically phenotypes associated with 1q21.1 distal duplication/deletion syndrome, are critical to performing a complete analysis. Our approach for sequencing variable regions of NOTCH2NL has been to develop molecular inversion probe assays (MIPs). MIPs as a suitable assay for genotyping because of the capture method that allows us to selectively sequence regions we are interested in.

MIP assays successfully catch NOTCH2NL variant regions. To accurately assess the frequency of our paratypes we have been optimizing our assays. We tested the efficiency of our MIPs to capture variant regions, and found a large amount of no-template MIPs present in our reactions at both the quality control stage and in our sequencing data. No-template MIPs skew depth of sequencing and our allele frequencies. Though we believed these no-template MIPs were a product of inefficiency, the capture performance stayed consistent and continued to use all available unique barcodes. This caused us to reevaluate our efficiency problem and conclude it was likely the opposite, as we’re creating more molecules than the molecular barcodes can keep up with. Instead of trying to eliminate the no-template MIPs during the reaction we decided going forward to alter the clean up to more stringently remove all sequences smaller than the ESR.

Data analysis is in progress on our full set of probes. That data will be used to see if there is a difference in MIP performance among different designs. We don’t expect to see issues with design performance of this particular set, however if this does arise, we have pre-emptively done pairwise probes reactions that are ready for sequencing. We would like to use the pairwise probe data to identify the problematic probes.

Fixing our molecular barcode problem is the final step to assay validation. While data
analysis of our recent barcode length titration is also currently in progress, we believe the optimal length will be between 6-3 Ns and 6-6 Ns. As soon as we confirm our optimal barcode length, we would like to do another full probe set reaction. Presumably then we will be able to accurately calculate paratype frequency within our known cell lines. If we can accurately recreate these frequencies we will be able to use our assays as a diagnostic tool on patient DNA from the Simons Foundation Autism Research Initiative (SFARI), and possibly a diagnostic tool on other patients with suspected duplication/deletion events.

The data that will come from those samples will be used in further analysis to find whether or not there is a distinct correlation between NOTCH2NL genotype and 1q21.1 distal duplication/deletion syndrome phenotype.
Bibliography


Appendix

Oligo sequences

A.

mipgen -regions_to_scan /path/to/bed/file.bed -project_name name_of_project -min_capture_size 120 -max_capture_size 250 -bwa_genome_index /path/to/reference_genome.fa

B.

mipgen -masked_arm_threshold 0 -ext_min_length 23 -tag_sizes N,N -regions_to_scan /path/to/bed/file.bed -bwa_genome_index /path/to/reference/genome.fa -project_name name_of_project -lig_min_length 23 -max_capture_size 300 -min_capture_size 200 -snp_file /path/to/VCF/file.vcf -arm_length_sums 46,47,48,49,50,51,52,53,54,55 -seal_both_strands on -max_mip_overlap 0

Table 1: (A) The command used for the very first MIP designs. (B) The most updated command used to design MIPs. Improved primarily to changed N length of molecular barcodes (-tag_sizes) and avoid designing MIPs with our variants in targeting arms (-snp_file).

<table>
<thead>
<tr>
<th>MIP name</th>
<th>size</th>
<th>5’-target arm/barcode/sequence/barcode/target arm-3’</th>
</tr>
</thead>
<tbody>
<tr>
<td>MII2bpdel</td>
<td>313</td>
<td>GCATGGTGAGTATCGGGCTGAGGCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNGGAGGAGAGAG</td>
</tr>
<tr>
<td>R113*</td>
<td>249</td>
<td>GCTTTGTGTCCTCGACCCTGCCTGAATGGCCTTCAGCTTCCCAGATATCCGACGGTAGTGTNNNNNTGGTGGGACTTGTGTG</td>
</tr>
<tr>
<td>A154T</td>
<td>314</td>
<td>GCTTCACAGGGCCAGAAATGTGAGACTGCTTCAGCTTCCCGAGATATCCGACGGTAGTGNNNNTGGTGGGACTTGTGTG</td>
</tr>
<tr>
<td>T197I</td>
<td>369</td>
<td>GCTTCACAGGCCAGTACTGTGACAGACTTTCAGCTTCCCGATATCCGACGGTAGTGNNNNTGGTGGGACTTGTGTG</td>
</tr>
</tbody>
</table>

Table 2: Four original MIPs designed for NOTCH2NL paratypes.
<table>
<thead>
<tr>
<th>MIP variant name</th>
<th>size</th>
<th>5'-target/barcode/backbone/barcode/target-3'</th>
</tr>
</thead>
<tbody>
<tr>
<td>M1I_2bpdel_C19W_T21A</td>
<td>322</td>
<td>GGGCTGCCACCTGGGGCGACCCTNNCTTCAGC</td>
</tr>
<tr>
<td></td>
<td></td>
<td>TTCCCCGATATCCGACGGTAGTGTGNNNNNCCAT</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GTGGATCTGCCCAGGGCGGC</td>
</tr>
<tr>
<td>R65Q</td>
<td>322</td>
<td>GTGTGCGCCAGCCATGCTGGGAAAAGNNCTTT</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CAGCTTCCCCGATATCCGACGGTAGTGTGNNNNN</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GTGCTTTCTTGTAGTTGGACGTGTATGTG</td>
</tr>
<tr>
<td>S106P_R113Q_R113*</td>
<td>322</td>
<td>CCTACCTTCAGCAGATACCTTTTATTTAGCNNTTT</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CAGCTTCCCCGATATCCGACGGTAGTGTGNNNNN</td>
</tr>
<tr>
<td></td>
<td></td>
<td>TGTGCTTCAGGGTTTACAGGAGG</td>
</tr>
<tr>
<td>P152L_A154T</td>
<td>322</td>
<td>CATCCTCAATACCTCATGACATCGAGANNTTT</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CAGCTTCCCCGATATCCGACGGTAGTGTGNNNNN</td>
</tr>
<tr>
<td></td>
<td></td>
<td>TCTGCCTCAGGGTTTACAGGAG</td>
</tr>
<tr>
<td>T197I</td>
<td>322</td>
<td>GCAGACCTGGTGACTTACCTTTTGAGTGNNNCTTC</td>
</tr>
<tr>
<td></td>
<td></td>
<td>AGCTTCCCCGATATCCGACGGTAGTGTGNNNNN</td>
</tr>
<tr>
<td></td>
<td></td>
<td>AGACTGATGTCAATGAGTGAC</td>
</tr>
<tr>
<td>NOTCH_ENDING</td>
<td>318</td>
<td>ACTGAATTTGACATTAAACAGTAGTGATCNNNTTT</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CAGCTTCCCCGATATCCGACGGTAGTGTGNNNNN</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GGGTTATGATGATGAGTGCGGTAG</td>
</tr>
<tr>
<td>PMS1-E2</td>
<td>322</td>
<td>GGGTCTGGACACATTTTTCTCTTTNNTCTCA</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GCTTTCCCCGATATCCGACGGTAGTGTGNNNNN</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ATGAAACATTTTGCCCGCGCAACAGTT</td>
</tr>
<tr>
<td>BTBD8-E6</td>
<td>272</td>
<td>GCTCTCCTTTTCACTGGACCTTCTCCTGAGGCNNTTT</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CAGCTTCCCCGATATCCGACGGTAGTGTGNNNNN</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CATCAGGGTCCTTATCACAACCTT</td>
</tr>
<tr>
<td>PLXNA2-E18</td>
<td>257</td>
<td>GCCAGGGCAGGTGAGGGTGTTGTGTTNNTCTTC</td>
</tr>
<tr>
<td></td>
<td></td>
<td>AGCTTCCCCGATATCCGACGGTAGTGTGNNNNN</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GTGTCGTTGTAAAATTGCAAGG</td>
</tr>
<tr>
<td>FOXJ2-E9</td>
<td>322</td>
<td>GCACGTGACCGTCTGAAGTGAAGNNCTTCAGC</td>
</tr>
<tr>
<td></td>
<td></td>
<td>TTCCCCGATATCCGACGGTAGTGTGNNNNNNTCTC</td>
</tr>
<tr>
<td></td>
<td></td>
<td>AGTCTTTGCTTGCCTTTCTCCT</td>
</tr>
</tbody>
</table>

Table 3: Full current set of MIPs with 6-2 Nmer barcode. For barcode titration experiments these designs were used with varying barcode lengths. The P152L_A154T design was solely used for most optimization experiments.
<table>
<thead>
<tr>
<th>PCR primer</th>
<th>5'- adaptor/backbone-3'</th>
</tr>
</thead>
<tbody>
<tr>
<td>universal_for_MIP_PCR</td>
<td>AATGATACGGCGACCACCGAGATCTACACATACGAGAT CCGTAATCGGGAAGCTGAAG</td>
</tr>
<tr>
<td>PCR primer</td>
<td>5'- adaptor/sample ID/backbone-3'</td>
</tr>
<tr>
<td>1_rev_MIP_PCR</td>
<td>CAAGCAGAAGACGGCATA CGAGATcttatcgtACACG CACGATCCGACGGTAGTGT</td>
</tr>
<tr>
<td>2_rev_MIP_PCR</td>
<td>CAAGCAGAAGACCGCATA CGAGATggccagagACACG CACGATCCGACGGTAGTGT</td>
</tr>
<tr>
<td>3_rev_MIP_PCR</td>
<td>CAAGCAGAAGACGGCATA CGAGATttacttccACACG CACGATCCGACGGTAGTGT</td>
</tr>
<tr>
<td>4_rev_MIP_PCR</td>
<td>CAAGCAGAAGACGGCATA CGAGATcggcccttACACG CACGATCCGACGGTAGTGT</td>
</tr>
<tr>
<td>Sequencing primer</td>
<td>5'- backbone-3'</td>
</tr>
<tr>
<td>fwd_MIP_SEQ</td>
<td>CATACGAGATCCGTAATCGGGAAGCTGAAG</td>
</tr>
<tr>
<td>rev_MIP_SEQ</td>
<td>ACACGCACGATCCGACGGTAGTGT</td>
</tr>
<tr>
<td>index_MIP_SEQ</td>
<td>ACACGCACGATCCGACGGTAGTGT</td>
</tr>
</tbody>
</table>

Table 4: Sequences of MIP PCR and sequencing primers.

**Materials**

1. Reagents
   - 10x T4 DNA ligase reaction buffer, *NEB*
   - T4 polynucleotide kinase, *NEB*
   - Hemo KlenTaq DNA polymerase, *NEB*
   - Exonuclease I, *NEB*
   - Exonuclease III, *NEB*
   - 10x Ampligase reaction buffer, *Lucigen*
   - Ampligase DNA ligase, *Lucigen*
   - 2x iProof Hi-Fidelity Master Mix, *Bio-Rad*
   - SPRIselect reagent, *Beckman Coulter*
   - Buffer EB elution buffer, *Qiagen*
   - Quibit dsDNA HS assay kit, *Thermofisher*
   - Bioanalyzer High Sensitivity DNA kit, *Agilent*
- Agarose for molecular biology, *Sigma Aldrich*
- DNA ladder
- DNA loading dye
- Ethidium bromide solution
- 1X TBE buffer
- 200 proof ethanol
- .25nM dNTPs
- Nuclease-free water