Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

_____ _____

Karyn Meltz Steinberg Date

Identifying Variants in Neuroligin Pathway Genes Using Next Generation Sequencing Technologies

By

Karyn Meltz Steinberg
Doctor of Philosophy

Graduate Division of Biological and Biomedical Sciences
Program of Population Biology, Ecology and Evolution

_____
Michael E. Zwick, Ph.D.
Advisor

_____
David J. Cutler, Ph.D.
Committee Member

_____
Stephanie L. Sherman, Ph.D.
Committee Member

_____
Yun Tao, Ph.D.
Committee Member

_____
James W. Thomas, Ph.D.
Committee Member

Accepted:

_____
Lisa A. Tedesco, Ph.D.
Dean of the James T. Laney School of Graduate Studies

_____
Date

Identifying Variants in Neuroligin Pathway Genes Using Next Generation Sequencing Technologies

By

Karyn Meltz Steinberg
B.A., Northwestern University, 2001

Advisor: Michael E. Zwick, Ph.D.

An Abstract of
a dissertation submitted to the Faculty of the James T. Laney School of Graduate Studies
of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy

Graduate Division of Biological and Biomedical Sciences
Program of Population Biology, Ecology and Evolution

2009

Abstract

Identifying Variants in Neuroligin Pathway Genes Using Next Generation Sequencing Technologies

By Karyn Meltz Steinberg

The fields of population genomics and evolutionary quantitative genetics provide a framework in which one can best pursue the heritable component of complex human phenotypes. A central challenge lies in the comprehensive ascertainment of all the relevant genomic variation, irrespective of their population frequency, in a large collection of human samples. Autism Spectrum Disorder (ASD) is a complex human neurodevelopmental disorder, characterized by a high heritability and a nearly 4:1 male excess. Applying this comprehensive genomic variation detection paradigm poses two main challenges. The first lies in developing and applying technologies that can efficiently detect the relevant genomic variation. The second lies in applying these methods in the context of a testable genetic hypothesis that might elucidate the etiology of ASD. Here I report on a series of studies that have pursued both of these challenges. While sequencing technologies have advanced rapidly in the past decade, the ability to rapidly isolated target DNA for sequencing has lagged. I first describe a novel technique for isolating target DNA for downstream resequencing applications. This protocol, named Microarray-based Genomic Selection, is able to efficiently select user-defined sequence that is then hybridized to resequencing arrays. Two experiments that used Microarray-based Genomic Selection for resequencing are described. In the first experiment the technology was used to isolate all of the exons on the X chromosome, while the second experiment used it to isolate specific genes in the neuroligin pathway that are hypothesized to contribute to ASD. Advantages and limitations of MGS are discussed. To address the genetic basis of ASD, I first selected X-linked neuroligin pathway genes thought to harbor ASD susceptibility alleles that may help explain the male excess in ASD. Using samples of male individuals with ASD obtained from the Autism Genetic Resource Exchange (AGRE), I performed paired-end multiplexed sequencing on the Illumina Genome Analyzer to comprehensively sequence the genomic regions containing the neuroligin pathway genes. This study identified a series of candidate variants that may contribute to ASD susceptibility. Finally, I will highlight the importance of using quantitative evolutionary genetics when analyzing and interpreting sequence data.

Identifying Variants in Neuroligin Pathway Genes Using Next Generation Sequencing Technologies

By

Karyn Meltz Steinberg
B.A., Northwestern University, 2001

Advisor: Michael E. Zwick, Ph.D.

A dissertation submitted to the Faculty of the James T. Laney School of Graduate Studies
of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy

Graduate Division of Biological and Biomedical Sciences
Program of Population Biology, Ecology and Evolution

2009

ACKNOWLEDGEMENTS

I would first like to thank my advisor, Michael Zwick, who has supported me throughout this entire process. He has always been there to answer a question or guide me through an experiment. Mike has treated me with respect and truly been my mentor.

I would also like to thank the members of the Zwick Lab (Team Zwick) for assisting me with experiments or analysis.

Finally, I would like to acknowledge my family: my parents, Susan and Gary Meltz, my husband, Scott Steinberg and my daughter, Zoe. They have been there for me when times were good and when times were not so good. They were always supportive and never judgmental, and I could not have finished this dissertation without them.

TABLES

FIGURES

INTRODUCTION

QUANTITATIVE GENETICS AND COMPLEX HUMAN DISEASE

A central challenge of contemporary human genetics lies in identifying and characterizing the genetic variation that underlies the heritable component of complex human phenotypes. The tools and concepts of population genomics that aim to understand the forces that influence allele frequencies provides a framework in which this research program can be pursued. Human geneticists have been highly successful in identifying the genetic basis of Mendelian disorders that are caused by very rare mutations in one or a few genes. These highly penetrant, rare alleles can be identified through traditional family-based linkage analysis and with the completion of the human genome project and the broad availability of genomics technologies, these approaches have become routine.

In contrast, human geneticists have faced considerably more difficulty mapping genes that contribute to heritable complex diseases [1]. Limitations in both the cost and throughput of existing DNA sequencing technologies, combined with the potential of high-throughput genotyping at sites known to vary in the human genome, let to the development of the whole genome association studies paradigm. These studies were only powered to detect common variants that might contribute to common disorders. Large scale whole genome association studies, such as the HapMap Project [2] and the Wellcome Trust Case Control Consortium, identified a few common alleles that are associated with diseases such as Crohn's Disease [3] and hypertension [4]. Yet, these alleles contribute only a small fraction of the total variance observed in these traits. Additionally, researchers have failed to find any genome-

wide statistically significant common alleles of large effect for many complex disorders, such as bipolar disorder. In total, these findings imply that there are no common alleles of ***large effect*** underlying complex human disease traits. This is an important although perhaps underappreciated finding about the genetic structure of these complex human disorders. Yet at the same time, these studies do suffer from limitations inherent in their design. This is in part due to the fact that methods like whole genome association lack the power (and the ability to experimentally) detect multiple, low frequency alleles with moderate phenotypic effect [5]. Clinical heterogeneity and poor phenotyping pose further challenges in pinpointing the genetic contribution to complex traits.

Evolutionary Quantitative Genetics

Evolutionary quantitative genetics provides a statistical framework that may provide insight and help us identify the causes of these complex human traits. The phenotypic distribution of a quantitative trait is a function of both the genotypes present in a population and the effects of the environment that are usually difficult to measure or observe over long periods of time. Barton and Turelli provide a summary of the challenge we face:

> "The fundamental difficulty is that the phenotypic distribution of the few characters we observe depend on genotype frequencies over many loci, a problem analogous to that in thermodynamics, where bulk properties reflect the hidden motions of many molecules…[and] while the immediate response to selection is predictable, long-term changes depend on genetic parameters about which we know very little" (Barton and Turelli [6], pp. 347-8).

Much as in thermodynamics, the hope is that the application of a statistical model can provide insights without requiring specific measurements of every genotypic and environmental effect in the given population. The genetic contribution to a complex quantitative trait can be estimated as the sum of all of the effects from segregating loci, which approximates a Gaussian distribution [7]. Most quantitative genetic variation can be statistically attributed to the additive component of variation although the effects of epistasis and/or dominance can also contribute to genetic variation. The process of mutation introduces new alleles to a population, and much of this variation is selectively neutral [8,9]; however, very deleterious alleles are quickly removed from the population while positive selection can potentially fix beneficial alleles. The distribution of mean fitness of a phenotypic trait is Gaussian, and selection acts to shift the mean to the optimal fitness (Figure I.1).

Understanding longer-term trends, as indicated by Turelli and Barton, is the real challenge and can be approached with different types of models. The first type of model is named *mutation-selection balance*. In this model, at equilibrium, genetic variation eliminated by selection is balanced by the variation introduced by mutation. If variation in a quantitative trait is maintained by mutation-selection balance, the distribution of allelic effects will be dominated by primarily rare alleles, each with extremely small effect [10,11]. This is paradoxical because there is ample empirical evidence of polygenic variation as well as evidence of stabilizing selection, which should eliminate this variation. One explanation is that many genes affect many traits, and an individual trait is affected by many genes. Pleiotropically related traits are concomitantly acted upon by stabilizing selection, and variation in the trait in

question is simply a by-product of polymorphisms that are maintained for completely different reasons. This could account for both common and rare variation that is associated with quantitative characters.

Alternatively, one may consider the role of fluctuating environments on the underlying patterns and levels of genetic variation. Gillespie modeled the effects of temporal and spatial fluctuations on the phenotypic distribution of quantitative characters. His simulations suggest that alleles that have different additive effects in different environments as heterozygotes will have lower phenotypic variance than homozygotes who will have lower mean fitness than heterozygotes [12]. Therefore strong fluctuating natural selection could act to maintain variation in quantitative traits. Unfortunately, this hypothesis is difficult to test, and, in general, accounting for the effects of environmental stochasticity over long periods of time remains a challenge.

Most quantitative characters are polygenic; however, the actual number of loci involved in any given trait is still largely unknown. Sewell Wright's work on guinea pig coat coloring led to the hypothesis that complex traits are controlled by many independent sets of loci acting together additively. He demonstrated that alleles with small effects could impact changes in phenotype if there were large enough sets of loci [13]. On the other hand, early studies on *Drosophila* bristle number were suggestive of a relatively small number of loci with large effects that contributed to the trait [14]. The number of loci and the magnitude of allelic effects are of great importance when examining complex human disease. The underlying hypothesis of

the number of alleles and the distribution of allelic effects guides the experimental design and subsequent analysis.

Under this framework, we can treat complex human diseases as quantitative traits. Yet, disease incidence data are discrete variables (affected and normal), and quantitative genetics models are based upon continuous variables. Falconer [15] developed theory to analyze heritability of disease susceptibility for polygenic, multifactorial diseases. An individual's liability is the likelihood of developing disease based upon the genetic contribution to disease susceptibility as well as external circumstances that affect disease susceptibility. Liability is a graded scale, and the threshold is the point above that all individuals are affected and below which all individuals are normal (Figure I.2).

The field of quantitative genetics provides models to answer questions related to how many loci contribute to complex traits, how selection affects the distribution of alleles related to complex traits and the role of environmental changes in shaping complex traits. With Falconer's work, disease data can be examined using a quantitative genetics framework. This provides a foundation for the current genetic models of complex human diseases.

Common Disease Common Variant (CDCV) Hypothesis

The CDCV hypothesis predicts that most disease susceptibility variants are relatively common, few in number and have modest phenotypic effects [16,17]. Common ancestral alleles that were previously positively selected may contribute to disease susceptibility due to changing environmental and selective pressures. For example,

the "thrifty gene" hypothesis proposes that the ability to store fat was advantageous to hunter gatherer populations who underwent periods of feast interspersed with famine [18]. However, with the adoption of a high fat/high carbohydrate diet and sedentary lifestyle this trait contributes to the development of Type II Diabetes. Empirical data suggest that the common, ancestral alleles in genes that are associated with Type II Diabetes contribute to disease susceptibility while the derived, rarer alleles are protective and show the signature of recent positive selection (reviewed in [19]).

In addition to selective forces, demographic processes have shaped the current allele frequency spectrum over many generations. The human lineage has most certainly undergone major bottlenecks and expansions. Data suggest that the European/Caucasian population has experienced bottlenecking and founder effects, while the Sub-Saharan African population expanded rapidly [19]. These demographic changes shape the genomic architecture of both protein coding regions as well as non-coding elements [20].

In a single disease locus model, Reich and Lander [21] modeled the allelic spectrum of common disease and demonstrated that with a high equilibrium disease allele frequency (approximately 0.2 for a common disease) and a mutation rate of 3.2×10^{-6} per site, per generation, the modern allelic spectra is relatively simple even when taking into account the rapid expansion of the human lineage approximately 100,000 years ago. They explain the modern allelic spectra quantitatively by accounting for the "half-life of the ancestral allelic spectrum," or in other words, how quickly ancestral alleles are replaced by modern alleles in the approximately 3000 generations since the human population expansion. Because their model assumes that the

population is in equilibrium state both rare and common diseases have a diverse allelic spectra, but rare diseases reach this equilibrium faster than common diseases based on their "half life." However, this model does not account for the effects of genetic drift, population substructure or changing selection pressures. It also assumes that all of the disease alleles are selectively equivalent which is unlikely for a complex trait attributed to many genetic loci. Additionally the assumption that the population is in an equilibrium state is not likely the actual state of the human population given such a drastic demographic change in recent evolutionary history [19]. Despite these shortcomings, the Lander-Reich model provided one rationale for studies such as the HapMap Project [2] as well as Genome Wide Association Studies (GWAS).

Using a selection model stochastically based on individual genotype, Peng and Kimmel test the Reich/Lander model with forward time simulations. Their model assumes an infinite allele model [22] and that the fitness of an individual is based upon the fitness at all disease susceptibility alleles. Their simulation results support the common disease common variant hypothesis. One caveat is that they simulated mostly rare diseases with allele frequencies close to or at equilibrium for the Reich/Lander model. Additionally their model does not hold up under a model of polygenic disease caused by rare alleles at numerous loci that can each cause disease singlehandedly (see Smith and Lusis' discussion of genetic heterogeneity [23]).

Although these simulations provide support for the CDCV hypothesis, the empirical evidence is contradictory. Large case-control GWA studies focusing on SNPs with minor allele frequencies (MAF) greater than 5% have identified a few

common variants associated with common diseases such as Type I Diabetes [3] . Yet, these common variants account for only about 5% of the heritability of these diseases. That leaves the majority of genetic variability unexplained by common variation. Thus one reasonable conclusion is that rare variation must be significantly contributing to heritability for complex human disease traits.

Common Disease Rare Variant (CDRV) Hypothesis

The alternative hypothesis, the CDRV hypothesis, predicts that disease susceptibility is a result of multiple, rare alleles with moderate to large phenotypic effect [5,24]. Non-synonymous sequence variation may be individually rare, but, as a class, affected individuals may have variants that are at the same locus. Pritchard [25] modeled complex disease using a coalescent simulation, assuming a multilocus model with a constant population size where the current population is at equilibrium. Susceptibility alleles are in mutation-selection balance, and selection acts independently at each disease susceptibility locus. The model predicts that the majority of genetic variance can be attributed to loci with high overall mutation rates, but at loci where the mutation rate is low common variation can explain susceptibility to disease. This model also generates numerous mildly deleterious alleles that rarely reach high frequencies in a population, which are acted upon by weak purifying selection. Like Reich and Lander, Pritchard assumes that the human population is at equilibrium, which is unlikely. Yet, the empirical evidence is overwhelmingly in support of the CDRV hypothesis.

If the central challenges in human genetics are to be addressed, the comprehensive detection of all classes of genomic variation is seemingly the optimal method to pursue. Ideally, we would sequence the entire human genome of large numbers of individuals. While sequencing technologies have experienced a dramatic improvement in accuracy and cost, this experiment is still not practical for large populations. Direct sequencing of portions of genomes, on the other hand, is becoming technically feasible. Like all sequencing based approaches, in principle, it can easily identify rare as well as common variants. While the statistical power to test any single rare variant is low [26], grouping of rare variants into classes can provide sufficient numbers to assess their role in complex disease traits [27].

Systems biology assumes that most genes will function within complex networks, and mutations in strongly connected genes may lead to the same or similar phenotypes that potentially result in genetically heterogeneous syndromes [28]. One plausible hypothesis supposes that genes within a particular pathway work together as a single biological module and that genes controlling predisposition to a human trait may be primarily involved in the physiological pathway that regulates that trait. For example, Henneh et al [29] found that distinct allelic haplotypes for the DISC1 gene and its binding partner NDE1 were over-transmitted in females with schizophrenia. Additionally, Lesnick et al [30] demonstrated the contribution of axon guidance pathway genes in Parkinson Disease. These authors concluded that although mutations in single genes within this pathway would show only slight phenotypic effects, the combined effects of mutated genes could explain severe phenotypes of complex diseases. The genomic pathway approach, where a set of candidate genes is

chosen from various linkage analysis studies or gene-gene interaction data sets, represents a major paradigm shift away from traditional candidate gene studies [30]. Investigating multiple genes within a physiological pathway may elucidate the underlying genetic architecture of complex traits such as autism spectrum disorders.

AUTISM AS A COMPLEX TRAIT

Autistic Disorder [OMIM 209850] is a pervasive developmental disorder (PDD) characterized by a lack of reciprocal social relations, poor verbal communication skills, and repetitive behaviors presenting clinically within the first three years of life. PDDs include Autistic Disorder, Rett Syndrome, Childhood Disintegrative Disorder, Asperger's Disorder, and PDD-NOS (not otherwise specified). The term Autism Spectrum Disorder (ASD) is used to reflect the phenotypic and genotypic heterogeneity observed among PDDs (with the exception of Rett Syndrome, which is known to arise as a consequence of mutations at the MECP2 locus).

Patterns of Inheritance

ASD is highly heritable; using a broad definition of ASD, concordance rates among monozygotic twins are estimated as 91% and 10% for dizygotic twins [31-33]. Additionally, sibling relative risk has been estimated to be approximately 4.5% [34]. The distribution of allelic sharing for markers from a genome wide screen was most consistent with a model of multigenic inheritance of at least 15 susceptibility loci [35]. Additionally, results from a complex segregation analysis of ASD were inconsistent

with the major-locus inheritance model, and the multifactorial threshold model was proposed to explain the inheritance patterns seen among ASD families [34]. The model makes three predictions about recurrence risks: (1) among second and third degree relatives they will decline rapidly, (2) they increase with multiple affected offspring, and (3) they are lower for relatives of the more frequently affected sex. For example, in ASD the recurrence risk for siblings of affected males is 3.7% while the recurrence risk for siblings of affected females is 7.0% [34]. These statistics suggests that females with ASD have more severe disease phenotypes. However, both the prevalence of disease and frequencies of susceptibility variants are expected to be extremely rare among females. On the other hand, the prevalence of males with ASD is relatively common and males are expected to be highly informative in genetic studies of ASD. These studies highlight the considerable genetic heterogeneity of ASD. One plausible hypothesis predicts that that ASD may be caused by many, possibly rare susceptibility variants that are widely distributed throughout the genome.

The X Chromosome and Cognitive Disorders

There is a significant sex bias towards males for mental retardation and other cognitive disorders, such as ASD. For example, in the general population the prevalence of X-linked MR is 2.6 cases per 1000 which accounts for more than 10% of all cases of MR [36]. The hemizygosity of males for virtually all X chromosome genes exposes recessive phenotypes. Additionally, more than 500 of the genes located on the X chromosome are expressed in human brain [36], and the X chromosome appears to be enhanced for "cognition genes" [37]. Zechner et al [38] has

proposed that for the past 300 million years the X chromosome has played a role in the development of sexually selected traits and natural selection has favored X-linked genes that are related to higher cognitive abilities. The fourfold excess of males affected with ASD compared to females is consistent with a model of X linkage.

The majority of the human male X chromosome does not recombine during meiosis; only small portions at the distal ends of each arm, the pseudoautosomal regions, recombine with homologous loci on the Y chromosome. Gene content, predicted transcription of exons and the frequency of CpG islands are significantly lower than expected given the length of the X chromosome [39]. Although the human X chromosome contains only 4% of all human genes, approximately 10% of all disorders with Mendelian inheritance are due to mutations in genes on the X chromosome [39]. There are at least 16 X-linked genes associated with mental retardation (MR), and mutations in genes that cause or contribute to MR could be targets of selection for human cognitive abilities [40]. Comparative genomics provides an ideal platform to examine the questions of human lineage specific selection. Genes that have undergone recent human specific positive selection may shed light on human specific biological processes and specializations. The human X chromosome, overall, shows an excess of positively selected genes when compared to our closest great ape relatives [41]. Other data suggest that genes related to brain function did evolve under positive selection [42].

A study of 10 X chromosome genes that are associated with mental retardation found that nucleotide diversity in these genes was lower than in chimpanzees, which could signal positive selection, however, the ratio of

nonsynonymous to synonymous substitutions indicated a selective constraint on these genes in the great ape lineage [40]. This ratio is also significantly lower in brain-specific genes when compared to tissue specific genes suggesting strong purifying selection [43]. Additionally, genes with maximal expression in the brain are highly conserved, most likely due to functional constraints of the nervous system, and are less likely to show any marks of positive selection [41]. It is likely that cognitive differences between humans and apes may be due to small changes in gene expression and regulation [44]. Indeed, genes that are differentially expressed in human brain tissue, when compared to chimpanzee brain tissue, are upregulated in the human lineage [45,46]. Non-coding regions, specifically non-coding DNA upstream of genes that may control expression and regulation, show higher rates of nucleotide divergence between humans and chimpanzees than coding and downstream non-coding regions [47] suggesting that regulatory regions have diversified in the human lineage. Resequencing candidate genes on the X chromosome may expose the underlying genetic architecture of ASD.

SEQUENCING TECHNOLOGY DEVELOPMENT

During the past decade, large industrial genome sequencing centers, using Sanger sequencing chemistries, have been able to automate these steps and increase throughput 50-fold while at the same time reducing costs 100-fold [48]. This technical achievement has been remarkable. Yet today, we stand on the cusp of a revolution in DNA sequencing. Novel DNA sequencing chemistries that offer drastic cost reductions, increased data production and high accuracy are now available in single

instruments that require far fewer people and less laboratory space to operate [49]. Collectively, these recent innovations are beginning to raise the question as to whether the traditional industrial genome sequencing model is reaching the end of its utility. We can think of sequencing technology in terms of their "generations." Traditional Sanger sequencing was the first generation of technology, while second generation technology included advances in sequencing by hybridization (resequencing arrays), pyrosequencing and other massively parallel sequencing by synthesis methods. The third generation of technology expands on the second generation with single molecule sequencing. Here I review all three of these generations.

First Generation Sequencing Technology

In traditional Sanger sequencing the genome is fragmented and clonal libraries are produced to isolate and amplify single fragments. Determining the fragment sequence with Sanger sequencing, also known as dideoxy sequencing, involves synthesis of a complementary DNA template using deoxynucleotides (dNTPs) and termination of synthesis using unnatural 2',3'-dideoxynucleotides, ddNTPs, by DNA polymerase [50]. Fragments are produced and separated by gel electrophoresis for analysis. This process has been automated by tagging each ddNTP with a different fluorescent dye. As labeled fragments pass through the DNA sequencer the dye is excited by a laser, and the resulting fluorescence emission of one of the four colors is used to determine basecalling and sequence assembly [51]. The integration of multiple capillary arrays per instrument has allowed the sequencing to proceed in parallel [52].

Automated data quality assessment and sequence assembly algorithms are then used to reconstruct the original genome sequence. For large projects that sequence millions to billions of bases, error rates must be low to minimize errors in the final sequence. The Bermuda standard, the community accepted quality level for finished genome sequencing, is equal to an error rate of less than 1 error per 10,000 bases sequenced (or 99.99% accurate). The accuracy of single reads is typically lower than this stringent requirement for Sanger sequencing. To achieve very high accuracy, multiple reads of the same base are necessary. Phred quality scores are calculated for each sequenced base to determine the probability of a basecalling error (phred = $-10 \log_{10}$ * (error probability) [53,54]). A phred score of 40 is accepted as the Bermuda standard. Achieving a quality score of 40 for an entire genome with this technology usually requires a random ten-fold, or 10X, coverage; this increases the costs involved in whole genome sequencing. Currently, the approximate cost for a complete draft sequence at 4X coverage (corresponding to a Phred score of 20—far below the Bermuda standard) is approximately US $ 0.008 per base pair [55]. While the industrial implementation of Sanger sequencing is still considered the "gold standard" of DNA sequencing, particularly for *de novo* genome sequencing and assembly, the relatively high cost of obtaining the final sequence suggests that this methodology will not prove sufficiently economical for routine whole genome resequencing [49]. Furthermore, the vast infrastructure requirements and costs to even establish a genome sequencing center in the first place, preclude the ability of individual laboratories to perform routine sequencing on this scale. Consequently, the need for

ever greater sequencing throughput is driving the development of more efficient sequencing technologies.

Second Generation Sequencing Technologies

Resequencing arrays (RAs) are a second generation sequencing-by-hybridization technology. Overlapping oligonucleotide probes that are typically 25 base pairs long are immobilized on an array and are tiled at a 1 base pair resolution. Each targeted base has 4 identical features for the forward strand and 4 features for the complementary strand. Features are 25 bases long with position 13 as the query base that contains either A, C, G, or T and contains approximately 1,000,000 copies of the particular oligonucleotide. Target DNA is fragmented, fluorescently labeled and then hybridized to the RA. The two features (one forward and one reverse) that are complementary to the test sequence will provide the brightest signal. If the sample DNA happens to be heterozygous at position 13, the two features with the appropriate complementary base will provide the highest signal.

Resequencing array data is analyzed using the ABACUS (Adaptive Background genotype Calling Scheme) algorithm implemented in RATools [56]. ABACUS is a fully automated statistical algorithm that determines individual base/genotype calls with high accuracy regardless of the nature of the site. The algorithm employs likelihood models for each of the possible base calls that are tested independently for the forward and reverse strands. A quality score is assigned based on the difference between the best fitting model and the second best fitting model for each genotype. In the initial application of Affymetrix RAs using ABACUS more

than 80% of genotypes were called with greater than 99.9999% accuracy. To improve upon the 80% call rate, additional software was developed to perform more accurate, automated grid alignment [57].

Pyrosequencing is a sequencing by synthesis technology where the sequence of the target DNA is determined through a series of four enzymatic reactions. In the first reaction, a single nucleotide is added to the end of the sequencing primer; the four nucleotides are added one at a time. If the complementary base is added, polymerase extends the primer; however, if a noncomplementary base is encountered the reaction pauses until the proper complementary base is added. In the second step inorganic pyrophosphate is released which acts as a substrate for the ATP which is then converted to light by luciferase in the third reaction. The light signal produced by luciferase indicates base incorporation and is detected with a photon detector and recorded on a pyrogram. The sequence is then inferred by reading the signals across the pyrogram. Finally, apyrase is used to remove the unincorporated nucleotides and ATP, and the process cycles through the next addition of nucleotides.

Pyrosequencing has been used for *de novo* sequencing, resequencing, genotyping and sequence determination of secondary DNA structures (reviewed in [58]). Pyrosequencing was introduced for whole genome sequencing by Roche454 on the FLX machine (http://www.454.com). The 454 method fragments the entire genome to 300 base pair long fragments that are then ligated to adapters and individually captured on Sepharose beads. The beads are then added to an oil emulsion and clonally amplified on the PicoTiterPlate that can amplify 300,000 templates of a single DNA molecule. As reagents and nucleotides are passed over the

plate the sequence can be inferred from the pyrogram. Read lengths average about 250 base pairs, which makes this technology ideal for bacterial and viral genome sequencing [59].

The ABI SOLiD technology (http://www.solid.appliedbiosystems.com) utilizes a DNA preparation pipeline similar to Roche454 with adapter-ligated fragments and emulsion PCR with magnetic beads. The beads are then loaded onto a slide with up to 8 chambers and primers hybridize to the P1 adapter sequence. ABI's technology differentiates itself from Roche454 and Illumina in the di-base probe chemistry; four fluorescently labeled di-base probes competitively ligate to the primer. Multiple competitive ligation reactions, detection and cleavage are performed. The extension product is removed and the template resets with a primer that is complementary to the second base position. The competitive ligation reactions, detection, cleavage, and primer reset are performed for five more rounds allowing for each base to be interrogated by two different primers and ligation reactions. According to ABI, the di-base probe chemistry which leads to two base encoding allows for a more accurate sequence in resequencing applications because base-calling errors can be distinguished from true polymorphisms or single base deletions.

The Illumina technology [60] (http://www.illumina.com) uses a single molecule bridge amplification step that takes place on a flow cell that contains 8 sealed channels. Multiple copies create clusters that contain approximately 1 million copies of the original fragment. The flow cell with the clusters are then placed in the Illumina Genome Analyzer which then flows all four nucleotides simultaneously;

each base incorporation is a unique event which is captured during an imaging step. After imaging, the base is chemically removed and the nucleotides are flowed across the cell for the next incorporation. Illumina now supports up to 75 base pair reads, paired end sequencing (75 base pair reads each direction), and multiplexed sequencing in which a unique 6 base tag is added to different samples allowing up to 12 samples per lane [61].

Harismendy et al [62] assessed Roche 454, Illumina GA and ABI SOLiD technologies with regards to SNP discovery. When compared to SNP calls from the Illumina Hap550 BeadChip, the genotype accuracy was 97.4%, 100%, and 99.7% for Roche454, Illumina GA, and ABI SOLiD, respectively. When SNP calling was compared to independent Sanger sequencing, variant calling accuracy was 95%, 100% and 96% for the three technologies above respectively. The false positive rates were approximately 2.5%, 6.3% and 7.3% while the false negative rates were 3.1%, 0% and 3% for the three technologies above respectively. These data indicate that although Illumina has a higher false positive rate, accurate variant detection is superior to the other technologies.

Third Generation Sequencing Technology

Second generation sequencing by synthesis technologies are limited in their sequencing costs, complexity of library preparation and use of PCR amplification [63]. To overcome these issues, single molecule sequencing, which does not rely upon PCR was introduced for whole genome sequencing. Briefly, Poly(dT)

oligonucleotides are randomly attached to a glass slide which capture single stranded Poly(dA)-tailed template DNAs. Sequencing cycles consist of adding polymerase and one labeled nucleotide at a time, rinsing, imaging and dye cleavage [63]. For Helicos BioScience technology (http://www.helicosbio.com), each molecule is independent allowing for asynchronicity in synthesis phases. This translates into low rates of misincorporations and accurate homopolymer calling. The average read length is between 23 and 27 bases with a 98% mutation detection rate [63]. In addition to Helicos' technology, Pacific Biosciences (http://www.pacificbiosciences.com) has developed a single molecule real time sequencing technology and nanopore sequencing technology is on the horizon [64,65].

SCOPE OF THE THESIS

Sequencing technologies have advanced rapidly in the past decade. As a consequence, it is reasonable to suppose that we will be able to sequence the human genome for $1000 in the very near future. However, for many applications such as diagnostic testing it may not be useful to sequence the entire genome. It may be sufficient to target a subset of genes that contribute to the disease in question and resequence those loci. Yet generating this target DNA efficiently and accurately is still a challenge in the field.

In Chapter 1, I describe a novel technique for isolating target DNA for downstream resequencing applications. This protocol, called Microarray-based Genomic Selection (MGS), is able to efficiently select user-defined sequence that is then hybridized to resequencing arrays (RAs). In Okou, Steinberg et al [66] we

demonstrate that variant detection is highly accurate and that enrichment using this technology is sufficient for identifying clinically important variation in the FMR1, FMR2 and FMRNB region on the X chromosome.

In Chapter 2, I describe an experiment that utilizes MGS to select all of the exons on the X chromosome for downstream resequencing on RAs (so called MGS/RA resequencing). In this experiment, MGS selected only a portion of targeted DNA, and I will explore some possible reasons for its failure.

In Chapter 3, I describe an experiment in which MGS was applied to select target DNA from X chromosome loci that have been associated with a complex trait, Autism Spectrum Disorder (ASD). Extremely high rates of genetic variation were observed that suggested that MGS/RA resequencing was generating too many false positive genotype calls. I will examine the reasons for its failure and the limitations of MGS as a technology.

In Chapter 4, I present the results from an experiment using the Illumina GAII to resequence genes in the neuroligin pathway that have previously been associated with ASD in a large clinical sample.

Finally, in the Conclusion, I discuss the future of detecting variants in genes in biological pathways and how population genetics must guide these experiments and analyses.

FIGURE LEGENDS

Figure I.1. The effect of stabilizing selection on trait distribution. Extreme phenotypes in the original population are selected against leading to an increase in the frequency of the mean phenotype.

Figure I.2. Distributions of the general population and of relatives of affected individuals. The threshold value of liability remains constant. In the general population, affected individuals have a liability above the threshold (top), while in the population of relatives of affected individuals the mean is shifted. The average liability among siblings of affected individuals is higher than the general population while the threshold is shifted demonstrating that affected sibs are at a higher risk of developing the disease [15].

Figure I.3. Resequencing arrays. Four forward strand and four reverse complement strand features are associated with every site. One feature is a 25-base oligonucleotide where the 13th base is the query base. Features are then divided into 56 equal pixels and scanned individually [56].

Figure I.4. Pyrosequencing Chemistry. In the Pyrosequencing enzyme reactions, when an added dNTP forms a base pair with the template DNA, it is incorporated via Klenow polymerase (-exo), which releases pyrophosphate (PPi). ATP Sulfurylase converts PPi into ATP. Luciferase production leads to a light signal, which is measured and then analyzed [58].

Figure I.5. Illumina Genome Analyzer Workflow (adapted from www.illumina.com)

a) Genomic DNA is fragmented and the ends are repaired to convert overhangs into blunt ends. Then dATP adds "A" bases to the blunt ends to allow for the ligation of adapters. After ligation, 300 base pair fragments are selected and amplified.
b) These fragments are then added to the flowcell and bridge amplification is performed. This creates millions of clusters with each cluster containing many copies of a single DNA fragment. The sequencing primer is then attached.
c) The flowcell is then placed in the Genome Analyzer. All four fluorescently labeled nucleotides are flowed simultaneously. After the first base is incorporated and read, the base is deblocked, and the process is repeated for the remaining bases in the read.

Figure I.1

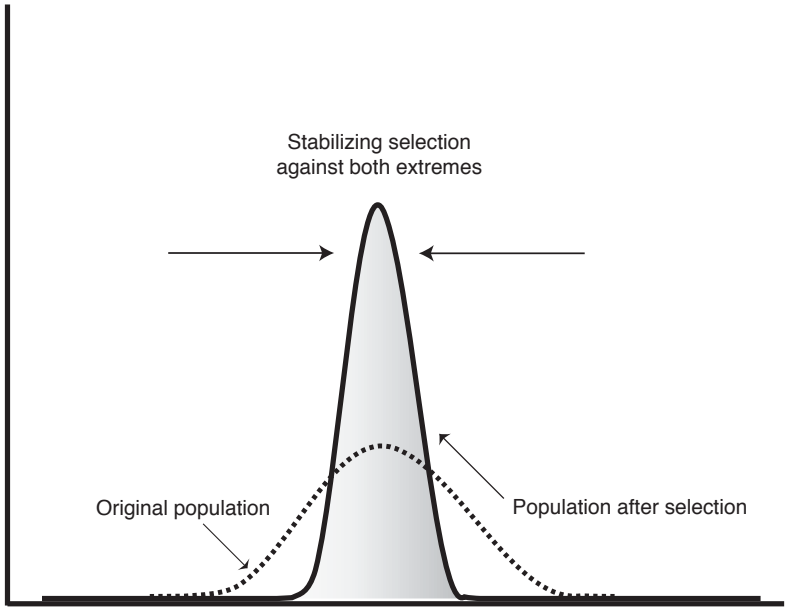

Stabilizing selection
against both extremes
Original population
Population after selection

Figure I.2

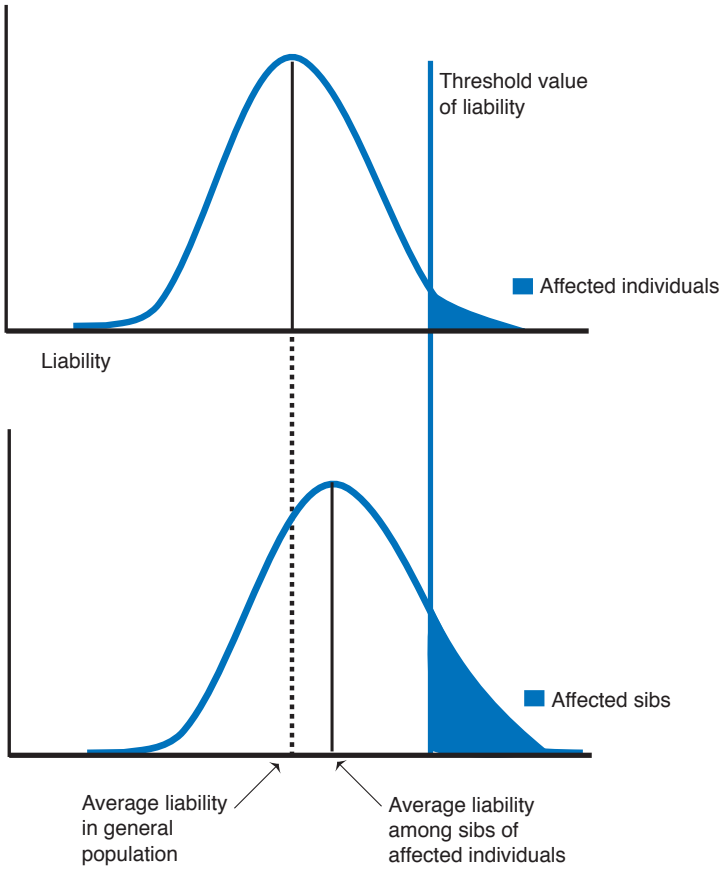

Threshold value
of liability
Affected individuals
Liability
Affected sibs
Average liability
in general
population
Average liability
among sibs of
affected individuals

Figure I.3

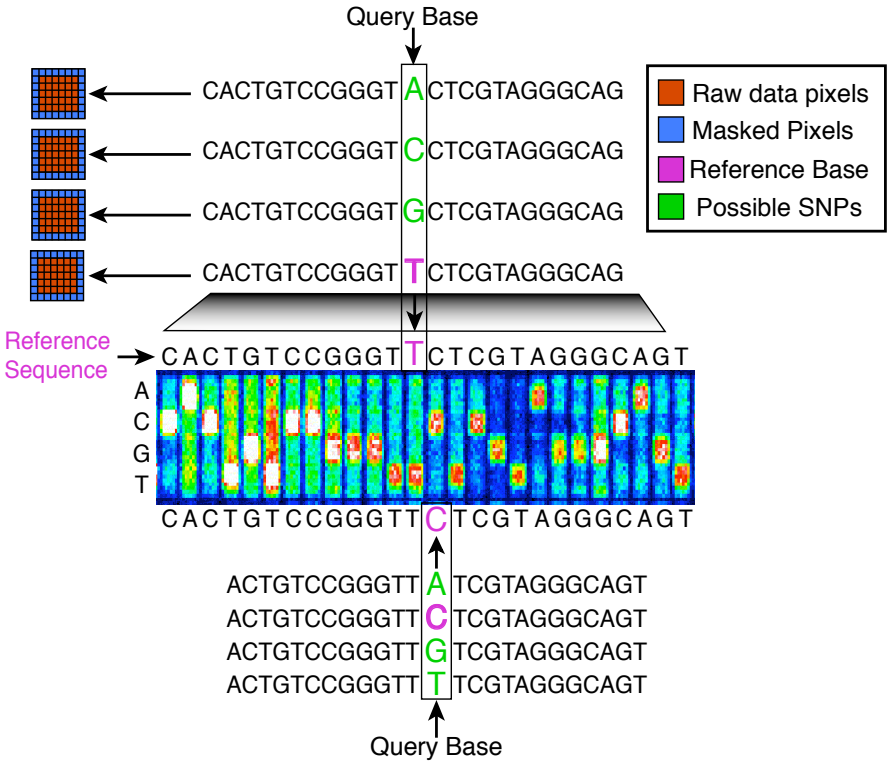

Query Base
CACTGTCCGGGTACTCGTAGGGCAG
CACTGTCCGGGTCCTCGTAGGGCAG
CACTGTCCGGGTGCTCGTAGGGCAG
CACTGTCCGGGTTCTCGTAGGGCAG
Raw data pixels
Masked Pixels
Reference Base
Possible SNPs
Reference Sequence
CACTGTCCGGGTTCTCGTAGGGCAGT
A
C
G
T
CACTGTCCGGGTTCTCGTAGGGCAGT
ACTGTCCGGGTTATCGTAGGGCAGT
ACTGTCCGGGTTCTCGTAGGGCAGT
ACTGTCCGGGTTGTCGTAGGGCAGT
ACTGTCCGGGTTTTCGTAGGGCAGT
Query Base

Figure I.4

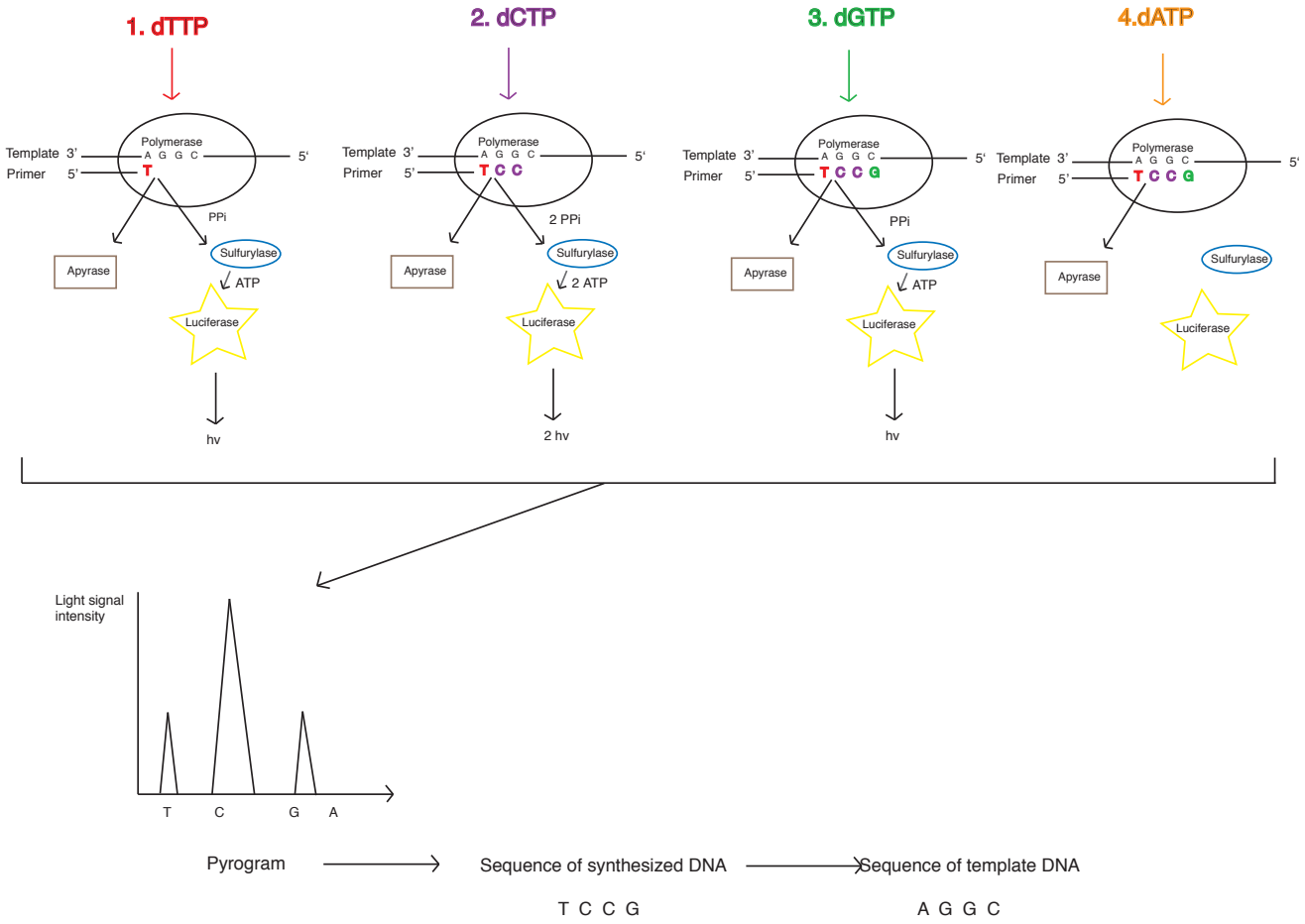
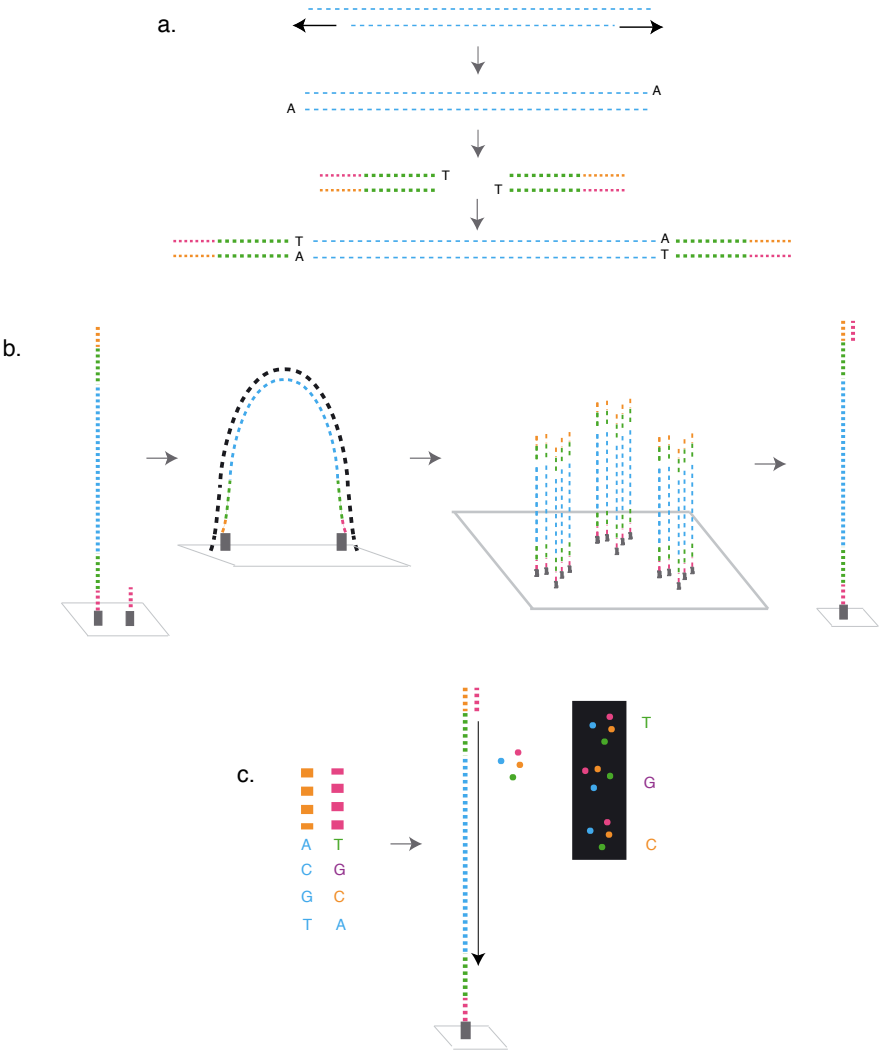

1. dTTP
2. dCTP
3. dGTP
4.dATP
Template 3'
Primer 5'
Polymerase
A G G C
T
T C C
T C C G
5'
PPi
2 PPi
Apyrase
Sulfurylase
ATP
2 ATP
Luciferase
hv
2 hv
Light signal intensity
T C G A
Pyrogram
Sequence of synthesized DNA
T C C G
Sequence of template DNA
A G G C

Figure I.5



CHAPTER 1

Microarray-based Genomic Selection for High Throughput Resequencing

David T. Okou[1], Karyn Meltz Steinberg[1,2], Christina Middle[3], David J. Cutler[1], Thomas J. Albert[3], Michael E. Zwick[1,2]†

[1]Department of Human Genetics, Emory University School of Medicine, 615 Michael Street, Suite 301, Atlanta, GA 30322, USA. [2]Program in Population Biology, Ecology and Evolution, Emory University, Atlanta, GA 30322

[3]NimbleGen Systems, Inc., Madison, Wisconsin 53711, USA.

†To whom correspondence should be addressed. E-mail: mzwick@genetics.emory.edu

ABSTRACT

We developed a general method, named Microarray-based Genomic Selection (MGS), capable of selecting and enriching targeted sequences from complex eukaryotic genomes without the repeat blocking steps necessary for BAC-based genomic selection. We demonstrate that large genomic regions, on the orders of hundreds of kilobases, can be enriched and resequenced with resequencing arrays. MGS, when combined with a next-generation resequencing technology, can enable large-scale resequencing in single investigator laboratories.

Published in *Nature Methods*. 2007 Nov;4(11):907-9.

K.M.S. assisted D.T.O. in performing experiments and assisted D.T.O. and M.E.Z. with data analysis and writing the paper.

Technological innovation in DNA sequencing offers the promise of a more comprehensive, cost effective, and systematic ascertainment of genetic variation.[49,56,57,59,67] A major bottleneck, however, lies in isolating the target DNA to be sequenced. Complex eukaryotic genomes, like the human genome, are too large to explore without complexity reduction using methods that directly amplifies specific sequences. Current approaches for target DNA isolation include short PCR[68,69], long PCR[56,57], fosmid library construction and selection[70], TAR cloning[71,72], selector technology[73], and direct genomic selection with bacterial artificial chromosomes (BACs)[74]. PCR using primer pairs complementary to specific genomic regions of interest is still the most common method sample preparation, but it is difficult to scale to large genomic regions, is labor intensive, and when primers are multiplexed, is subject to failure or artifacts. Random clone-based methods offer the advantage of obtaining complete haplotypes, but remain relatively expensive to scale.

Direct genomic selection, using BAC clones as hybridization “hooks”, has previously demonstrated the ability to isolate specific genomic regions without requiring specific amplification[74], but its adoption has been limited. Because BAC clones consist of a great deal of highly repetitive sequences, a number of protocol steps are required to minimize the enrichment of these types of sequences. Furthermore, because a single BAC is the unit of selection, isolating discontiguous unique sequence regions from across the genome would require multiple BACs. Finally, the existing protocol depends upon the presence of restriction sites adjacent to the targeted regions of interest that produce sticky ends for the ligation of generic adaptors. This acts to limit coverage in regions lacking these restriction sites. While random shearing followed by repair was mentioned as a possible alternative approach, it was not demonstrated [74].

To address these challenges, we have developed a method, *M*icroarray-based *G*enomic *S*election (MGS), capable of isolating user-defined unique genomic sequences from complex eukaryotic genomes. The MGS protocol consists of five main steps: (1) Physical shearing of genomic DNA to create random fragments with an average size of 300bp, (2) End repair of the fragments, that includes adding 3'-A overhangs, followed by ligation to unique adaptors with a complementary T nucleotide overhangs, (3) Fragment hybridization and capture using a custom high-density oligonucleotide microarray consisting of complementary sequences identified from a reference genome sequence, (4) Fragments bound to the probes are eluted, and (5) Selected fragments are amplified through one round of PCR using the adaptors as a single set of primers/template. Figure 1 provides a schematic overview of the method, starting with genomic DNA and ending with finished sequence across the targeted regions. The complete protocol is outlined in detail in the Supplementary Methods.

To demonstrate MGS, we captured and resequenced two X-linked genomic regions (Figure 2). The initial experiment examined a region 50Kb in size and included coding and non-coding sequences surrounding the fragile X mental retardation gene (FMR1). In a second, larger scale experiment we isolated and resequenced 304Kb of unique coding and non-coding sequences contained within a 1.7 MB genomic region that includes FMR1, FMR1NB and the AFF2 genes. Each custom MGS array consisted of ~385,000 long oligonucleotide capture probes (50-93bp) covering the regions of interest and were manufactured by NimbleGen Systems, Inc. Capture probe sequences included both the forward and reverse strands manufactured on a standard commercially available microarray to our specifications (Supplementary Text Files 1-2). For the 50 Kb region,

there were four pairs of probes for every targeted base, while the 304 Kb region had one pair of probes for every 1.5 targeted bases. The capture oligonucleotides were between 50 and 93 basepairs long and were designed to achieve optimal isothermal hybridization across the microarray.

Twenty micrograms of whole genome amplified genomic DNA were processed for each sample using the MGS protocol. Upon eluting the selected target from the capture MGS chip, we obtained yields of between 700ng and 1.2 μg. The eluted sample was split into between 5 and 10 PCRs, each of which was carried out using high fidelity *Taq* polymerase at an optimal concentration of 3ng/μl of PCR template. We have been able to reuse the MGS capture chips at least one time with no apparent contamination or effect on data quality (data not shown).

To assess MGS, we first sought to resequence a 50kb genomic region containing the FMR1 locus in cell lines derived from 2 patients with known FMR1 mutations: Tr91 contains a disease causing point mutation (A>T) at position 146825745 on the X chromosome while DM316 harbors a large deletion of the FMR1 gene[75,76]. We designed a custom NimbleGen 50Kb resequencing array that covered the targeted regions, containing both coding and non-coding sequences in the vicinity of the FMR1 gene (Figure 2), and resequenced both patients in triplicate using MGS. Analysis of the TR91 sequence identified the expected A>T point mutation when compared to the human genome reference sequence in all three replicates. Six additional variants were detected in TR91, 5 of which were successfully validated by independent sequencing (Agencourt Bioscience; see Supplementary Methods and Supplementary Table 1). As we expected, each of the three DM316 samples exhibited an absence of hybridization on the

resequencing array (RA) in the regions corresponding to the known deleted sequences (Supplementary Figure 1).

A total of 304 Kb was selected from 10 individual genome represented by two populations of different ancestry: a European descent (ED) population (n=5) selected from the Centre d'Etude du Polymorphism Humain (CEPH) panel and an African descent (AD) population (n=5) selected from the HapMap (Coriell Cell Repository numbers provided in Supplementary Methods). MGS was replicated twice for each of the ten samples. Using quantitative PCR, we estimated that MGS enriched targeted sequences ~1000-fold (Supplementary Figure 2).

Our resequencing results provide three lines of evidence demonstrating the efficacy of our MGS protocol. First, our total basecalling call rate over all 20 replicates (10 samples each processed twice) was 99.1% (6,528,393 called out of 6,585,832 total). This very high level of coverage implies that our MGS protocol efficiently enriches for the variety of sequences contained in the genomic regions we targeted. Second, for each sample, we counted the number of bases called identically and differently between both replicates. The reproducibility of RA base calls was 99.98%. Third, for each sample, to assess accuracy of basecalls, we compared our RA basecalls with genotype calls generated by the HapMap project (www.hapmap.org). We initially observed 39 discrepancies between RA and HapMap genotype calls. In order to identify the nature of the discrepancy, we had each of them independently resequenced via conventional ABI chemistry (Agencourt Bioscience, Beverly, MA). The resulting sequence data showed that 27 of the discrepancies agreed with our RA call, while 12 agreed with the HapMap genotype call. Hence, more than two thirds of the discrepancies we observed arose due to

errors in HapMap genotyping. Our final accuracy at segregating sites was thus 99.81% (Table 1).

The MGS protocol we describe uses routine enzymatic reactions and protocols that increase efficiency while minimizing risk of contamination and artifacts. The capture arrays are standard high-density long oligonucleotide arrays and are commercially available. The user can design the array to select multiple unique sequence fragments located throughout the genome for resequencing, or to comprehensively resequence genomic regions without the repeat blocking step necessary for BAC genomic selection. We are continuing to further pursue the tradeoff between probe density and sequence coverage and to increase the level of enrichment. Our current MGS microarrays contain ~385,000 capture probes. Current state of the art microarrays containing 2.1 million features and arrays with 4.5 million probes will be available in the near future. We believe that obtaining high coverage from genomic regions on the scale of megabases will soon be feasible.

Owing to the quality and comprehensive coverage of the data obtained, we believe that MGS will significantly contribute to a future where single investigator laboratories, using limiting infrastructure and requiring relatively few personnel, will be able to generate genome sequences at levels comparable to a conventional genome sequencing center. The ability of MGS to select multiple targets enables a comprehensive large-scale resequencing of user defined genomic regions that provide potentially important clues to the pathogenesis of complex diseases[69] or to find human genetic variation and functional sequences in both coding and non-coding regions[73]. Our method is useful for candidate gene studies that have been limited by sequencing capabilities and

offers the opportunity to select hundreds of genes in known pathways for resequencing. MGS would be useful in other eukaryotic model systems (i.e. mouse, zebrafish, Drosophila) to speed the sequencing of regions known to contain induced mutations. Finally, while we chose to use RAs, our approach is quite general. With continuing improvements in levels of enrichment, MGS should be able to be incorporated into existing sample preparation pipelines for instruments from Solexa[77] and 454[59], enabling even greater throughput at lower costs in the near future.

ACKNOWLEDGMENTS

Funding for this work was provided by the National Institutes of Health/National Institute of Mental Health RO1 MH076439-03 (MEZ) and by the Gift Fund, NIH (GIFT).

COMPETING INTERESTS STATEMENT

The authors declare no competing financial interests.

SUPPLEMENTARY METHODS

1. Array Design

We used the UCSC Table Browser function with repeats masked on the latest human genome build (March 2006) to identify the unique sequences within a selected genomic region[78]. The CGG repeat sequence of FMR1 from the human genome reference sequence was included in the design. Since genetic variants in regulatory elements away from the coding sequences may influence the expression of a gene [79], unique sequence upstream and downstream of the target genes were also be included. We then selected among the unique sequence to obtain ~50 Kb or 304 Kb of unique sequence. We excluded unique sequences 100 bp or less and in some cases, we added short (<100 bp) stretches of previously masked sequence, to avoid breaking up long stretches of genomic regions.

The FASTA format sequences were then provided to chip design engineers at NimbleGen to select oligonucleotides for the microarray-based genomic selection (MGS) array. Standard bioinformatics filters that check for genomic uniqueness against an indexed human genome (15mers) were used to select capture oligos. The capture oligonucleotides were between 50 and 93 basepairs long and were designed to achieve optimal isothermal hybridization across the microarray. No other optimization of oligos was performed. For the 50 Kb region, there were four pairs of probes for every targeted base, while the 300Kb region has one pair of probes for every 1.5 targeted bases.

Resequencing arrays were designed from the FASTA format sequences provided to design engineers at Affymetrix (FMR1/FMR2) and NimbleGen (FMR1 only). Resequencing Arrays (RAs) query a given base by using overlapping oligonucleotide

probes, tiled at a 1-basepair (bp) resolution. The oligonucleotide probes, referred to as features, are typically 25 basepairs long. Both the forward and reverse strands are interrogated, so sequencing a single base requires a total of 8 features. A set of four features contains oligonucleotides identical to the forward reference strand, except at position 13 (the base to be queried), where there is either A, C, G, or T. The remaining four features are similarly designed for the complementary strand. When a labeled DNA sample, called a target, is hybridized to these eight features on the array, the two features complementary to the reference sequence (forward and reverse complement) will yield the highest signal. If, however, the target DNA contains a variant base at position 13, the two features complementary to that variant base will yield the highest signal. Given eight features for each base, interrogation of an L-length duplex strand would require 8L oligonucleotide probes.

2. Sample Selection

DNA samples were purchased from the Coriell Cell Repository (http://ccr.coriell.org) and included 10 individual genomes represented by two populations of different ancestry: a European descent (ED) population (n=5) selected from the Centre d'Etude du Polymorphism Humain (CEPH) panel with the Coriell Cell Repository numbers: NA07029, NA07048, NA10846, NA10851 and NA10860; and an African descent (AD) population (n=5) selected from the HapMap with the Coriell Cell Repository numbers: NA18500, NA18503, NA18506, NA18515 and NA18521. MGS was replicated twice for each of the ten samples. Other samples used in this study were extracted from cell lines representing fragile X patients with either disease causing point

mutation (A>T) at position 146825745 on the X chromosome (Tr91) or deletion (DM316) in the fragile X mental retardation (FMR1) gene [75,76].

3. Adaptor and Primer Design

All oligonucleotides used in this project were obtained from Invitrogen Corp. The adaptor was prepared by annealing the forward (21 bp) and reverse (22 bp) oligonucleotides to generate a 21 bp dsDNA fragment with single and double base "T" overhangs at the 3 prime and 5 prime end respectively (Supplementary Figure 3). Annealing of the oligos was performed by mixing both oligonucleotide to a final concentration of 1.5 μg/μl each oligo, heating to 95'C for 10 minutes in a heating block, turning off the heating block and allowing the mixture to slowly cool back to room temperature. The primers used for the enrichment were made by preparing a 20 μM of each oligonucleotide used for the adaptor.

4. Genomic DNA preparation

Whole genome amplification was performed on 100 ng of genomic DNA using the RepliG Kit (Qiagen Inc.). Following amplification, the unpurified samples were quantified using a spectrophotometer (NanoDrop). Twenty-five micrograms of each sample was aliquoted into sterile Eppendorf tubes for a final concentration of 100 ng / μl (250 μl).

5. Target DNA isolation

Samples were sonicated (Misonix sonicator 3000) in Eppendorf tubes with a microtip probe using the following parameters: 3 pulses of 30 seconds each with 2 minutes of rest and a power output level of two. After fragmentation, approximately 750 ng of each sample was run on a 1.5% TAE agarose gel against 750 ng of a 1 Kb plus ladder to verify that fragments average 300 bp in size. The samples were then dried down in a SpeedVac at medium heat to 47 µl (75° C).

6. Repairing Ends of Sheared DNA

To the 47 µl fragmented DNA we added 8 µl of dNTPs (2.5 mM, TaKaRa), 8 µl of 10X T4 DNA Polymerase Buffer (NEB), 1 µl of 100X BSA (NEB), 1 µl 100mM ATP, 14 µl of T4 DNA Polymerase (3U/µl, NEB), and 1 µl of T4 Polynucleotide Kinase (10U/µl, NEB). We then incubated in a thermocycler at 12°C for 20 minutes followed by 37°C for 30 minutes and 70°C for 5 minutes. After incubation we directly added 2 µl of 10X T4 DNA Polymerase Buffer (NEB), 2 µl 100mM dATP (Sigma), 3 µl of 50mM MgCl2, 8 µl of VWR H2O, and 5 µl of Taq DNA Polymerase (5U/µl, NEB). Samples were incubated in a thermocycler at 72°C for 45 minutes. After incubation we used the Promega Wizard® SV Gel and PCR Clean-Up System following the manufacturer protocol. Each column was eluted with 70 µl of water, the volume adjusted to 71 µl and 1 µl removed to perform NanoDrop quantification.

7. Ligation of Adapters

The following reaction(s) were performed in a 0.2 ml PCR tube. To the 70 µl repaired reaction 10 µl of 10X T4 DNA Ligase Buffer (NEB), 15 µl of Adapters (1.5 µg/µl) and 5 µl of T4 DNA Ligase (2000U/µl, NEB) was added. This was incubated at room temperature for 2 hours. The insert to vector ratio was calculated in terms of insert ends to vector ends. The number of ends available for ligation in pmoles can be calculated as follows:

pmol ends/µg of DNA = (2 x 10^6) / (number of base pairs x 660)

The ratio of adapter to DNA should be at least ~12:1. While this increases the chance of getting some adapter concatamer (which should not hybridize to the array), all of the fragments will likely get adapters, which is very important. When the ligation was complete, the sample was transferred to a 1.5 ml tube and 100 µl of VWR water was added. The Promega Wizard® SV Gel and PCR Clean-Up System was used following the manufacturer protocol. Each column was eluted with 50 µl of water and 1 µl was removed to perform NanoDrop quantification.

8. Hybridization

To the ligated sample we added a 5-fold amount (in µg) of human Cot-1 DNA (Invitrogen). The sample was dried in the Speed-Vac at medium heat (75°C) for 45 minutes. The sample was vortexed for 3 minutes and drying continued to the pellet. The following reactions were performed in a 1.5 ml tube. To the pellet from dried sample 7.2 µl of VWR water, 8.25 µl of 2X Hybe Buffer (NimbleGen) and 1.43 µl Hybe Component A (NimbleGen) was added. The samples were vortexed 3 minutes and then heated at

95°C for 10 minutes. The samples were quickly spun down and placed in the MAUI heat block at 42°C until ready to use. Once the samples were applied to the chip surface, we began the mixer on program B and hybridized for 60 hours.

9. Elution

After hybridization, the MGS arrays were first prewashed at 42ºC in NimbleGen Buffer 1 followed by two 5 min washes at 47.5ºC with NimbleGen Stringent Buffer. The arrays were then washed at room temperature for 2 min with NimbleGen Buffer 1, 1 min with NimbleGen Buffer 2 and 30 seconds with NimbleGen Buffer 3. We placed the washed chip on the Hybriwheel (NimbleGen) at 100°C and secured with a Hybe Puck (NimbleGen). We added 400 µl of 95°C VWR water and incubated 5 minutes. After the 5 minute incubation we removed as much water as possible and pipetted it into a labeled 1.5 centrifuge tube (placed on ice). We repeated this process one more time beginning with the addition of 400 µl of 95°C VWR water to the puck. When this was complete, we added 350-400 µl of 95°C VWR water and removed it immediately and pipetted it into the 1.5 ml tube.

After elution, the sample was placed in the Speed-Vac at medium heat (75°C) for 45 minutes. The sample was vortexed for 3 minutes and drying continued until the sample was to the pellet. We then hydrated the pellet in 33 µl of VWR water and vortexed for 3 minutes. We performed NanoDrop quantification of single strand DNA (DNA -33) to determine the concentration of the sample (picogreen and ethidium bromide quantification are inefficient for single stranded DNA). Upon eluting the

selected target from the capture MGS chip, we obtained yields of between 700ng and 1.2 μg.

10. Amplification (LMPCR)

From the NanoDrop quantification we determined the number of LMPCR reactions to carry out using 110 -150 ng of template per 50 μl reaction. For example, if the eluted sample has 960 ng of total DNA, then we had 8 reactions with 120 ng of template DNA per reaction to amplify the entire eluate. The expected yield is 4 - 8 μg of product / 50 μl reaction. It is important to use at least 110ng of template to maximize yield; 120ng is optimal.

To the 50 μl reaction we added 5 μl of 10X LA PCR buffer (TaKaRa), 5 μl of 2.5 mM dNTPs mix (TaKaRa), 2 μl of 20 μM FWD LMPCR primer, 2 μl of 20 μM REV LMPCR primer, and 2 μl of LA Taq (5U/μl, TaKaRa), and VWR water to 50 μl. This reaction was incubated in a thermocycler at (1) 95°C for 2 minutes, (2) 95°C for 60 seconds, (3) 58°C for 60 seconds, (4) 72°C for 60 seconds, (5) Repeat step 2 30 times (35 cycles), then at 72°C for 5 minutes and finally hold at 4°C.

All PCR reactions were pooled by sample and transferred into a 1.5 ml tube. We used the Promega Wizard® SV Gel and PCR Clean-Up System following the manufacturer centrifugation protocol. For spin steps we used 13000 g, and for the elution spin we used 16000 g and 1.5 minutes. Each column was eluted with 50 μl of water.

Three to 5 μl were used to verify size distribution on 1.5 % TAE agarose gel against 500 – 750 ng of 1 Kb plus ladder and positive control (6 X xylene cyanol loading dye for samples). Then the samples were quantified using NanoDrop and sonicated.

11. Resequencing of selected DNA

NimbleGen's Comparative Genomic Sequencing protocol was used for the 50K RA. Briefly, 1 μg of sample was denatured at 98ºC for 10 min in random primer buffer and labeled in the dark with Cy3-9mer primers (TriLink BioTechnologies) in the presence of dNTP mix and 100 units of Klenow (50U/μl, NEB) for 2 hours. To guarantee at least 20 μg of label sample for resequencing, 2 labeling reactions were done per sample (2 μg total). Labeled samples were purified using ethanol precipitation method and dried down to the pellet in the dark to avoid bleaching of the Cy3 dye. After rehydrating the pellets with 20 μl total of VWR H_2O, ten to thirty micrograms of labeled DNA was mixed with NimbleGen's Hybridization cocktail (2X hybe buffer and hybe component A) and denatured at 95ºC for 5 min. The arrays were loaded and incubated overnight at 42°C on MAUI Hybridization System (BioMicro). The signal was detected by measuring Cy3-chrome fluorescence using Genepix 4000B (Molecular Devices Corp.).

For Affymetrix RAs, 30 μg of enriched samples were digested to 20 to 100 bp for 3 min in a 42μl reaction comprised of 10X Phor-All_Buffer (Amersham Biosciences), 10X Acetylated BSA and 3 units of DNAse1 (Promega). Reactions were heated at 75° C for 10 minutes to inactivate the DNAse then to 95° C for 15 minutes to separate the strands. The reactions were then cooled at 4° C for 45 minutes. The fragmented DNA was labeled using 17.13 nmol of a biotinylated proprietary labeling reagent (Affymetrix), 4.5 units of terminal deoxynucleotidyl transferase (Affymetrix) and terminal deoxynucleotidyl transferase buffer (Affymetrix) at a final concentration of 1X. The

reactions were brought to a volume of 60µl with nuclease free water (VWR). Each reaction was incubated at 37°C for 4 hours followed by heat- inactivation for 15 minutes at 95°C and stored at 4° C until ready to use.

The labeled DNA samples were combined with 160 µl Hybridization buffer comprised of 1M Tris HCl pH 7.8 (Sigma), 5M TMACL (Sigma), 0.10% Tween 20 (Pierce Biotechnology), 100 µg/µl of Herring Sperm DNA (Promega), 500ug/ml Acetylated BSA (Invitrogen), and 200pM biotinylated SNPHy948B (Invitrogen). The hybridization mix was then heated to 95°C for 5 minutes, equilibrated at 49°C and hybridized to the high-density oligonucleotide array at 49°C for 16 hours. All signal detection steps were performed using an Affymetrix fluidics. The arrays were washed in 6X SSPE, 0.01% Tween 20 solution (wash A) 6 times at 25°C then in .6X SSPE, 0.01% Tween 20 solution (wash B) 6 times at 45°C. For signal detection, the arrays were incubated with stain 1 (6X SSPE, 0.01% Tween 20, 1X Denhardt's solution (Sigma), and 10ug/ml SAPE (Invitrogen), final concentration) for 10 minutes at 25°C, followed by 6 washes with wash A at 25°C. Incubation with stain 2 (6X SSPE, 0.01% Tween 20, 1X Denhardt's solution (Sigma), and 10ug/ml anti-streptavidin antibody (Vector), final concentration) was done for 10 minutes at 25°C. A second incubation with stain 1 was done for 10 minutes at 25°. The arrays were rewashed 10 times in wash A at 30°C and filled with a holding buffer (5M NaCl, 10% Tween 20, MES hydrate and MES sodium salt). They were stored at 25°C until they were ready to be scanned. The signal was detected by measuring Cy-chrome fluorescence using a G7 Genechip scanner (Affymetrix). For both the NimbleGen and Affymetrix resequencing arrays, all bases calls were made with the RATools program RA_PopGenCaller (http://www.dpgp.org/).

12. Validation Sequencing

Discrepancies between RA data and HapMap data were evaluated using independent sequencing (Agencourt). PCR primers were designed using Primer 3 (http://frodo.wi.mit.edu/). PCR Reactions were composed of 400 ng of sample DNA was mixed with 8 μl of dNTP mix (TaKaRa), 5 μl of 10X LA Taq buffer (TaKaRa), 1.5 μl LA Taq (TaKaRa), 0.8 μl of each forward and reverse primers and VWR water to 50 μl total reaction volume. DNA was amplified using the following parameters: 94°C for 4 min, 30 cycles of 94°C for 20 sec, 58°C for 1 min, and 72°C followed by 72°C for 5 minutes. This method was also used to validate discrepancies in the Tr91 RA data. The primers that amplified the SNP discrepancies are listed in Supplementary Table 2. PCR products were run on a 1% TAE agarose gel, excised from the gel and purified using the Promega Wizard® SV Gel and PCR Clean-Up System.

13. Long PCR Control

To minimize the number of amplifications, we used long PCR to amplify genomic regions that contain one or more unique sequence blocks tiled onto the variant resequencing array. A total of 14 primer pairs spanning 48 Kb (including the 39 kb FMR1 genome region) were used. Except for one primer close to the CGG repeat (20 bp) Long PCR primers were 31 to 34 base pairs long and were selected by using Amplify 3.1.4 [80] to ensure that they bound uniquely within a 48 kb region and had a primer stability value between 70 and 80. Primers had GC content between 45% and 60%.

Amplification of genomic DNA was accomplished in 50 μl reactions carried out

in thin-walled polypropylene tubes using LA Taq (TaKaRa). The manufacturer's recommendation was followed. LPCR amplification of the human samples employed either a standard or a modified mixture where 5% DMSO (or manufacturer GC Buffer) was added to aid the amplification of GC rich regions. The standard conditions for the LPCR were: (1) 94˚C for 2 minutes, (2) 94˚C for 10 seconds, (3) 68˚C for 1 minute per kb fragment size, (4) repeat to step 2, 30 times, and (5) final extension time equal to step 3 plus five minutes. Each LPCR required a minimum of 200ng of human genomic DNA and most fragments were between 3.4 and 11 kb long. To obtain optimal performance across the microarray, we pooled equal molar concentration of PCR product, to ensure that an equal number of targets existed for each probe on the array. The primer sequences that amplified each fragment are listed in Supplementary Table 1.

14. Quantitative PCR

We performed quantitative PCR on sample DNA with two treatments: (1) whole genome amplified, ligated and then amplified using LMPCR protocol and (2) eluted from genomic selection with LMPCR. We used the iQ SYBR® Green Supermix (Bio-Rad) and the following primer pair:

FW: ACAGTAGGGCTGTGCTTACTGC

REV: CTCATTTTCAGCCTCAATCCTC

The primers amplify 156 bases from exon 10 in the FMR1 gene. Reactions contained 12.5 µl of 1X iQ SYBR® Green Supermix, 1 µl of FW Primer (10mM), 1µl of REV Primer (10mM), 9.5 µl of VWR water and 1 µl of DNA template (30 ng/µl) for a total volume of 25 µl. The standard curve was created using whole genome amplified DNA at

concentrations ranging from 500 ng/µl to 7.8 ng/µl. The reactions were performed in triplicate. The reactions were incubated in a Bio-Rad iQ5 Multicolor Real Time PCR Detection Light Cycler using the following parameters: (1) 94˚C for 3 minutes, (2) 94˚C for 10 seconds, (3) 58˚C for 30 seconds, (4) 72˚C for 30 seconds, and (5) Repeat steps 2-4 for 40 cycles. From our quantitative PCR result we conservatively estimate at least 1000X enrichment of DNA used for resequencing (treatment 2) when compared to whole genome amplified DNA that underwent LMPCR amplification (treatment 1). The DNA from treatment 2 has a cycle threshold of 15 while the cycle threshold for treatment 1 is 25. If we assume that DNA doubles every cycle then enrichment can be calculated by 2^N, with N equaling the difference between the cycle thresholds of the two treatments (Supplementary Figure 3).

Table 1.1

Table 1 Assessment of 304kb RA data quality

Replication Experiment	
Total number of bases called identically in both replicates	6,492,426
Total number of bases called differently in both replicates	1115
Percent of Bases Called Identically	99.98%
Accuracy Estimation	
Total number of bases called identically	6280
Total number of bases called differently (before validation)	39
Total number of bases called differently (after validation)	12
Accuracy at segregating sites	99.81%

Table 1.2 (Supplementary Table 1) Primer sequences used in independent sequencing validation of HapMap and Tr91 discrepancies.

HapMap Samples	
rs16994908_FW2	**CTTCACCATTTTTGCATGTACC**
rs16994908_REV	**TTGCAACCACATTTGAAGTGAC**
rs12688573_FW	**AAAGTCGCACAGATACCCTCTC**
rs12688573_REV	**CTTTTCTGTCTTGCCATTAGCC**
rs11117557_3_FW	**ACTGCATCTGCAGAGAAACAAC**
rs11117557_3_REV	**AACAGTTGTGAAACTACGTCAGG**
rs7052829_FW	**TTATGGGAAGAATCCACTCCAG**
rs7052829_REV_2	**AGTAGCAGCAACAGCAACAAAG**
rs7052654_rpt_FW	**CAGGGCAGGGATGATTAGAG**
rs7052654_rpt_REV	**AGAAAGGAAGAGATGCATGGAC**
rs6626955_6_FW	**TCCCTTGTGTTCATGGAGTATG**
rs6626955_6_REV	**AACAGGAGCTTCTTCCTGATTG**
rs2761622_2_FW	**AAATGAAATGCACCTTCCAGAG**
rs2761622_2_REV	**GCACTTGTTTCACAGGTACAGC**
rs1805422_FW	**GTAGCAGTAGTGCGTTTGTTGG**
rs1805422_REV	**TTTCCTATAGCCAAACGTGTCC**
rs1265401_FW	**GGGTATGGGTTTAACATAGGACAG**
rs1265401_REV	**GACTTACGGGCTGCTTCTCAC**
rs1265397_FW	**GCATGCGTGTCTTACTCCATAG**
rs1265397_REV	**AAGCTCTGTCAGTGTGATGTGG**
rs25699_FWD	**GCCAGAGGCTATTTCCCTAACTTAC**
rs25699_REV	**TGATGACGAACTCTGGAATTTGAC**
rs4949_FWD	**AGAGTGCTTTTGTTGGGATGTAC**
rs4949_REV_2	**attacacacataGGTGGCACTA**
rs1442280_FWD	**AGACATTGCAAACATCCAGAAC**
rs1442280_REV	**ATGCAGTCAGCCAGGTAATAGA**
rs16994869_FWD	**tgAACAGTCACTTGACATCCAAAG**
rs16994869_REV	**GATTGGAGGAGGCAGAGAAATAGT**
Tr91	
rs29284_int9_FW	**CTCTGGTACCTGACCAAAGGAG**
rs29284_int9_REV	**AAAGCAGTAAGCACAGCCCTAC**
rs29288_int13_FW	**CATGCCATTCATTCTTATGGTG**
rs29288_int13_REV	**AATCCTAACTCTCCAGGCCTTC**
rs25707_ex5_FW	**CCTGCCACAAAAGATACTTTCC**
rs25707_ex5_REV	**TTCTCCATTGCTCTTGCAAAC**
I304N_ex10_FW	**ACAGTAGGGCTGTGCTTACTGC**
I304N_ex10_REV	**CTCATTTTCAGCCTCAATCCTC**
rs29286_int12_FW	**GTGGCTTCATCAGTTGTAGCAG**
rs29286_int12_REV	**CACATACCCACAAACACTCCTC**
rs5904816_int14_FW	**GCACATCAAGGTTTGAACTTAGG**
rs5904816_int14_REV	**CAGAGACGTTTCAGGGGTAATC**
rs25704_ex17_FW	**GGAAGGTCATTTCCATGTATGC**
rs25704_ex17_REV	**AAAACCAAACCCCAACACTTC**

FIGURE LEGENDS

Figure 1.1. (Figure 1) Microarray-based genomic selection and resequencing of complex genomes. Sheared genomic fragments (A, B) are repaired and ligated to generic adaptors (C). Hybridization to a custom designed high-density oligonucleotide microarray allows the capture of the target DNA regions (D). The selected target is eluted (E) and amplified using a one step PCR and a single primer pair /template (F). We resequenced the amplified target with resequencing arrays (G) analyzed with RATools.

Figure 1.2. (Figure 2) Genomic regions (50kb, 304kb) resequenced in the two MGS validation experiments. Targeted sequences included both coding and unique non-coding genome sequences.

Figure 1.3. (Supplementary Figure 1) RA hybridization results for TR91 (A) and DM316 (B) samples. The large absence of hybridization on the DM316 array is the result of a large deletion of much of the FMR1 locus.

Figure 1.4. (Supplementary Figure 2) Results of quantitative PCR assay measuring the extent of enrichment after a single round of microarray-based genomic selection (MGS).

Figure 1.5 (Supplementary Figure 3) Schematic showing design of adapter oligonucleotides used in the MGS ligation and later PCR amplification steps.

Figure 1.1

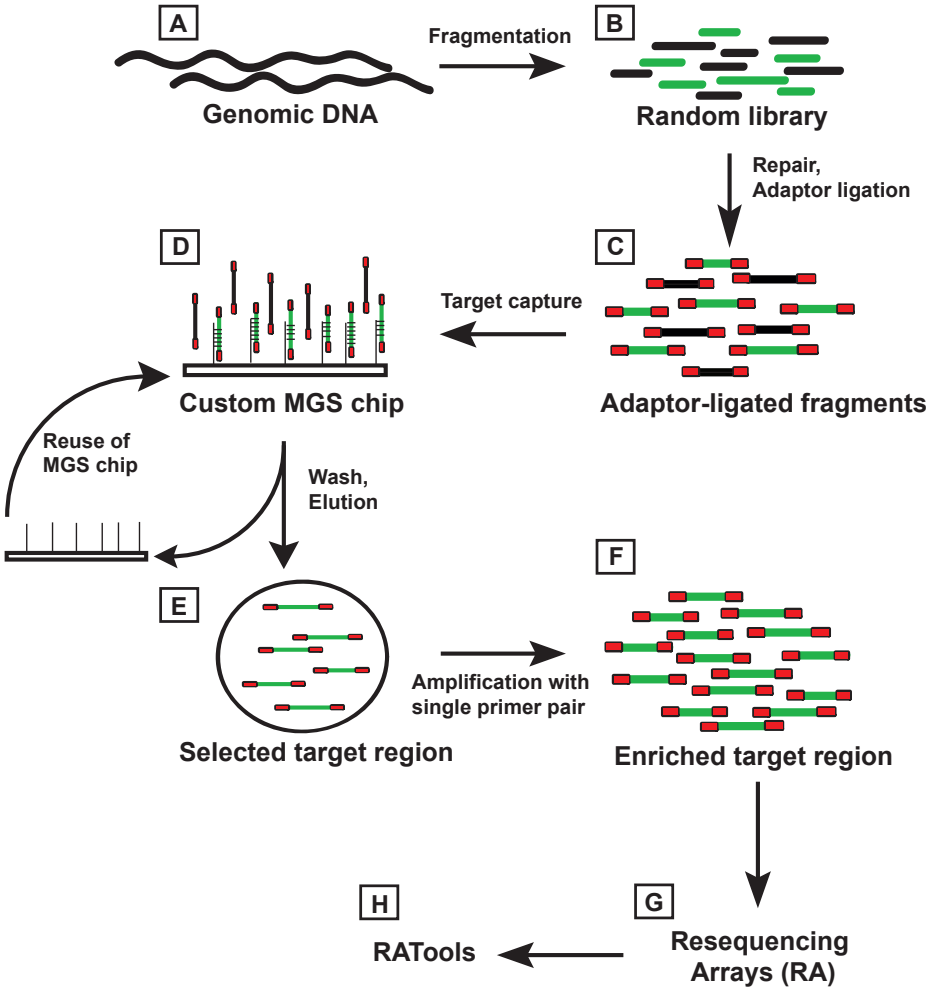

A
Genomic DNA
Fragmentation
B
Random library
Repair,
Adaptor ligation
C
Adaptor-ligated fragments
Target capture
D
Custom MGS chip
Reuse of
MGS chip
Wash,
Elution
E
Selected target region
Amplification with
single primer pair
F
Enriched target region
G
Resequencing
Arrays (RA)
H
RATools

Figure 1.2

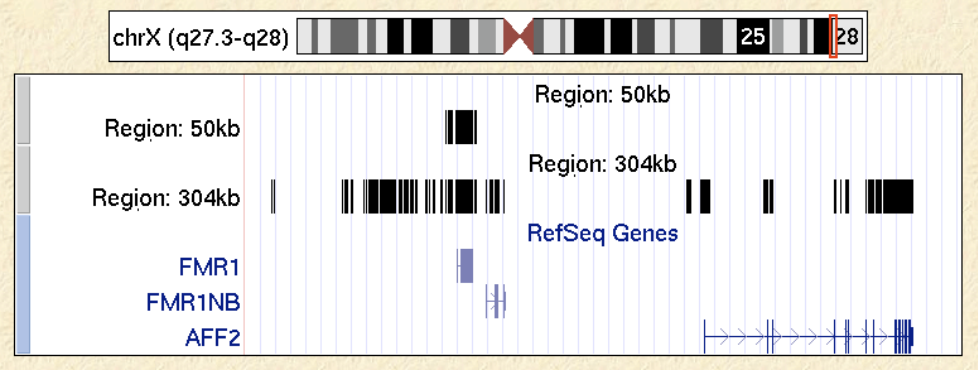

chrX (q27.3-q28)
25
28
Region: 50kb
Region: 304kb
Region: 50kb
Region: 304kb
RefSeq Genes
FMR1
FMR1NB
AFF2

Figure 1.3
(Supplementary Figure 1)

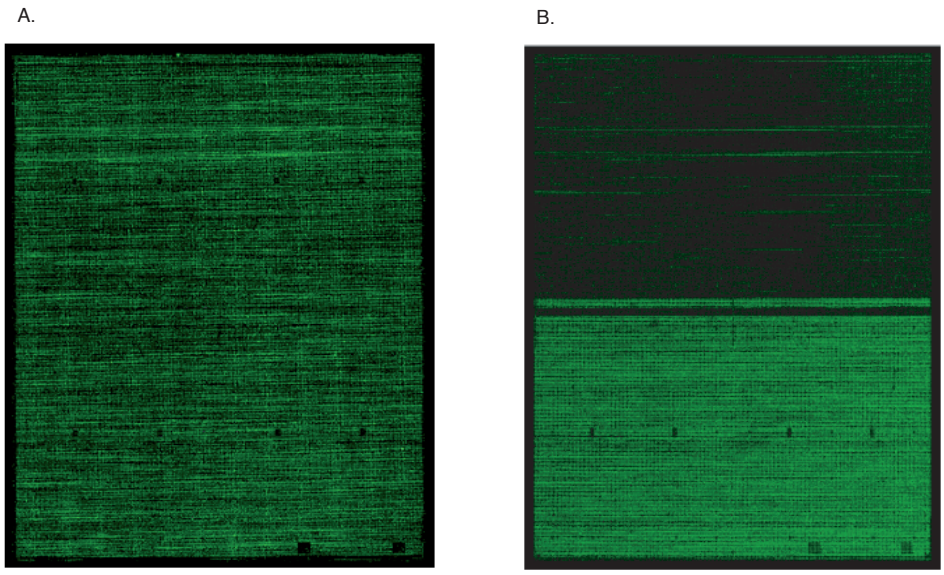


Figure 1.4
(Supplementary Figure 2)

Quantitative PCR Results

Treatment 1: WGA Genomic DNA --------> Ligation ---------> LMPCR

Treatment 2: WGA Genomic DNA --------> Ligation ---------> GS Chip Hybridization ----------> Elution ----------> LMPCR

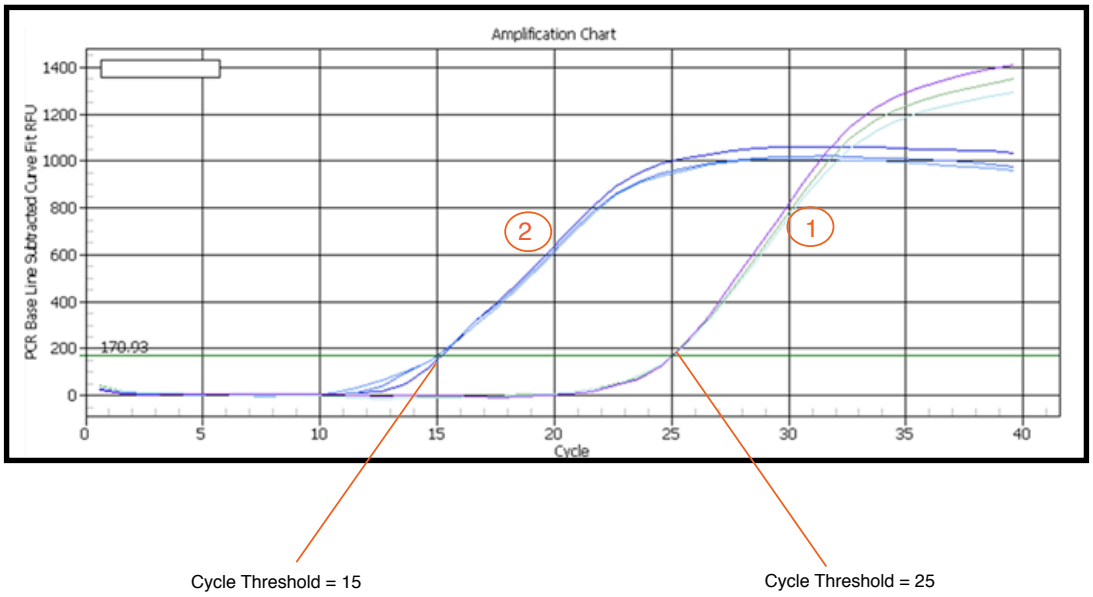
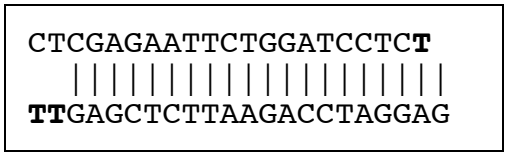


Figure 1.5
(Supplementary Figure 3)



CHAPTER 2

INTRODUCTION

Recent advances in high throughput DNA sequencing technologies, which have applied novel-sequencing chemistries in highly parallel and miniaturized formats, have resulted in major advances in two areas [58,62,81]. First, these approaches have fueled an exponential increase in the amount of DNA sequence obtained in a single experiment. Second, the cost per high quality base pair generated has been dramatically reduced. While these technological developments continue to move us closer to rapidly and inexpensively sequencing entire human genomes, clinical applications of whole genome sequencing will not be broadly adopted until accuracy and scalability in single, diagnostic laboratories improves. Targeted capture and resequencing of protein coding regions, or exomes, is a viable clinical alternative to whole genome sequencing [82-85]. For example, exome sequencing is a method that could be used in cancer genetics to identify mutations in tumor tissue when compared to the patient's germline sequence [86]. Additionally, complete exome resequencing in unaffected human populations can be expected to help determine which variants are truly associated with disease and which are normally present in the population. [87]

The conventional method of generating target DNA for sequencing using short PCR requires too many pairs of primers, it is too expensive, and too slow to efficiently amplify large target regions. For example, a recent report of 22 tumor samples required over 135,000 primers and 3 million PCR reactions [69]. As described in the previous chapter, we developed Microarray-based Genomic Selection (MGS) to address issues of target DNA isolation in large resequencing studies [66]. MGS is able to isolate specific unique genomic sequence from complex, eukaryotic genomes by using capture

oligonucleotides bound to a solid surface. Rather than focusing on a small candidate region implicated in a disease phenotype, here we evaluated MGS on a larger scale by first selecting and then resequencing all of the exons on the X chromosome (the X exome) in ten HapMap males (Coriell numbers listed in Methods).

RESULTS

We chose 7066 exons (ranging in size from 12 to over 9000 bases) to capture and resequence on resequencing arrays in ten male samples. Arrays were designed and manufactured by NimbleGen Systems, and probes were tiled every 5 bases. Fasta files containing desired target sequences were provided to NimbleGen for array design. NimbleGen then used bioinformatics filters that assessed genomic uniqueness against an indexed human genome (15mers) to select oligos between 50 and 93 basepairs long.

Statistical analysis of MGS probes

Across all ten samples, approximately 30% of exons had basecalling rates over 90% while approximately 74% of exons had basecalling rates over 70% (see Figure 2.1). While these results were promising there were a significant number of exons that had basecalling rates below 90%. The causes of this failure could arise from two sources: the MGS array could have failed to select the exons that had low basecalling, or, alternatively, the MGS array could have selected the exon, but the resequencing array did not accurately call the bases. To begin to tease these two confounding hypotheses apart, I performed univariate analyses to compare the 219 exons with the highest basecalling rates and the 216 exons with the lowest basecalling rates. I compared basecalling with

three main variables: number of MGS probes, GC content of probes and MGS probe melting temperature (Tm). The number of MGS probes acts as a proxy for exon length as the probes were tiled evenly every 5 bases regardless of coding sequence length. The results demonstrate that longer exons, exons with lower GC content and exons with lower melting temperature (Tm) all had higher basecalling rates (see Figure 2.2).

Chip redesign

To address the issue of Tm and GC content I redesigned an MGS array to have capture probes with an optimal Tm (mean Tm=70.8°C, range 56°C -89°C). I then designed a 50 Kb NimbleGen resequencing array to contain 25 Kb of sequence that performed well in the initial experiment and 25 Kb of sequence that failed in the initial experiment. We predicted three different possible outcomes. If the initial experiment was solely a failure in target DNA capture by the MGS array and not influenced by resequencing on the RA, we expected that our redesigned capture array should have improved the capture efficiency and that this would be reflected by improved basecalling on the RA. Alternatively, if the experiment had failed because the RAs performed poorly, even in the presence of sufficient target DNA, we would have expected to continue to see poor base calling at those sequences that we previously failed to resequence. Finally, it is possible that both the capture and resequencing arrays failed in a similar fashion - that would lead to poor sequence data. We note if this latter hypothesis were true, we would not be able to distinguish the second or third hypotheses in the case of a negative outcome.

Each sample was prepared using the same MGS protocol and resequenced once. Average basecalling was 45% over the entire chip. Exons that had previously high basecalling rates had high basecalling rates in this experiment, while exons that had previously failed also failed in this experiment. These results suggest that adjusting for Tm could not fully eradicate the variance in probe performance. There may be other variables for which we did not account that could be contributing to variance in probe success or failure. Additionally, these results do not account for success or failures in the RA features, which may also be confounding these results. RA probes were generated identically to MGS probes, although they were only 25 bases long and tiled at a 1 base resolution. Therefore, it is impossible to resolve the confounding hypotheses that either or both the MGS and/or RA probes failed based upon these results. Currently, the lab is pursuing new sequencing technologies, such as Illumina sequencing, paired with MGS target DNA isolation. Results from these experiments may shed some light on whether redesigning MGS probes can decrease the variance in probe success and increase the amount of sequence generated over a large set of genomic targets.

DISCUSSION

The results from this X exome resequencing experiment indicate that standard MGS array designs from NimbleGen cannot accurately select small exons and those exons with high GC content and high Tm. In addition, simply adjusting probe Tm could not increase target DNA selection alone. To overcome these issues we are currently exploring probe designs generated by software designed by Viren Patel. This software chooses probes of specific length and within a specific range of GC content and Tm. The

program is also able to densely tile probes for smaller exons to increase the likelihood of capturing target DNA in these genomic loci. In addition, a new protocol for eluting target DNA off of the array using NaOH may help to eliminate this temperature effect (data not shown).

One study that examined long probes (45-85-mer) used for copy number detection demonstrated that probe performance was inversely correlated with probe uniqueness and positively correlated with length [88]. This study also identified that higher Tm is associated with lower probe performance while lower and more uniform Tm is associated with higher probe performance. Another study of expression array probe performance demonstrated that the number of probes per gene significantly affected signal with a greater density of probes per gene giving a more reliable signal than fewer probes [89].

A study by Hodges et al [90] demonstrates the difficulty in capturing exons with high accuracy and efficiency. In attempting to select 44 MB of sequence (coding exons and adjacent splice sites) over 7 different arrays and sequence using Illumina, they were only able to achieve 237-fold enrichment and only 36-55% of the reads mapped within the exon boundaries.

Ng et al [82] demonstrated microarray-based capture and highly accurate whole exome sequencing of 8 HapMap individuals and 4 individuals with an autosomal dominate Mendelian disease. The sequence data from the HapMap individuals show a high concordance with heterozygous genotypes (greater than 99%), and the patient/proband data suggest that the causative gene could be identified using this method. The ability to accurately identify heterozygous variants is vital to medical

genetics as many disease causing variants are likely to be rare and in the heterozygous state.

The results from the univariate statistical analyses performed here are currently being used to further improve the probe design and MGS protocol for selecting and resequencing the X exome.

METHODS

We used the UCSC Table Browser function with repeats masked on the latest human genome build (March 2006) to identify the coding sequences on the X chromosome. This amounted to 7066 exons ranging from 12 to 9638 bases. One pair of probes was tiled for every 5 targeted bases. The custom MGS arrays were designed and synthesized by NimbleGen Systems, Inc., and contained approximately 385,000 probes that were 50 to 93 basepairs long.

DNA samples were purchased from the Coriell Cell Repository (http://ccr.coriell.org) and included 10 individual genomes represented by two populations of different ancestry: a European descent (ED) population (n=5) selected from the Centre d'Etude du Polymorphism Humain (CEPH) panel with the Coriell Cell Repository numbers: NA07029, NA07048, NA10846, NA10851 and NA10860; and an African descent population (n=5) selected from the Hapmap with the Coriell Cell Repository numbers: NA18500, NA18503, NA18506, NA18515 and NA18521.

Samples were prepared using Microarray-based Genomic Selection as outlined in Okou et al [66]. Briefly, genomic DNA was amplified using whole genome amplification and then fragmented using a sonicator to 200-600 base pair fragments. The ends were

then repaired and an A-tail was added on the 3' ends. Unique adaptors were then ligated to the 3' overhangs. Ligated fragments were hybridized to a custom MGS array composed of 2.1 million features for 60 hours. Fragments that did not bind to the array were washed using a series of washes produced by NimbleGen Systems, Inc. The array was then placed on the HybriWheel at 95ºC and fragments that were bound to the probes were eluted. These fragments were then amplified in one round of PCR using the unique adapters as primers.

NimbleGen's Comparative Genomic Sequencing protocol was used. Exon sequences were tiled across 27 2.1 million-feature arrays containing probes approximately 25 basepairs long. Briefly, 1 μg of sample was denatured at 98ºC for 10 min in random primer buffer and labeled in the dark with Cy3-9mer primers (TriLink BioTechnologies) in the presence of dNTP mix and 100 units of Klenow (50U/μl, NEB) for 2 hours. Labeled samples were purified using ethanol precipitation method, dried down to the pellet and rehydrated in water. Ten to thirty micrograms of labeled DNA was mixed with NimbleGen's Hybridization cocktail and denatured at 95ºC for 5 min. The arrays were loaded and incubated overnight at 42°C on MAUI Hybridization System (BioMicro). The signal was detected by measuring Cy-3 chrome fluorescence using a 2 micron scanner at NimbleGen, and the sequence was inferred using NimbleScan software.

All analyses were performed using the R software package.

FIGURE LEGENDS

Figure 2.1. Basecalling metrics. The percent of bases called in a total of 7066 exons. 30% of the targeted exons had over 90% basecalling while 74% of the targeted exons had over 70% basecalling.

Figure 2.2. Comparing MGS probes that failed or succeeded for a) GC content (mean failed = 71.4%, mean succeeded = 41.9%), b) Probes per exon (mean failed = 32.9, mean succeeded = 41.9%), and c) Probe melting temperature (Tm; mean failed = 91.2%, mean succeeded = 79.1%). All p values $< 2.2 \times 10^{-16}$. Failure or success was determined by basecalling rates across the targeted exon. "Probes per exon" reflects the length of the targeted exon as probes were evenly spaced across targeted sequence.

Figure 2.1

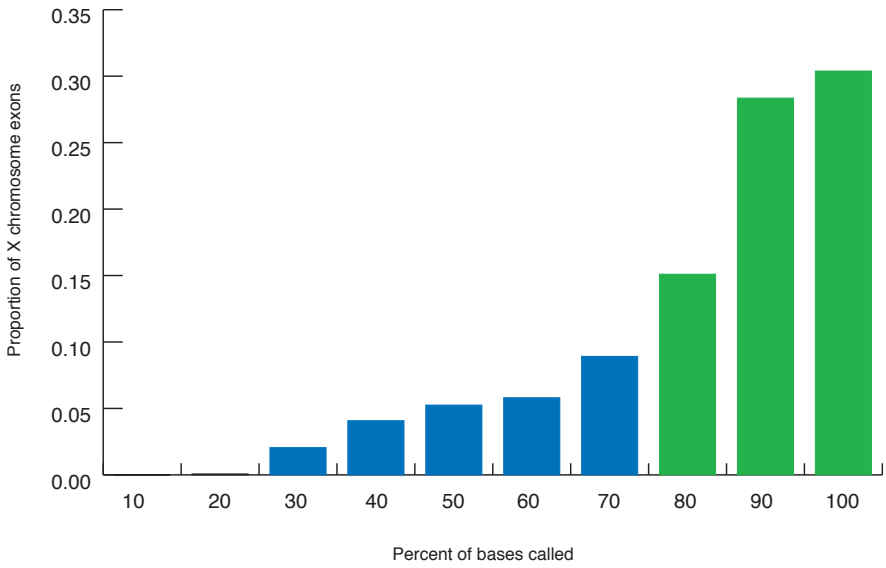
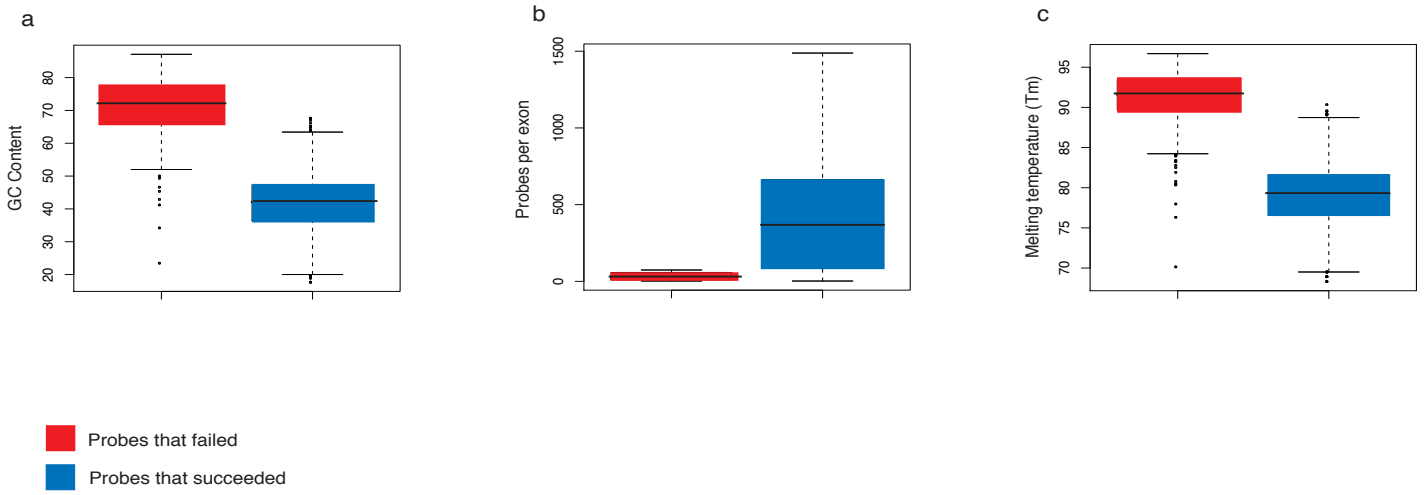

0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
Proportion of X chromosome exons
10
20
30
40
50
60
70
80
90
100
Percent of bases called

Figure 2.2


a
GC Content
20
30
40
50
60
70
80
b
Probes per exon
0
500
1000
1500
c
Melting temperature (Tm)
70
75
80
85
90
95
Probes that failed
Probes that succeeded

CHAPTER 3

INTRODUCTION

Autism spectrum disorder (ASD) is a highly heritable pervasive developmental disorder that affects four times as many males as females. This epidemiological observation is consistent with that expected of a disorder influenced by X-linked susceptibility alleles. Because males only have one X chromosome, any recessive alleles that are hidden maternally are revealed in male offspring. Providing further support for this hypothesis, candidate gene studies have identified a number of X chromosome loci that contribute to ASD susceptibility [91-95].

Association of autism and neuroligin pathway genes

Two candidate genes recently associated with ASD are the neuroligin genes, NLGN3 and NLGN4X on Xq13 and Xp22.3, respectively (see Table 3.1 for a complete list of published mutations). In two multiplex families from Sweden, Jamain et al [94] identified a mutation in NLGN3 leading to a change from a conserved arginine residue to a cysteine and a frameshift mutation in NLGN4X leading to a premature stop codon in two sets of affected sibpairs. Laumonnier et al [96] also demonstrated that a deletion causing a premature stop codon in NLGN4X was associated with ASD in a French cohort, and Yan et al [97] demonstrated that missense mutations in NLGN4X were associated with ASD in a mixed American and Portuguese Caucasian cohort. Mutations in NLGN3 and NLGN4X that have been found in autism patients lead to the retention of neuroligin proteins in the endoplasmic reticulum and, therefore, less protein on the surface of the cell [98]. Additionally, cells expressing the mutated form of NLGN3 had significantly decreased binding to neurexin1β and syntrophin-γ2, another scaffolding

protein in the post-synaptic domain [99,100], respectively. On the other hand, a de novo mutation in the promotor region of NLGN4X in a patient with autism and nonsyndromic profound MR increased gene expression [101]. These data strongly support the hypothesis that altered binding and synapse function can contribute to ASD.

Because the interactions between neuroligins and their binding partners are so critical to synaptic function, mutations in genes encoding neurexin and Shank3 may play a role in ASD. Neurexins are composed of three autosomal genes that are extensively alternatively spliced; neurexin1β which directly binds to the neuroligins is located on chromosome 2p16.3 [102,103]. Missense mutations in the neurexin1β (NRXN1β) gene were associated with ASD in a population of American patients [102]. Recent studies of copy number variation (CNV) in ASD patients found a deletion of coding exons from NRXN1 in multiple, independent probands and a set of affected sibpairs [104-106]. Results from another CNV study have demonstrated the significant contribution of rare exonic deletions of NRXN1 to ASD risk [106]. It appears that disrupting NRXN1 can lead to multiple cognitive phenotypes; deletions and duplications are associated with schizophrenia [107,108]. Additionally, linkage and gene expression studies suggest that the Contactin Associated Protein-like 2 (CNTNAP2), which is a member of the Neurexin family, is associated with autism and language disorders [109]. The anchoring protein, Shank3, is encoded by a gene located on 22q13.3, a region that is often associated with microdeletion syndrome [110]. Mutations and deletions in the SHANK3 gene have also been found in multiplex families with ASD [111,112]. There is substantial evidence that genes in the neuroligin pathway, which regulates synapse function and development, contribute to ASD. Examining genes in this pathway provides an opportunity to look at

sequence variants in multiple, related genes to look for patterns among affected individuals to identify susceptibility variants and possible genetic modifiers.

Evolutionary history of neuroligins

There are 5 total NLGN genes in the human genome. NLGN4X lies within a region of the short arm that is homologous to a single block of chicken 1q, and NLGN3 is located on the long arm in a region almost completely syntenic to chicken 4p [39]. It has been hypothesized that human Xq represents a conserved region descended from the mammalian proto-X chromosome while the syntenic region from chicken 1q represents sequence that was added via translocation before the eutherian radiation (*ibid*). NLGN4X has a single copy Y homologue at Yq11.221 and is approximately 97.5% similar at the amino acid level [94]. Like its homolog, NLGN4Y is expressed in the fetal brain, brain, prostate and testis [113], and expression in male brains is similar to NLGN4X expression [94]. NLGN4X was part of the pseudoautosomal region of the X chromosome; however the pseudoautosomal region has been moving in a stepwise manner distally through the short arm. It has been estimated that movement that created the region encompassing NLGN4X occurred 38-44 Myr ago via a series of inversions [39]; these dates correspond to the divergence of Old World from New World Monkeys.

Neuroligins are composed of a highly conserved esterase domain with a variant transmembrane region [114]. They are postsynaptic cell adhesion proteins, anchored in the postsynaptic density by Shank3 proteins, and they bind to the presynaptic protein neurexin-1β [115-120] (Figure 3.1). Binding between neuroligins and neurexins is controlled via alternative splicing mechanisms in both genes [121]. Neuroligins are necessary for the

maturation and function of proper synaptic activity [122], and some evidence suggests that expression of neuroligins is essential for the balance between excitatory and inhibitory synaptic activity. Mutations in neuroligin genes that result in altered levels of protein may change global circuitry and neuronal connectivity which could be a major contributing factor to the ASD phenotype [123].

Mouse models

A loss of function mutation in the murine ortholog of NLGN4X results in mice that have reduced reciprocal social interactions and communication similar to that of human ASD [124]. The R451C mutation in NLGN3, however, behaves as a gain of function mutation in the mouse causing impaired social interactions but enhanced spatial learning [125]. Deletion of this locus did not result in these behaviors, and the gain of function mutation was associated with an increase in inhibitory synaptic activity.

Further evidence of Xp22.3 involvement in cognitive disorders

Perhaps the most convincing evidence of NLGN4X involvement in ASD comes from studies on females with ASD or autism-like behavioral phenotypes. Three out of eight females with deletions at Xp22.3, which is the region containing NLGN4, showed features of autism suggesting a possible haploinsufficiency [92]. In a family with an interstitial deletion within Xp22.2-22.3, encompassing NLGN4X, the two offspring (one male and one female) had autistic like behavior and severe cognitive defects [126]. The female phenotype was variable (i.e. the mother was unaffected and the daughter affected) suggestive of skewed X inactivation. Skewed X chromosome inactivation is common in

female X-linked mental retardation patients, and females with ASD were shown to have higher rates of skewed X chromosome inactivation than normal sibpairs [127]. Additionally, in a study of females with ASD, 15% of X-linked genes, including NLGN4X, escaped X inactivation [128]. It has also been demonstrated that Turner syndrome females that maternally inherited their X chromosome had more social-cognitive defects (especially in verbal tests), and Turner syndrome females that also had ASD all inherited the X maternally [129]. These results suggest that in females with ASD portions of the X inherited maternally that may contain recessive, ASD contributing alleles might be escaping X inactivation. In males, these ASD susceptibility alleles of NLGN4X may contribute to disease phenotype due to recessivity.

In light of the positive associations of the neuroligin pathway genes with ASD, there is still considerable debate as to the contribution of mutations in these genes to ASD as many studies failed to find associations of NLGN3 and NLGN4X mutations in individuals with ASD [130-134]. Some of these studies were only searching for previously reported mutations while others only sequenced the coding exons of these regions. Many of the studies suffer from small sample sizes, often one or two families. Studies that focused on common SNPs likely overlooked rare variants that underlie complex traits such as ASD. Sequencing is the ideal technology to identify these rare variants in addition to common variants in a large sample population.

These data taken all together suggest a hypothesis for at least some classes of ASD, mainly, that disregulation of the neuroligin pathway is important in the susceptibility to ASD. The experiments in this chapter use Microarray-based Genomic Selection (MGS) to select the unique coding and non-coding sequences from NLGN3,

NLGN4X, NRXN1β and SHANK3 in males with ASD from multiplex families with 2 or more affected males and control (unaffected) males. Selected, enriched DNA was then sequenced using Affymetrix resequencing arrays (RAs). Okou et al [66] demonstrated that high quality DNA can be generated from MGS arrays as well as a high accuracy of sequences generated from the Affymetrix 300kb RA. Because previous attempts to identify common alleles with large effects in these regions have largely failed, we have focused on detecting relatively rare susceptibility variants with large effects. Sequencing both the coding and non-coding DNA, rather than limiting efforts to coding exons, increases the probability of finding rare alleles that may contribute to this complex disorder.

RESULTS

Sample preparation by MGS and sequencing on resequencing arrays was completed for 64 cases and 64 controls for a total of 38.4 MB of sequence generated. The total average basecalling rate was 90.9%. The following analysis focuses on NLGN3 and NLGN4X for the first 21 cases and 20 controls that had the highest basecalling rates (average basecalling rate=98.9%). Single base variants were annotated using a novel bioinformatics algorithm, SeqAnt. This software identifies the genomic position, amino acid change, dbSNP identification, phastCon score [135] and PANTHER score [136] for all variant bases in a given data file when compared to the reference file. The phastCon program assigns a score based upon the level of conservation of that particular base across a multiple alignment of 44 species. PANTHER is a database that scores protein changes at a specific location using information from homologous sequences to

annotate the overall biological function within the context of the functional divergence and molecular properties of that residue using a library of Hidden Markov Models (HMMs) [137]. One can then infer whether the amino acid change is more likely to change the protein function based on these scores.

Most variants are rare [8] and resequencing technology is ideal to find these variants. First, variants were partitioned into functional classes (UTR, silent, replacement, intron, intergenic) and compared within and between these classes (see Table 3.2). Replacement sites are predicted to be the most rare and show the least polymorphism, while silent sites are predicted to be more common and exhibit high levels of polymorphism [138]. Nucleotide diversity (θ) and average heterozygosity per site (π) was calculated for each variant within these classes to characterize any heterogeneity in diversity by functional class. It is expected that increased nucleotide diversity will correspond to a decrease in functional constraint.

Under the infinite sites neutral allele model most variants are expected to have frequencies less than 10% [139]. To test this model, the allele frequency spectrum was examined for all functional classes together as well as for each individual functional class. Tajima's D, a test statistic used to determine if the distribution of alleles fits the infinite sites model, was used. It is expected that Tajima's D will overall be negative indicating an excess of rare alleles, based on previous studies of human variation on the X chromosome [138], but will vary between functional classes depending on functional constraints. For example, replacement sites are predicted to have more rare variants than silent sites. Additionally, an excess of rare replacement alleles among affected cases

compared to controls would be consistent with the hypothesis that these mutations decrease gene function [5].

If rare sequence variants contribute to ASD these variants should be significantly more common among individuals with ASD than those without ASD whereas alleles found in both cases and controls are likely to be neutral. I compared whether the SNPs found in ASPs are found in dbSNP or the control males; if the variants found in ASPs are deleterious they are not expected to be found in dbSNP or control males.

Overall, sequences from both NLGN3 and NLGN4X have an excess of rare segregating sites as indicated by negative Tajima's D in all functional categories. Although we expect negative D for the X chromosome [138], θ is significantly higher for all functional categories than expected in both cases and controls (see Table 3.3). These data suggest a high rate of false positive SNP calling.

To identify the source of the false positives I examined the individual RA fragments that contained the most variants. An RA fragment is a fragment of DNA sequence containing unique sequence that is used for creating oligonucleotide probes. These fragments are similar to the fragments provided to NimbleGen for MGS array design. Therefore, I was examining both the success of the RA as well as the efficiency of MGS in target DNA selection. By using the BLAT function on the UCSC Genome Browser I was able to align fragments to the rest of the human genome and determine the percent identity (in other words, amount of sequence homology) between that fragment and other genomic loci.

As seen in Table 3.4, when the sequence of an RA fragment was not highly homologous to the other neuroligin genes, fewer variants were identified and none of

those variants were validated as being the paralogous allele. However, when the sequence of an RA fragment was highly homologous to the other neuroligin genes, many variants were detected and a substantial fraction of those variants were confirmed to be the paralogous allele. None of the exonic variants identified in NLGN3 were validated and only 3 of the remaining variants in NLGN4X were validated. Independent Sanger sequencing confirmed one replacement SNP and two silent variants; however, the replacement site is in a control (Table 3.4).

DISCUSSION

We sequenced the unique coding regions of four genes in the neuroligin pathway that have been previously associated with ASD: NLGN3, NLGN4X, NRXN1β and SHANK3 using RAs. A preliminary bioinformatic analysis indicated that variants identified in the neuroligin genes could be due to cross-hybridization of sequence from paralogous genes to the MGS array, which resulted in false positive SNP calling. Independent ABI sequencing validated only 3 variants in NLGN4X; one replacement SNP in a control male and two silent SNPs in both cases and controls.

These experiments highlight the limitations of genomic selection technologies. Highly homologous sequences will be difficult to accurately select and sequence due to the constraints of probe design and construction as well as the methods to elute fragments off of the arrays. In this experiment, fragments were eluted off of the array using heated water after a series of stringent washes. Probes, therefore, must have an optimal melting temperature (Tm) across all of the targeted sequences for successful elution. This limits the length and uniqueness of probes; in this experiment probes were between 50 and 93

bases long. However, there are stretches of sequence much greater than 93 bases in both of the X-linked neuroligin genes that are homologous to each other as well as NLGN4Y.

NimbleGen currently uses a sodium hydroxide elution method for their capture arrays, which eliminates the limitations of Tm on probe length. With longer probes it may be possible to select sequences from paralogous genes for downstream sequencing applications, depending on the degree of homology. A recent study demonstrated that short sequence reads from highly homologous sequences could accurately be assembled for copy number variation detection [140]. Yet, it is unclear if single base variants could be correctly identified using this algorithm.

Due to the limitations of microarray-based genomic selection, it may be more useful to explore technologies such as RainDance, a microdroplet, emulsion PCR technology that uses a library of unique PCR primers that can be processed in a single tube in a multiplexed fashion (www.raindancetechnologies.com). This technology eliminates the issues associated with multiplexed hybridization and PCR and may be able to more accurately select targeted sequence from paralogous genes.

METHODS

Sample selection

The Autism Genetic Resource Exchange (AGRE) collection is publicly available to the scientific community and contains genotype, phenotype and pedigree data on 830 affected families. The 830 families can be divided into three categories: those that have two or more affected female sibpairs (56), those with two or more affected sibpairs that are different sexes (223), and families with two or more affected male sibpairs (551). I

will exclude monozygotic twins (46) and those individuals that have diagnoses from known etiology (e.g. cytogenetic abnormalities, Fragile X) (4). Of the 501 families that have two or more affected male sibpairs (ASPs) that are not MZ twins, 257 families have been genotyped at markers DXS9895 (Xp22.3) and/or DXS9902 (Xp22.2) near NLGN4X and 244 have not been genotyped at these markers (Table 3.5).

Male ASPs were chosen as sample cases to test the hypothesis that there are X-linked susceptibility factors. Since two of the genes I am examining are X-linked, I restricted the samples to male ASPs to test the hypothesis that variants in these X-linked genes contribute to ASD. I chose only male individuals that come from families that have two or more male ASPs that share identical maternal X chromosome markers near NLGN4X (n=101) to test the hypothesis that there are variants (common and/or rare) in the NLGN4X genomic region that contribute to ASD susceptibility. If the two male siblings do not share the same genotype at these markers, ASD cannot be attributed to identical variants in NLGN4X. On the other hand, male sibpairs that share the identical genotype at the two markers may share ASD susceptibility variants in NLGN4X. I chose to use the markers surrounding NLGN4X rather than NLGN3 for two reasons: (1) NLGN4X is a better candidate gene, and (2) the distribution of markers around NLGN4X allows for a greater confidence of obtaining 2 male ASPs with identical genes.

I focused on the 101 male ASPs with identical genotypes at DXS9895/9902. One male was randomly be chosen for resequencing if both affected siblings are equally affected; if they are not equally affected, I chose the male with autism, not quite autism (NQA) or broad spectrum in that order to maintain consistency. Control males were randomly selected from the population of unaffected fathers of affected male sibpairs.

Array design

I designed a 385,000 feature NimbleGen Genomic Selection to isolate target DNA. The array contained all of the unique sequence for NLGN3 and NLGN4 and the coding regions of NRXN1β and SHANK3. Alternative splice sites for all four genes were included on the array. The latest build of the human genome on UCSC (March 2006, Build 36) was queried for unique sequence in NLGN3 and NLGN4X using the RepeatMasker function [141]. Then, the coding sequences for NRXN1β and SHANK3 were obtained using the Table Browser interface set to output only exon sequences padded with approximately 500 base pairs on the 3' and 5' ends. Once FASTA files for these four genes were generated from UCSC, a program containing a bioinformatics algorithm for chip design created by Viren Patel (available at https://hgxserver.genetics.emory.edu/zwicklab/tiki-index.php) generated fragments that are greater than 100 base pairs long. Fragments generated by this program are composed of unique sequence flanked by 50 base pairs of repeat sequence on the 3' and 5' prime ends. The coordinates of the fragments as well as the FASTA file were then sent to NimbleGen array designers for engineering. Capture probe sequences include both the forward and reverse strands, and oligonucleotides are between 50 and 93 basepairs long.

I designed an 8μm, 300kb Affymetrix resequencing array (RA) that includes NLGN3, NLGN4X, NRXN1β and SHANK3. Chips were designed to include all of the unique sequence in the transcript for the X linked genes (NLGN3 and NLGN4X) and the coding regions, plus all alternative splice sites, of the autosomal genes (NRXN1β and

SHANK3). Sequences for this array design were generated identically to those for the genomic selection array. However, fragments of unique sequence generated for chip design had to be greater than 50 base pairs and were only flanked by 12 base pairs of repeat sequence on the 3' and 5' ends. Preliminary studies using RAs reliably resequenced 32 autosomal and 9 X-linked regions composed of approximately 50kb of unique sequence each [56].

Target DNA selection and resequencing

Target DNA from AGRE samples was prepared using the Microarray-based Genomic Selection protocol from Okou et al [66] (Chapter 1). Resequencing arrays were hybridized using the standard Affymetrix Chip Hybridization protocol. All basecalls were made with the RATools program RA_PopGenCaller (http://www.dpgp.org/).

Variants were confirmed using independent Sanger sequencing. Briefly, unique primer sets were designed using a bioinformatics algorithm developed by David Cutler (https://hgcc.genetics.emory.edu/~ashetty/PrimerPicker.html) that allows the user to specify primer size, GC content, melting temperature, genomic interval of interest and desired amplicon size. A list of primers used can be found in Table 3.6.

All primers were obtained from Invitrogen Corp. and resuspended to a final concentration of 40 nM. Approximately 400 ng of genomic DNA was combined with 1ul each of the forward and reverse primer, 8ul of dNTPs Mix (TaKaRa), 5ul of 10 LA Buffer (TaKaRa), 1.5ul of LA Taq (TaKaRa) and VWR water to 50ul. PCR reactions were run at the following parameters: 94C for 4 minutes; 30 cycles of 94C for 20 seconds, annealing temperature of primers for 1 minute then 72C for 1 minute; 72C for 5

minutes followed by 4C hold. Annealing temperatures of the primer sets were determined by optimizing the primers using gradient PCR with a range of annealing temperatures to obtain a single band when run on a 1% TAE gel. Products were sent to Agencourt Biosciences for sequencing.

Analysis

Analyses were performed using the Popgen software, SeqAnt, Microsoft Excel and the R software package.

Table 3.1. Published Variants in NLGN3 and NLGN4X

Gene	Mutation	rs Number	Reference
NLGN3	R451C		[94]
NLGN3	222C>T; Y74Y		[134]
NLGN3	2189G>A; T632A		[131]
NLGN3	G(813-41)A		[131]
NLGN3	G(1148+65)A	rs2233441	[131]
NLGN4X	5.8Mb deletion Xp22.33-p22.31		[108]
NLGN4X	deletion of exons 4-6		[142]
NLGN4X	1186insT; premature stop D396X		[94]
NLGN4X	1253del(AG); premature stop D429X		[96]
NLGN4X	G99S		[97]
NLGN4X	K378R		[97]
NLGN4X	V403M		[97]
NLGN4X	R704C		[97]
NLGN4X	deletion of exon 4		[128]
NLGN4X	A558 (Syn)		[143]
NLGN4X	5.5Mb deletion Xp22.31-p22.13		[144]
NLGN4X	G34C	rs2290488	[131]
NLGN4X	1397C>T; T311T	rs7049300	[131]
NLGN4X	2241C>T; L593L	rs3747333	[131]
NLGN4X	2243C>G; L593L	rs3747334	[131]
NLGN4X	G(1-54)A	rs2290487	[131]
NLGN4X	335G>A		[101]
NLGN4X	1597A>G; K378R		[145]
NLGN4X	R87W		[146]
NLGN4X	7.7Mb deletion Xp22.2-22.3		[126]

Table 3.2. Number of segregating sites by functional category for NLGN3 and NLGN4X in cases and controls

	NLGN3 Cases	**NLGN3 Controls**	**NLGN4X Cases**	**NLGN4X Controls**
Total Segregating Sites	148	134	1048	852
UTR	8	4	29	19
Silent	**8**	**2**	**26**	**22**
Replacement	**10**	**8**	**14**	**11**
Introns	119	115	959	788
Intergenic	3	5	20	12

Table 3.3. Comparing nucleotide diversity between cases and controls and between functional classes[a].

	Expected Value[b]	**NLGN3 Cases**	**NLGN3 Controls**	**NLGN4X Cases**	**NLGN4X Controls**
Total Nucleotide Diversity, θ (x10^{-4})	5.32 ± 1.82	15.83 ± 9.22	12.62 ± 7.25	15.83 ± 9.22	12.62 ± 7.25
UTR	5.64 ± 2.70	20.64 ± 14.70	9.78 ± 9.63	27.68 ± 18.97	19.89 ± 14.27
Silent	**9.80 ± 5.00**	**41.70 ± 9.21**	**15.39 ± 4.53**	**170 ± 85.4**	**113 ± 21.45**
Replacement	**1.86 ± 1.07**	**12.11 ± 10.13**	**10.62 ± 9.73**	**24.05 ± 7.85**	**7.37 ± 1.47**
Introns	5.29 ± 1.83	19.90 ± 9.22	22.85 ± 11.15	14.66 ± 9.11	11.66 ± 7.26
Intergenic	6.14 ± 2.26	15.53 ± 9.23	12.08 ± 7.04	17.42 ± 8.45	14.45 ± 6.90

[a]Total nucleotide diversity defined as Watterson's[147] estimate of $\theta = 4N\mu$, where N is the population size and μ is the per-site, per-generation mutation rate.
[b]Expected value based on sequence data from 8 X-linked loci in 40 unrelated, unaffected males.

Table 3.4. Analysis of exonic segregating sites in NLGN3 and NLGN4X from RA data

Gene	Fragment Start	Fragment End	Size (bp)	NLGN3[a]	NLGN4X[a]	NLGN4Y[a]	Number of Exonic Segregating Sites	Validated as paralogous allele	Validated as true variant
NLGN4X	5820806	5822081	1275	78.9	-	96.5	24	7	0
NLGN4X	5831028	5833725	2697	81.1	-	92.2	30	8	2
NLGN4X	6078931	6079987	1056	0	-	92	2	0	1
NLGN3	70283008	70285498	2490	-	0	0	8	0	0
NLGN3	70289365	70292970	3605	-	0	0	1	0	0
NLGN3	70300372	70302095	1723	-	71.2	71.2	3	0	0
NLGN3	70303052	70304637	1585	-	78.7	80.3	4	0	0
NLGN3	70305481	70306725	1244	-	89	87.1	8	0	0

[a]Values represent percent identity to gene

Table 3.5. AGRE Sample selection

Table 2: Selection of Male ASPs from the AGRE Collection	
Male ASPs with Identical Genotypes at DXS9895/9902	**101**
Male ASPs with Different Genotypes at DXS9895/9902	156
Male ASPs not Genotyped at DXS9895/9902	**244**
Male ASPs with Cytogenetic Abnormalities or MZ Twins	50
Total Male ASPs	**551**

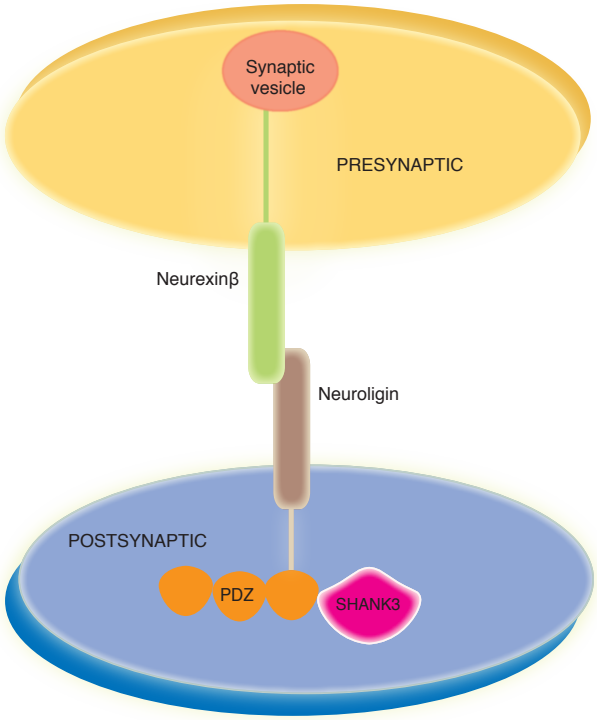
Table 3.6. Primers for independent ABI validation sequencing

ID	Primer Sequence
3.1_FWD	5'-CCTTTCTGAAGCTGTGGTGCTTG-3'
3.1_REV	3'-CGTTGGGCTCCTGGATGTAAGTAG-5'
3.2_FWD	5'-TGCTTCTAACCATCACTCTGC-3'
3.2_REV	3'-GCTCGCTACCTACCTCCTTTC-5'
3.3_FWD	5'-AGGACTTAGCGGATAATGACG-3'
3.3_REV	3'-GTAGCCTGTCAGTCCGTGAAC-5'
3.4_FWD	5'-CGTCATCACCCAAATCCTCCATCC-3'
3.4_REV	3'-TGTGACAACGTGAGGAGGCTG-5'
3.5_FWD	5'-TGTGCTGGACACCGTGGATATG-3'
3.5_REV	3'-GCGTAGAAGTAGGTAGGCGAGC-5'
3.6_FWD	5'-TCGGGCTGAAACCAAGGGTC-3'
3.6_REV	3'-ACGGTAGTAGAGGGCAGCGAAG-5'
3.7_FWD	5'-TCTACTACCGTAAGGACAAACGG-3'
3.7_REV	3'-GCTCTTAGCACTCATGGGTTG-5'
3.8_FWD	5'-CGCTGCCCTCTACTACCGTAAG-3'
3.8_REV	3'-AGCCTGGAGATTGGCTGTGC-5'
3.9_FWD	5'-CGCCTACCTACTTCTACGCCTTC-3'
3.9_REV	3'-TCCTTCCTAGTCCCGATGCTAAC-5'
4X.1_FWD	5'-GGTAGGGCAGAGGGATAGGAAGG-3'
4X.1_REV	3'-CGTCGCTCCTCTTCCTCAACATC-5'
4X.2_FWD	5'-GCGTCATAAGTGGGATGTCATCTGG-3'
4X.2_REV	3'-CAAACGCCCAGCAATCACTCC-5'
4X.3A_FWD	5'-AGGTGAGCCTCAGTGTGTCG-3'
4X.3A_REV	3'-AGGTTCCTCCACCAGACATGAC-5'
4X.3B_FWD	5'-GTGATTGCTGGGCGTTTGGTG-3'
4X.3B_REV	3'-TGGAGCAAATCAGTCCTGGATGAG-5'
4X.4_FWD	5'-CCAACACGAAGATGAACGTACCCAG-3'
4X.4_REV	3'-ACTTCTCCGTGTCCAACTTCGTG-5'
4X.5_FWD	5'-GAGGACACAAACAAGTGGCAAGG-3'
4X.5_REV	3'-AGCTCATCCAGCAGACCATCAC-5'
4X.6_FWD	5'-GTGGCAAGGATAGTGATACCC-3'
4X.6_REV	3'-AAGTACACTCGGATATTGGCAG-5'
4X.7_FWD	5'-ATCTCTCTACACACTGCACAAGAGG-3'
4X.7_REV	3'-TGGAGAGAGGCGGTTTCAGC-5'
4X.8_FWD	5'-TTGTAGTTCTTGTTCCGCAGG-3'
4X.8_REV	3'-AGCAAAAACGACACTAAATTGTGG-5'
4X.9_FWD	5'-AGTAAGGATCTCTCATCCAGGTG-3'
4X.9_REV	3'-GTTCTAGGTGGTCGTTGTGTG-5'
4Y.1_FWD	5'-CACCTCACTTAGACAGCTTCGG-3'
4Y.1_REV	3'-GTAGCATTTCGGATGCCAGTC-5'
4Y.2_FWD	5'-TCACATGCTTGCGAAACAATC-3'
4Y.2_REV	3'-CGTCACACCGTCCTCGTTATC-5'
4Y.3_FWD	5'-TCACATGCTTGCGAAACAATC-3'
4Y.3_REV	3'-GATGCCGAAGACATAGGGGAC-5'
4Y.4_FWD	5'-ATCCAAACCAACCAGTTCCTC-3'
4Y.4_REV	3'-GGTGTTTGGCGTCATAAATGG-5'
4Y.5_FWD	5'-CACAACTTGAACGAGATATTCC-3'
4Y.5_REV	3'-CCTAGTCACACGAAACATCTG-5'

FIGURE LEGEND

Figure 3.1. Neuroligin interacting with neurexin in the synaptic cleft. Neuroligins are anchored in the post-synaptic density by Shank3 and bind to the pre-synaptic protein, Neurexin 1β.

Figure 3.1


Synaptic vesicle
PRESYNAPTIC
Neurexinβ
Neuroligin
POSTSYNAPTIC
PDZ
SHANK3

CHAPTER 4

INTRODUCTION

In the previous chapter I described an experiment in which I selected target DNA from genes within the neuroligin pathway using Microarray-based Genomic Selection for downstream resequencing on arrays. Due to the high sequence homology of these genes, MGS selected target DNA from both loci resulting in a high rate of false positive SNP detection. Additionally, sequencing technology improved almost exponentially in the time since the start of the experiment. For these reasons, I chose to use Illumina paired-end multiplexed sequencing. Multiplexed sequencing allows one lane of the Illumina flow cell to be "shared" by 12 different samples [61]. During the amplification step prior to cluster generation, a unique 6 base tag is attached to each of 12 samples (Figure 4.1). These 12 samples are then pooled in equimolar concentration and added to one lane of the Illumina flow cell. After sequencing, the sequences from each sample are identified by their 6 base tag and parsed for downstream alignment and assembly. For targeted resequencing applications, sufficient depth of coverage can be achieved even when split among multiple samples.

In this study I sequenced the exons, 5'UTR, 3'UTR and a portion of the surrounding intronic and intergenic regions of NLGN3, NLGN4X and NRXN1β in 144 males with ASD from the Autism Genetic Resource Exchange (AGRE) collection. These males are from families with 2 or more affected male sibpairs and share identical sequence from the Xp22.3 region with their affected brothers. Using this selection criteria I am able to specifically test the hypothesis that maternally inherited variants at Xp22.3 contribute to autism in males but not in their mothers due to recessivity. I am

also able to test the hypothesis that rare variants in genes from this important neurological pathway contribute to autism in males.

RESULTS

Evaluation of Mapping and Assembly Algorithms

I sequenced approximately 85kb of long-range PCR products from 144 samples using Paired-End Multiplexed Sequencing on the Illumina Genome Analyzer II generating a total of over 12.2Mb of sequence. Sequences were parsed by their 6 base index tag to identify individual samples from multiplexed lanes.

To identify the best mapping, assembly and SNP calling pipeline, I analyzed my data with three different open source programs: MAQ [148], Bowtie [149] and BWA [150]. MAQ is an algorithm that rapidly aligns short reads to a reference based upon *phred*-like quality scores using a hash-table based method. BWA is much like MAQ, although BWA is able to support gapped alignment for single-end reads and uses Burrows-Wheeler Transform [151] and backward searching for exact matches. Bowtie is an alternative algorithm that also uses Burrows-Wheeler indexing and backward searching. Aligned sequences were then assembled as shown in Figure 4.2; read statistics for two of the three alignments are in Table 4.1. BWA does not report the number of paired end reads that mapped. A paired end sequence is considered mapped when both sequences are mapped to a unique genomic location. Otherwise, it is considered unmapped. This stringent requirement accounts for the observed frequency of unmapped reads.

Consensus fasta files generated from all three pipelines were then analyzed using Popgen (v 2.0.5). I expected that approximately 5 out of every 10,000 bases on the

human X chromosome will be variant, and approximately 1 our of every 1,000 bases on human autosomes will be variant [138]. To assess the level of genetic variation observed in my samples, I first calculated Watterson's estimate of *theta* [147] and then compared the amount of variation detected in our sample population with this expected frequency. Substantial deviations from this expectation could reflect real biological differences or could arise as a consequence of assembly and SNP calling errors.

All of the pipelines had higher than expected amount of variation (Table 4.2). The BWA pipeline had nearly ten times the expected amount of variation. This high level of variation probably arises as a consequence of sequence mismapping and, for this reason, I decided to not to pursue any further analyses using BWA. The variation detected from the MAQ haploid pipeline (0.0008) was closer to the expectation (0.0005), but the diploid pipeline was substantially higher (0.03 compared to 0.001 expected). Again, this is likely caused by mapping errors with MAQ's alignment algorithm or poor SNP calling algorithm. The Bowtie pipeline generated 2-6 times more variation than expected; however, I believe that the Bowtie pipeline provided the best data compared to the other two pipelines. I chose to continue with Bowtie rather than MAQ based on the levels of variation detected for both haploid and diploid data as well as the fact that Bowtie uses a more sophisticated alignment algorithm than MAQ (as discussed in the Introduction).

To improve upon the MAQ SNP calling algorithm utilized by the Bowtie pipeline and attempt to reduce the number of false positives, I used a novel SNP calling algorithm. This approach uses a population genetics based framework based on the value of *theta*, $\theta = 4N_e\mu$, where N is the effective population size, and μ is the per site mutation rate (see

Methods). The assembled Bowtie alignments were then analyzed using this novel SNP caller. The resulting population genetic data that includes the number of segregating sites, Watterson's estimate of θ [147] and Tajima's D are presented in Table 4.3 for each gene region and each functional class. The new SNP caller has reduced the number of false positives as indicated by values of θ that are between 1.2 and 2 times the expected values. There are still likely some false positives, but the number of false positives is significantly less than the previous assemblies.

Annotation of Single Nucleotide Variants

I then annotated the single nucleotide variants using SeqAnt, a software program that compares sequences to the reference genome to determine if there is a change in amino acid, if the variant is in dbSNP, the evolutionary conservation (as determined by phastCon scores [135]), and relative impact of amino acid change (as determined by PANTHER scores [136,137]). I predicted that the best candidates for ASD susceptibility alleles are those variants found at highly conserved sites.

To determine if our approach identified the common variation contained within the sequenced genomic regions, I compared the number of SNPs called in this experiment that had already been catalogued in dbSNP (Table 4.4). If the average heterozygosity of the SNP in dbSNP was greater than 0.10, the SNP calling algorithm detected it in this experiment. If the average heterozygosity was less than 0.10 it was detected 45% of the time. These results suggest that this experiment successfully identified all of common variation in our samples. The only variants not detected were lower frequency ones that may not have been present in the samples that we sequenced.

I identified 99 total variants in NLGN3, 121 variants in NLGN4X and 260 variants in NRXN1β. The exonic variants and non-coding variants with high phastCon scores for the two haploid NLGN genes are in Table 4.5. Twelve out of the 99 NLGN3 variants are highly conserved (phastCon scores > 0.90). Three of these 12 intronic variants are common and found in dbSNP, while 9 are rare in this sample population (frequencies between 0.7% and 4.2%). I identified 4 exonic variants (3 silent and 1 replacement variants) that are not highly conserved and are found in dbSNP. Two out of the 121 NLGN4X variants are highly conserved (phastCon scores > 0.90) and are found in the 3'UTR and intron. The exonic and non-coding variants with high phastCon scores for NRXN1β are in Table 4.6. Fifty-one out of 260 NRXN1β variants are highly conserved; 11 of these 51 are common and found in dbSNP. None of the replacement variants are in dbSNP and range in frequency in this population from 0.7% to 11%. The SIFT (Sorting Intolerant from Tolerant) algorithm predicts that 5 of the 7 replacement substitutions would be damaging to the protein [152].

Indel Analysis

Each of the assembly algorithms described first map sequences against a know reference sequence, and then perform an assembly. Mapping a sequence against the reference sequence is performed under conditions that require at most a set number of mismatches. In the case of Bowtie, the default value used is 2. As a consequence, the assembly process will not properly identify indel variation in individual samples. Instead, we predict that sequences at the end of indels will map successfully and sequences that

span larger indels will fail to map (exceed 2 differences). Thus, accurately identifying indels requires a different procedure.

To remedy this and to enable accurate indel identification, I developed an alternate procedure. First, reads that aligned to each individual PCR amplicon from the Bowtie alignment of a given individual were identified. Sequences from each fragment were then separately assembled *de novo* using the Velvet algorithm [153]. Velvet utilizes de Bruijn graphs, which are representations based on short (25-50bp) *k-mers*, to construct highly accurate contigs. The contigs for an individual sample and the original reference sequence were then input into MUMmer [154,155] a program that rapidly aligns two DNA sequences using suffix trees to create a representation of the sequences. The contigs were then stitched together based on the MUMmer alignment generating a new sequence for that sample for that particular amplicon. The new sequences for all 144 samples were then individually aligned against the human genome reference sequence using ClustalW [156]. The GDE file containing the aligned sequences was then parsed to generate a list of all indels.

Previous studies on insertion and deletion rates in the human genome estimate that there should be approximately 1 indel every 10,000 bases [157]. Therefore, if I am sequencing approximately 85 kb per sample, each individual should on average have 8 or 9 indels, or I should identify approximately 1,100 total indels. I indentified almost 5,000 total deletions, or an average of approximately 34 deletions per individual. This is clearly much higher than expected and likely contains many deletions that are false positives due to mismapping of repetitive regions or low depth of coverage for that region. I identified

105 total insertions, which may be an underestimate of the actual number of small insertions (Table 4.7).

I then analyzed the 29 indels that were mapped to coding regions. Twenty deletions were identified in NLGN4X and 9 deletions were identified in NRXN1β; there were no coding insertions and no coding indels in NLGN3. After removing samples that had more than 1 deletion there were 4 samples with 4 different deletions (Table 4.8). Three of the deletions were in NLGN4X and 1 was in NRXN1β. These deletions all cause frameshifts and would truncate between 40 and 76% of the protein likely resulting in a deleterious phenotype.

DISCUSSION

In this experiment, I successfully sequenced all of the coding and a portion of non-coding regions from NLGN3, NLGN4X and NRXN1β from 144 affected individuals from the Autism Genetic Resource Exchange using paired-end, multiplexed Illumina sequencing. I identified a number of replacement and silent variants as well as many highly conserved, non-coding variants.

Non-coding regions, such as UTRs, may contain important elements for proper gene expression. For example, microRNAs may target 3'UTR sequences and cause translational repression or cleavage of target messages [158]. The highly conserved 3'UTR variant identified in NLGN4X lies within two predicted miRNA binding sites, hsa-miR-561 (predicted by miRANDA algorithm [159] and DIANA-microT [160]) and hsa-miR-1287 (predicted by MICROINSPECTOR [161]). However, these are not highly conserved and the base change does not dramatically change the amount of free energy. One of the highly

conserved 3'UTR variants identified in NRXN1β (chr2: 50002181) lies within four predicted miRNA binding sites, hsa-miR-518b, hsa-miR-518c, hsa-miR-518d-3p, and hsa-miR-518e. These are highly conserved and a fairly significant prediction based on the miRANDA algorithm [159]. Future experiments might include testing whether these miRNAs actually suppress expression.

This study highlights the importance of understanding population genetics when analyzing results from a deep sequencing experiment. By understanding the expectations of nucleotide diversity in different functional classes one can determine whether particular alignment and/or assembly algorithms are erroneous in basecalling. Additionally, it is important to understand the evolutionary conservation of a variant when pursuing validation strategies and functional assays. One can prioritize those variants that are highly conserved across mammals, vertebrates, etc. as ones that may be contributing to the disease phenotype.

Finally, the results from comparing the variants identified in this experiment with those already annotated in dbSNP suggest that much of the common variation in the human population has already been discovered. A corollary is that if common variation were contributing to complex diseases whole genome association studies would have already made a significant genotype/phenotype association. As discussed in the Introduction Chapter of this thesis, most association studies have failed to find alleles that significantly contribute to the genetic variance of any complex, common disease. It is more likely that rare variants in relevant genomic pathways contribute to complex traits such as Autism Spectrum Disorder. This study identified many rare variants in the neuroligin pathway genes that may contribute to ASD in males. Further validation of

these variants is ongoing in a large sample of unaffected males from the NIMH as well as the affected brothers and unaffected fathers from the AGRE collection. By obtaining an estimation of allele frequencies in the population I will be able to confidently assert that variants only found in the affected population are associated with the disease phenotype.

METHODS

Sample Selection

The Autism Genetic Resource Exchange (AGRE) collection is publicly available to the scientific community and contains genotype, phenotype and pedigree data on over 900 affected families. Males from families with 2 or more male affected sibpairs (ASPs) that either share identical X chromosome markers, DXS9895 and DXS9902, or shared greater than 98% of 52 genotyped SNPs in the Xp22.3 region were chosen. There were a total of 152 families that fit these criteria. One male was randomly chosen for resequencing if both affected siblings were equally affected; if they were not equally affected, the male with autism, not quite autism (NQA) or broad spectrum were chosen in that order to maintain consistency. A total of 144 samples were processed; 6 samples had global PCR failure while 2 were unavailable from AGRE at the time of this experiment.

Primer Design

Long-range PCR primers were designed using an in-house program developed by D. Cutler (https://hgcc.genetics.emory.edu/~ashetty/PrimerPicker.html). We selected primers based on the following parameters: length between 29-32 bases, GC content between 45% and 60% and melting temperature of approximately 68°C. Ideal fragments

were between 6 and 8 kb in length, although they ranged from approximately 2 kb to 12kb. Primers were tested in silico using the USCS PCR function as well as Amplify 1.3 software. All primers were obtained from Invitrogen Corp; a list of all primers used in this experiment can be found in Table 4.8.

Long PCR

DNA from the AGRE repository was aliquoted into a master plate; 5ul (approximately 500ng) of each sample was further aliquoted into PCR plates using the BioMek FX robot. One plate was used per fragment. To the DNA we added 1X LA Taq buffer (TaKaRa), 250 μM dNTP Mix (TaKaRa), 400nM of both forward and reverse LMPCR primers and 0.1 U/μl of LA Taq (TaKaRa). If the fragment had a high GC content we used 1X GC Buffer (TaKaRa) in place of 1X LA Taq buffer. PCR parameters were as follows for 29 cycles: 94°C for 2 min, 94°C for 10 sec, and 68°C for 1 minute per kb (of fragment) with a final extension time of 5 min plus the time at step 3 at 68°C.

Amplification was confirmed using 1% agarose 96 well E-Gels (Invitrogen). If a sample failed it was removed from the plate and re-amplified. If it failed again, it was eliminated from the experiment and noted. We determined the concentration of each fragment using PicoGreen dsDNA Quantitation Kits (Invitrogen) and the Tecan Ultra Evolution plate reader. An equimolar concentration of each fragment was then pooled by sample using the following formula:

$$pM = \mu g \times (pmol/660pg) \times (10E6pg/1\mu g) \times (1/N)$$

where N is the number of nucleotides and 660pg/pM is the average molecular weight of a nucleotide pair; and then:

Volume to pool = (lowest pM value)/ (pM/volume)

The total DNA concentration per sample was 10ug. Pooled amplicons were then purified using the Invitrogen PureLink PCR Purification Kit with the HC buffer.

Fragmentation

Pooled, purified samples were dried using a SpeedVac at 75°C for 45 minutes and then resuspended in 100µl of Tris EDTA and transferred into glass microtubes (Covaris). The samples were then sheared to approximately 300 bp using the Covaris E210 with the following parameters: Duty cycle of 20%, Intensity of 4, 200 cycles per burst for 60 seconds. Fragmentation was validated using DNA 7500 chips (Agilent Biosciences) and the Agilent Bioanalyzer software (see Figure 4.3).

End Repair

To convert the overhangs resulting from fragmentation into blunt ends, we performed end repair using the NEBNext DNA Sample Prep Reagent Set 1 (New England BioLabs) with 0.4mM dNTP mix (4), 5µl of T4 DNA Polymerase, 1µl of DNA Polymerase I (Klenow) fragment, 5µl of T4 Polynucleotide Kinase, and 1X T4 DNA ligase buffer. The reactions were incubated in a thermal cycler for 30 minutes at 20°C. Following incubation, the reactions were purified using a QIAquick PCR purification Kit (Qiagen).

Add "A" Bases to 3' End of DNA Fragments

To the purified, blunt, phosphorylated DNA fragments we added 1X NEB Buffer2, 1mM dATP (NEB) and 3µl of Klenow fragment (NEBNext Set 1). Reactions were incubated for 30 minutes at 37°C. Following incubation, reactions were purified using a QIAquick MinElute Kit (Qiagen).

Ligation of Adapters

To the DNA we added 1X Quick Ligation Buffer (NEBNext Set 1), 10µl of Index PE Adapter Oligo Mix (from the Multiplexing Sample Preparation Kit; Illumina) and 5ul of Quick T4 DNA Ligase. The reactions were incubated for 15 minutes at room temperature and then purified using the QIAquick PCR Purification Kit (Qiagen). This protocol uses a 10:1 molar Adapter:DNA ratio based on the starting concentration of DNA.

Size Selection and Enrichment

We used the Size Select 2% E-Gels (Invitrogen) to remove all unligated adapters and to accurately select the 300bp band. When the 300bp band was successfully removed it was then selectively enriched using PCR to amplify the amount of DNA in the library and attach the 6-base index tag into the adapter. To 10ul of DNA we added 1X Phusion PCR Master Mix (Finnzymes; NEBNext Set 1), 1µl each of PCR Primer lnPE 1.0 and PCR Primer lnPE 2.0 and 1µl of PCR Primer Index (from Mulitplexing Sample Preparation Kit; Illumina. PCR parameters were as follows for 30 cycles: 98°C for 30 sec, 98°C for 10 sec, 65°C for 30 sec and 72°C for 30 sec with a final extension time of 5

min at 72°C. Following incubation, samples were purified using a QIAquick PCR Purification Kit (Qiagen), and enrichment was confirmed using the Agilent BioAnalyzer and the Agilent 7500 DNA chip (see Figure 4.4).

Cluster Generation and Paired End Multiplexed Sequencing

Enriched DNA was denatured and diluted to a concentration of 4pM. Cluster generation was performed in 2 steps using the Paired-End Cluster Generation Kits v1 from Illumina and the following recipes: Amplification_only_v3 and PE_2P_R1prep_Linearization_CombinedBlocking_PrimerHyb_v2. The flow cell was then transferred to the IGAII for sequencing. Sequencing reagents from the Illumina SBS Sequencing Kit v2 were used with the following recipe: GA2_MP_36+7+36Cycle_v4. Images were transferred to the server for downstream basecalling using the Illumina Pipeline.

Data Analysis

Raw basecalling data generated by Illumina was used as input for the following mapping and alignment programs: MAQ, Bowtie and BWA. To assemble and generate a consensus sequence file MAQ and Bowtie use the assembly algorithm from the MAQ software while BWA uses the algorithm from SAMTools. The MAQ and Bowtie data were run with a prior heterozygosity probability value of zero (BWA does not currently employ a prior probability of heterozygosity).

The new improved basecaller created by D. Cutler is based on a population genetics based framework. For any individual's sequence read data let

N_A = Number of A calls
N_C = Number of C calls
N_G = Number of G calls
N_T = Number of T calls

So that $N = N_A + N_C + N_G + N_T$

And let e = the per read genotyping error

The probability of the genotype given the data must be greater than 0.95 to call the genotype.

$\Pr\{G \cap Data\} = \Pr\{G \mid Data\} \cdot \Pr\{Data\}$ which is equal to $\Pr\{Data \mid G\} \cdot \Pr\{G\}$

Therefore $\Pr\{G \mid Data\} = \dfrac{\Pr\{Data \mid G\} \cdot \Pr\{G\}}{\Pr\{Data\}}$

We assume that $\Pr\{Data \mid G\}$ is multinomially distributed so that

$$\Pr\{N_A, N_C, N_G, N_T \mid A\} = \binom{N}{N_A, N_C, N_G, N_T} \cdot P_A^{N_A} \cdot P_C^{N_C} \cdot P_G^{N_G} \cdot P_T^{N_T}$$

Where for A homozygotes

$$P_A = 1 - e$$
$$P_C = \frac{e}{3}$$
$$P_G = \frac{e}{3}$$
$$P_T = \frac{e}{3}$$

And for AC heterozygotes

$$P_A = 0.5 - \frac{2e}{3}$$
$$P_C = 0.5 - \frac{2e}{3}$$
$$P_G = \frac{e}{3}$$
$$P_T = \frac{e}{3}$$

$$\Pr\{Data\} = \sum_{\substack{\text{All} \\ \text{genotypes}}} \Pr\{\text{Data} \mid \text{G}\}$$

To determine the probability of the genotype ($\Pr\{G\}$) assume that in a population at neutral equilibrium if we sequence k alleles the probability of the genotype is based on the value of theta, $\theta = 2N_e\mu$, where N is the effective population size, and μ is the per site mutation rate. The basecaller uses this population genetics based Bayesian framework to make calls such that the probability that a site is *not* segregating is equal to

$$1 - \theta \sum_{i=1}^{k-1} \frac{1}{i}$$

and the probability that there are j copies of the minor allele (the minor allele count) is equal to

$$\theta \cdot \left[\frac{j}{k} + \frac{k-j}{k} \right]$$

For these analyses we set $\theta = 1 \times 10e^{-3}$

For haploids

$$\Pr\{G\} = \sum_{\substack{\text{All possible} \\ \text{minor allele} \\ \text{counts}}} \Pr\{\text{G} \mid \text{minor allele count}\} \cdot \Pr\{\text{minor allele count}\}$$

where $\Pr\{\text{G} \mid \text{minor allele count}\}$ is simply binomial sampling.

For diploids

$$\Pr\{G\} = \sum_{\substack{\text{All possible} \\ \text{minor allele} \\ \text{counts}}} \sum_{\substack{\text{Number of} \\ \text{heterozygotes} \\ \text{given the minor} \\ \text{allele count}}} \Pr\{\text{G} \mid \text{minor allele count}\} \cdot \Pr\{\text{minor allele count}\} \cdot \Pr\{\text{Number of heterozygotes} \mid \text{minor allele count}\}$$

where $\Pr\{\text{Number of heterozygotes} \mid \text{minor allele count}\}$ where is the Hardy-Weinberg exact probability [162].

The consensus sequences generated were analyzed separately using the PopGen program version 2.0.5 developed by D. Cutler and annotated using SeqAnt. All statistical analyses were performed using the R software package.

Table 4.1 Paired-End read statistics for Bowtie and MAQ Alignments

	Bowtie	**MAQ**
Mean Total Reads per sample	1,359,012	1,359,012
Mean Mapped Reads per sample	874,903	912,463
Mean Unmapped Reads per sample	484,109	446,549
Mean Percentage Mapped per sample	64.38%	67.14%
Mean Percentage Unmapped per sample	35.62%	32.86%

Table 4.2 Number of Segregating Sites, Theta and Difference from Expected values of theta for different pipelines for Haploid and Diploid Data

	Haploid (NLGN3 and NLGN4X)			**Diploid (NRXN1β)**		
	Segregating Sites	θ **(x10^{-4})**	**Fold Difference from Expected**	**Segregating Sites**	θ **(x10^{-4})**	**Fold Difference from Expected**
BWA/SAMTools	2519	71.7 ± 32.2	14	913	68.8 ± 28.6	7
MAQ/MAQ	289	8.2 ± 3.8	1.6	3663	276.3 ± 113.7	30
Bowtie/MAQ	454	12.98 ± 5.8	2	741	55.9 ± 23.1	6

Table 4.3 Variation in NLGN3, NLGN4X and NRXN1β by functional class

a) NLGN3

	Segregating Sites	Theta	Tajima's D
UTR	4	0.0004	-1.72
Silent	0	0	0
Replacement	0	0	0
Introns	74	0.0007	-0.52
Intergenic	21	0.0006	-0.27
All Sites	99	0.0006	-0.41

b) NLGN4X

	Segregating Sites	Theta	Tajima's D
UTR	14	0.0008	-0.36
Silent	3	0.001	-0.3
Replacement	1	0.0001	0.24
Introns	86	0.0006	-0.47
Intergenic	17	0.0006	-0.32
All Sites	121	0.0006	-0.41

c) NRXN1β

	Segregating Sites	Theta	Tajima's D
UTR	11	0.001	-0.63
Silent	8	0.005	-1.07
Replacement	7	0.001	-0.81
Introns	199	0.002	-1.03
Intergenic	35	0.002	-1.17
All Sites	260	0.002	-1.21

Table 4.4 Number of variants identified in this study that are also found in dbSNP.

	AVERAGE HETEROZYGOSITY		
	Greater than 10%	**Less than 10%**	**TOTAL**
NLGN3	8/8	35/82	43/90
NLGN4X	32/32	39/97	71/129
NRXN1	33/33	24/39	57/72

Numerator is the number of variants in this study that are also in dbSNP and denominator is the number of SNPs in dbSNP in the target region.

Table 4.5 Exonic and Highly Conserved Non-Coding Variants in NLGN3 and NLGN4X

Gene Name	Coordinate	Reference	Mutation	Position	PhastCon Score	dbSNP_Name
NLGN3	70291748	g	A	Intron	1	no_dbSNP
NLGN3	70291656	c	T	Intron	1	no_dbSNP
NLGN3	70291342	g	A	Intron	0.941	no_dbSNP
NLGN3	70290929	a	G	Intron	0.921	rs7051529
NLGN3	70290163	c	T	Intron	1	no_dbSNP
NLGN3	70289941	c	T	Intron	1	rs2233440
NLGN3	70286263	g	A	Intron	1	no_dbSNP
NLGN3	70285256	g	A	Intron	0.986	no_dbSNP
NLGN3	70284973	t	G	Intron	1	no_dbSNP
NLGN3	70282170	g	A	Intron	1	rs62609614
NLGN3	70281630	c	A	Intron	0.969	no_dbSNP
NLGN3	70281629	c	G	Intron	0.992	no_dbSNP
NLGN4X	5831786	c	T	Silent	0.919	rs7049300
NLGN4X	5831468	g	C	Silent	0.099	rs61741754
NLGN4X	5821532	c	T	Replacement	0.531	rs3747333
NLGN4X	5821530	c	G	Silent	0.008	rs3747334
NLGN4X	5818136	t	C	3'UTR	1	no_dbSNP

Table 4.6 Exonic and Highly Conserved Non-Coding Variants in NRXN1β

Coordinate	Reference	Mutation	Position	PhastCon Score	dbSNP_Name
49999751	t	C	3'UTR	1	no_dbSNP
49999816	g	A	3'UTR	1	no_dbSNP
50000547	t	G	3'UTR	0.951	no_dbSNP
50001259	c	G	3'UTR	0.988	rs12998798
50001622	a	C	3'UTR	0.996	no_dbSNP
50001641	a	C	3'UTR	0.99	no_dbSNP
50001654	c	T	3'UTR	0.993	no_dbSNP
50001655	t	C	3'UTR	0.907	no_dbSNP
50001966	a	T	3'UTR	1	no_dbSNP
50002181	t	G	3'UTR	1	no_dbSNP
50002476	g	A	3'UTR	1	rs1045881
50002716	t	G	Replacement	1	no_dbSNP
50002845	a	T	Replacement	0.992	no_dbSNP
50002856	a	G	Silent	1	rs55923848
50020286	t	G	Intron	1	no_dbSNP
50024303	t	C	Intron	1	no_dbSNP
50024353	t	G	Silent	1	no_dbSNP
50024395	c	G	Silent	1	no_dbSNP
50024777	c	G	Intron	0.99	rs6753652
50024825	a	G	Intron	0.962	no_dbSNP
50024827	g	T	Intron	0.993	no_dbSNP
50024832	t	C	Intron	1	no_dbSNP
50130977	a	G	Intron	0.999	rs17039714
50131044	t	G	Intron	0.9	no_dbSNP
50131687	a	G	Intron	0.999	rs1452772
50131778	t	C	Intron	0.962	no_dbSNP
50133957	a	C	Replacement	1	no_dbSNP
50134020	a	C	Replacement	1	no_dbSNP
50134079	t	G	Replacement	1	no_dbSNP
50134108	a	G	Silent	1	no_dbSNP
50134281	t	C	Intron	1	no_dbSNP
50134380	a	G	Intron	1	no_dbSNP
50134475	t	G	Intron	0.983	no_dbSNP
50134592	t	C	Intron	1	rs17039730
50134611	g	A	Intron	0.998	rs13021036
50134899	t	A	Intron	0.998	rs11331484
50135142	a	G	Intron	0.934	rs4971644
50135274	a	T	Intron	0.999	no_dbSNP
50135302	c	A	Intron	1	no_dbSNP
50135306	t	A	Intron	1	no_dbSNP
50135308	a	T	Intron	1	no_dbSNP
50171630	t	A	Intron	1	no_dbSNP
50171638	g	T	Intron	1	no_dbSNP
50172000	t	G	Replacement	1	no_dbSNP
50172020	t	G	Silent	1	no_dbSNP
50173108	g	A	Intron	0.979	rs57137390
50317493	g	A	Replacement	1	no_dbSNP
50317497	g	A	Silent	1	no_dbSNP
50317569	g	A	Silent	0.97	no_dbSNP
50317839	c	T	Intron	0.935	no_dbSNP
50317997	c	A	Intron	0.962	no_dbSNP

Table 4.7 List of Deletions and Insertions

	DELETIONS		INSERTIONS	
	Number	**Average per sample**	**Number**	**Average per sample**
DNA Element	3	0.02	1	0.01
LINE	10	0.07	0	0.00
Tandem Repeat	10	0.07	0	0.00
Low Complexity Element	179	1.24	2	0.01
SINE	700	4.86	47	0.33
Simple Repeat	527	3.66	33	0.23
Non-repetitive DNA	3422	23.76	22	0.15
TOTAL	4851	33.69	105	0.73

Table 4.8 Coding deletions

Gene	Position	Number of bases	Bases deleted	Substitution Type
NLGN4X	chrX: 5831324-5831340	16	GGAGCCGTACTGCGCG	Frameshift; deletes 44% of protein
NLGN4X	chrX: 5831853	1	A	Frameshift; deletes 65% of protein
NLGN4X	chrX: 5957348	1	T	Frameshift; deletes 76% of protein
NRXN1β	chr2: 50134052	1	A	Frameshift; deletes 40% of protein

Table 4.9 Long PCR Primers

ID	Primer Sequence
NLGN3.0	5'-AAAGGTACCCAAAGTAGTGGTGAGCTAGGA-3'
NLGN3.1	3'-GACAGAGGTGTGTATGGCAGGAGTTACTAAA-5'
NLGN3.2	5'-CAACGAAGACTGTCTCTACCTGAACGTCTAT-3'
NLGN3.3	3'-GAATGGAGTTACCTGGAGTGCTAGGAGAAT-5'
NLGN3.4	5'-AGAGAGGAGGGAGGACTAAAAGAAGGACAG-3'
NLGN3.5	3'-GATGATAGAAGGCGTAGAAGTAGGTAGGCG-5'
NLGN3.6	5'-AGAGACTGTGTTCCTAGGTGACCATAGTGG-3'
NLGN3.7	3'-CTGCCCATCTCCAGTGTACCATATTAGTGT-5'
NLGN4.10	5'-GGAACCAGTGACCTCAAGAACTAGTCTGAA-3'
NLGN4.11	3'-CTCACTGCTTAATAGATGAGGTAGCCACACAT-5'
NLGN4.12	5'-CCTCATTTCTACTATGCGTACTCGCTGACT-3'
NLGN4.13	3'-GAACTCCTAGCATACTCATTACGCTAAGGTGA-5'
NLGN4.14	5'-GGAGCGCATTTCTACTTCTACCTTGAGTCTA-3'
NLGN4.15	3'-GTCGGATCTAGTGGAGTCTGTAACTTACGTTG-5'
NLGN4.16	5'-ATAGGGCATAGGTACTCAAGTGGGTAGGTG-3'
NLGN4.17	3'-CTCAAGTAGCTCTCTGAGAGATCTCCATTCTG-5'
NLGN4.18	5'-GACTTAGTATGTGAGACTGGAACTTCTCGGC-3'
NLGN4.19	3'-CAGGAGCAGCGACTTATGTAGGGATAGTTA-5'
NLGN4.20	5'-CTTCCAAACAACGGTGGTCTG-3'
NLGN4.21	3'-CTGCCCGTCCACAGACTATTG-5'
NRXN.0	5'-AGAGACGGATACTGTAATGGTTAAAGTTAGT-3'
NRXN.1	3'-CCTGATTCAGTCTAGAGTTAGACCTATGTT-5'
NRXN.2	5'-CTATCTCTCATGTGTCTACGCTTCTTAAAC-3'
NRXN.3	3'-CAAACTCTACATAGTCGTGAGTGTAAAAGG-5'
NRXN.4	5'-GTAGGAATATAAACAGGACTTTATGGAGGAT-3'
NRXN.5	3'-CACTAACTCAATCTATGAGGGTACTCACAG-5'
NRXN.6	5'-TTTCTCACTAGTTCTCTACTTATCTCACGG-3'
NRXN.7	3'-ATAAGAACCAGGTCTATAATAGACGAGACC-5'
NRXN.8	5'-CTGATAGAAGGAGAACTTATAGGGAAAGAC-3'
NRXN.9	3'-TGAATAGTTTGACAGCTACTAGATGTTGCT-5'

FIGURE LEGENDS

Figure 4.1. Schematic of Paired-End Multiplexed Sequencing on the Illumina Genome Analyzer. Changes to the basic Illumina protocol are highlighted in red.

A. Genomic DNA is fragmented and the ends are repaired to convert overhangs into blunt ends. Then dATP adds "A" bases to the blunt ends to allow for the ligation of adapters. After ligation, 300 base pair fragments are selected and amplified. During this amplification step a unique 6-base tag is attached to individual samples allowing for 12 samples to be concomitantly added to one lane of a flowcell.
B. The 12 pooled samples are then added to the flowcell and bridge amplification is performed. This creates millions of clusters with each cluster containing many copies of a single DNA fragment. The sequencing primer is then attached.
C. The flowcell is then placed in the Genome Analyzer. All four fluorescently labeled nucleotides are flowed simultaneously. After the first base is incorporated and read, the base is deblocked, and the process is repeated for the remaining bases in the read. After the end of the single end read (for this experiment, 36 base read) and the 6 base index read, the clusters are linearized and the process of bridge amplification is performed again in the Genome Analyzer machine. Then a second 36 base read is performed; this is called paired-end sequencing.

Figure 4.2. Analysis Pipelines. Basecalls generated by the Illumina Genome Analyzer are input into 3 different alignment algorithms, MAQ, BWA and Bowtie. The MAQ and Bowtie alignments are then assembled using the MAQ assembly while BWA is assembled using SAMTools. SNP calling identifies variants in each consensus sequence. FASTA files containing the consensus sequences are analyzed using the PopGen program, which measures the amount of variation by functional class, and SeqAnt, which annotates each variant by sample.

Figure 4.3. Trace result from Agilent Bioanalyzer showing the proper fragmentation profile. Most of the DNA is between 250 and 350 base pairs, which is ideal for paired-end sequencing on the Illumina Genome Analyzer.

Figure 4.4. Trace result from Agilent Bioanalyzer showing the proper size selection profile. Most of the DNA is about 350-375 base pairs, which is an ideal insert size for paired-end sequencing.

Figure 4.1

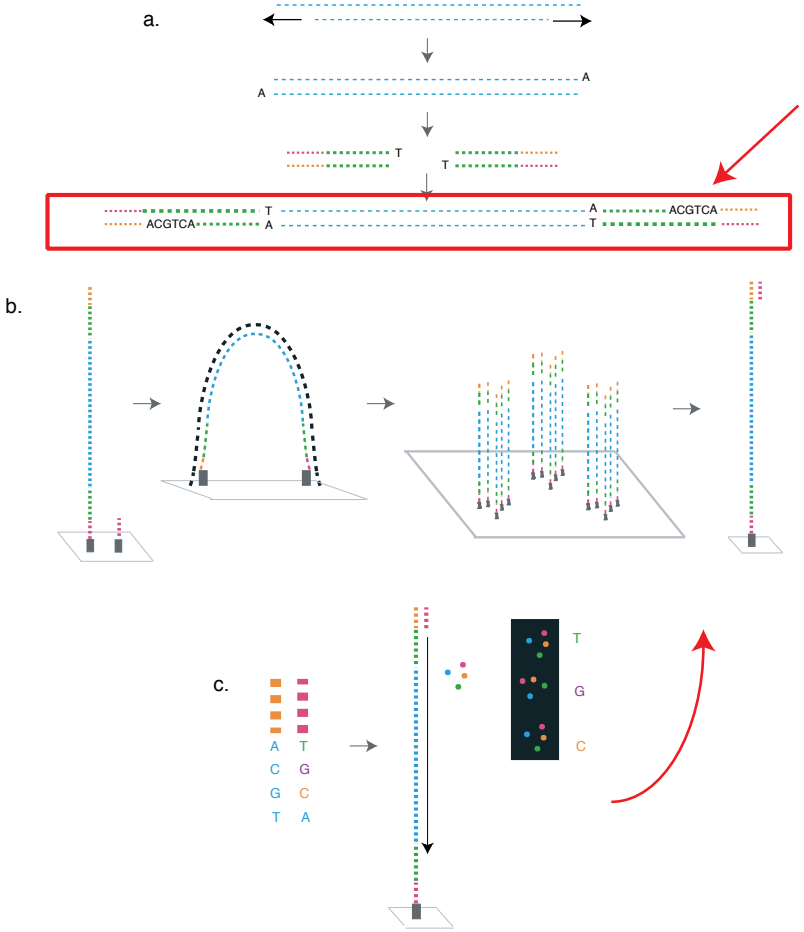


Figure 4.2

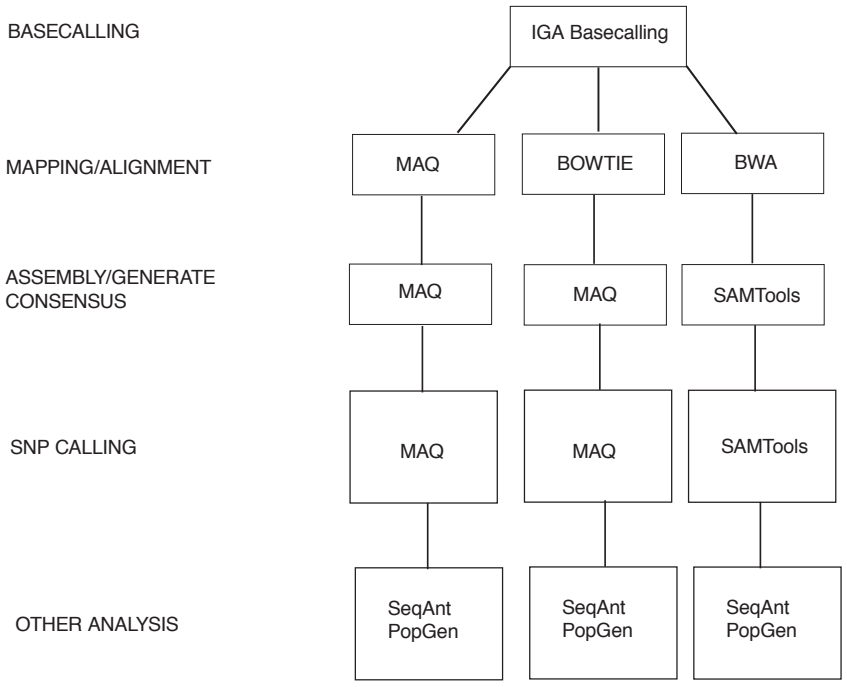

BASECALLING
MAPPING/ALIGNMENT
ASSEMBLY/GENERATE CONSENSUS
SNP CALLING
OTHER ANALYSIS
IGA Basecalling
MAQ
BOWTIE
BWA
MAQ
MAQ
SAMTools
MAQ
MAQ
SAMTools
SeqAnt PopGen
SeqAnt PopGen
SeqAnt PopGen

Figure 4.3

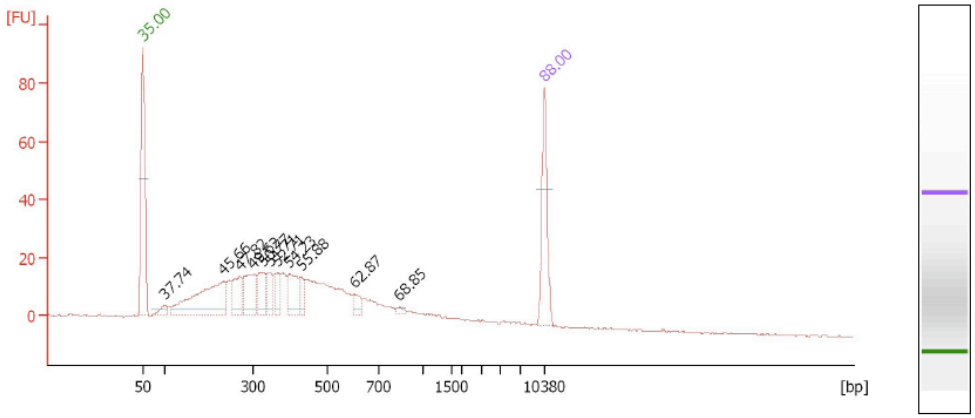
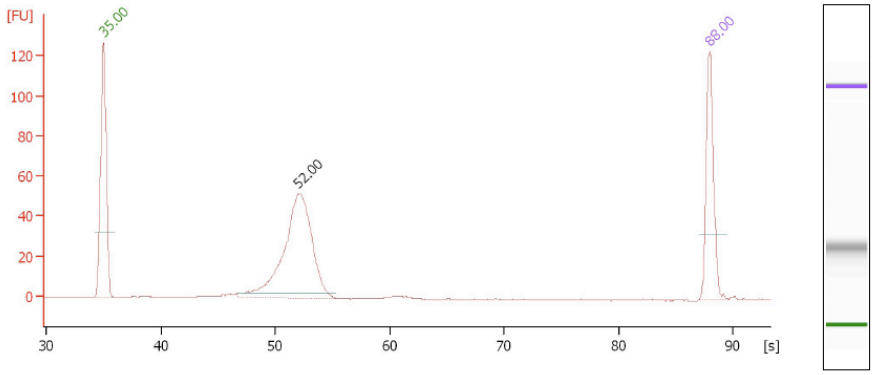

[FU]
80
60
40
20
0
35.00
88.00
37.74
45.66
55.88
62.87
68.85
50
300
500
700
1500
10380
[bp]

Figure 4.4



Overall Results for sample 2 : C2

Number of peaks found: 1

Peak table for sample 2 : C2

Peak		Size [bp]	Conc. [ng/µl]	Molarity [nmol/l]	Observations
1	◀	50	8.30	251.5	Lower Marker
2		349	9.97	43.3	
3	▶	10,380	4.20	0.6	Upper Marker

CONCLUSION

Identifying the heritable component of complex human diseases like Autism Spectrum Disorder (ASD) remains a central challenge for the field of human genetics. Evolutionary genetics provides an ideal framework with which to examine these types of diseases. Complex diseases can be explored as quantitative traits and the genetic contribution of common and rare alleles towards these traits can be studied. The ultimate goal of this research program is to identify and characterize the genetic variants that contribute to a complex trait, with the eventual goal of both explaining the great heterogeneity and understanding the biological systems disrupted in ASD.

Completing this ambitious research plan requires the ability to first find the relevant genomic variation in selected patient samples. Next generation sequencing technologies, which offer the opportunity to efficiently pursue this research program, are improving at a rapid pace. Even since beginning this project, technologies that were previously considered "next generation" are now outdated. As opposed to methods such as Genome Wide Association Studies (GWAS) which focus exclusively on common SNP variation, sequencing the entire genome, or even just selected target regions, allows for detection of all classes of genetic variation in a given set of samples, independent of the population frequency of these variants. Deep sequencing of the unique coding and non-coding regions, and eventually entire genomes, can capture all relevant genomic variation and provides a foundation for the future of human genetics. Recently new algorithms have been developed that use data from next generation sequencing to even examine copy number variation and structural variation in the human genome [140]. As the costs per base decrease, high throughput, deep sequencing may soon replace other methods for identifying

variation in the human genome. However, it may not always be necessary to sequence the entire genome. For many clinical and diagnostic applications, it may be sufficient, more efficient, and even desirable to just sequence a target region or regions containing known susceptibility loci.

In Chapter 1 I describe a method for isolating target DNA for downstream resequencing applications, Microarray-based Genomic Selection (MGS) [66]. In this publication, we demonstrated that with minimal sample manipulation target DNA was efficiently isolated for resequencing arrays. The sequence generated was highly accurate and complete. This method has also recently been adapted by our laboratory for the Illumina Genome Analyzer [163] with high accuracy and completeness relative to other genomic selection technologies. In Chapters 2 and 3 I explore the advantages and limitations of MGS. Probe composition and distribution is critical to the success or failure of MGS. For example, sequences with high GC content or highly homologous sequences may not be efficiently selected or the method may select the paralog rather than the sequence of interest. Yet, as seen in Chapters 3 and 4, genes that are part of larger gene families in gene networks or pathways have been associated with complex human diseases, such as the neuroligins and ASD. Alternative methods for target DNA isolation paired with next generation sequencing are viable options for smaller scale studies of candidate genes that contain highly homologous sequence, such as the one in Chapter 4. In Chapter 4 I demonstrate that multiplexed paired-end sequencing on the Illumina Genome Analyzer successfully produces reads that can then be assembled for analysis and variation detection. I also introduced a novel basecalling algorithm that was developed under a population

genetics framework and a new pipeline for identifying insertions and deletions. Using these methods, I was able to identify a series of promising variants, many of them in noncoding, but highly evolutionary conserved, sequences that may contribute to ASD susceptibility in our patient samples. This comprehensive genetic variation discovery will provide a solid foundation for future functional studies at these ASD candidate loci,

Throughout the thesis I have demonstrated the importance of utilizing population genetics and evolutionary quantitative genetics to interpret sequence data and guide experimental design. It is critical to understand the amounts of variation expected for different functional classes as well as the difference in the amount of variation expected for X-linked versus autosomal loci. Also understanding how allele frequencies change over time due to selection or environmental effects is essential for designing a successful experiment to examine complex human disease. Evolutionary genetics theory estimates that the genetic contribution to quantitative traits is additive; however, under different models of selection the distribution of allelic effects can be dominated by primarily common alleles or primarily rare alleles. For example, Genome Wide Association Studies (GWAS) rationalize their experimental design using a model of balancing selection where alleles that were previously neutral or beneficial during human evolution are now deleterious in the current environment. These alleles are common in the human population because they were not selected against in the past and rose to frequency due to drift or positive selection. However, common alleles of large effect have failed to explain the vast majority of the genetic contribution to complex common diseases. It is likely that rare variants with

moderate phenotypic effect significantly contribute to complex human disorders. If this hypothesis is true, then the development of methods that can efficiently detect and annotate all the relevant genomic variation, which was a major goal of this thesis, provide a critical foundation necessary if this research program is to be successful. Additionally, gene-gene interactions, epigenetics and structural variation most certainly play a role in these complex traits and will need to be incorporated in studies that aim to understand the genetic basis of complex human traits like ASD. The rapid pace of innovation in next generation sequencing and this thesis demonstrate that single laboratories can now identify both common and rare variants and evaluate their role in complex diseases.

REFERENCES

1. Xiong, M., Feghali-Bostwick, C.A., Arnett, F.C. & Zhou, X. A systems biology approach to genetic studies of complex diseases. *FEBS Lett* **579**, 5325-32 (2005).
2. A haplotype map of the human genome. *Nature* **437**, 1299-320 (2005).
3. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661-78 (2007).
4. Newton-Cheh, C. et al. Genome-wide association study identifies eight loci associated with blood pressure. *Nat Genet* (2009).
5. Cohen, J.C. et al. Multiple rare variants in NPC1L1 associated with reduced sterol absorption and plasma low-density lipoprotein levels. *Proc Natl Acad Sci U S A* **103**, 1810-5 (2006).
6. Barton, N.H. & Turelli, M. Evolutionary quantitative genetics: how little do we know? *Annu Rev Genet* **23**, 337-70 (1989).
7. Fisher, R.A. The correlation between relatives under the supposition of Mendelian inheritance. *Trans. R. Soc. Edinburgh* **52**, 399-433 (1918).
8. Kimura, M. *The Neutral Theory of Molecular Evolution*, (Cambridge University Press, Cambridge, 1983).
9. Ohta, T. Slightly deleterious mutant substitutions in evolution. *Nature* **246**, 96-8 (1973).
10. Lande, R. The Genetic Covariance between Characters Maintained by Pleiotropic Mutations. *Genetics* **94**, 203-215 (1980).

11. Lande, R. The maintenance of genetic variability by mutation in a polygenic character with linked loci. *Genet Res* **26**, 221-35 (1975).

12. Gillespie, J.H. *The Causes of Molecular Evolution*, (Oxford University Press, Inc., New York City, 1991).

13. Wright, S. *Evolution and the Genetics of Populations*, (University of Chicago Press, Chicago, 1968-78).

14. Shrimpton, A.E. & Robertson, A. The Isolation of Polygenic Factors Controlling Bristle Score in Drosophila Melanogaster. II. Distribution of Third Chromosome Bristle Effects within Chromosome Sections. *Genetics* **118**, 445-459 (1988).

15. Falconer, D. The inheritance of liability to certain diseases, estimated from the incidence among relatives. *Ann Hum Genet, Lond* **29**, 51-76 (1965).

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

17. Lander, E.S. The new genomics: global views of biology. *Science* **274**, 536-9. (1996).

18. Neel, J.V. Diabetes mellitus: a "thrifty" genotype rendered detrimental by "progress"? *Am J Hum Genet* **14**, 353-62 (1962).

19. Di Rienzo, A. & Hudson, R.R. An evolutionary framework for common diseases: the ancestral-susceptibility model. *Trends Genet* **21**, 596-601 (2005).

20. Zwick, M.E., Cutler, D.J. & Chakravarti, A. Patterns of Genetic Variation in Mendelian and Complex Traits. in *Annu. Rev. Genomics Hum. Genet.*, Vol. 1 387-407 (2000).

21. Reich, D.E. & Lander, E.S. On the allelic spectrum of human disease. *Trends Genet* **17**, 502-10 (2001).

22. Cantor, C.R. & Jukes, T.H. The repetition of homologous sequences in the polypetide chains of certain cytochromes and globins. *Proc Natl Acad Sci U S A* **56**, 177-84 (1966).

23. Smith, D.J. & Lusis, A.J. The allelic structure of common disease. *Hum Mol Genet* **11**, 2455-61 (2002).

24. Schork, N.J., Murray, S.S., Frazer, K.A. & Topol, E.J. Common vs. rare allele hypotheses for complex diseases. *Curr Opin Genet Dev* **19**, 212-9 (2009).

25. Pritchard, J.K. Are rare variants responsible for susceptibility to complex diseases? *Am J Hum Genet* **69**, 124-37 (2001).

26. Mitchell, A.A., Chakravarti, A. & Cutler, D.J. On the probability that a novel variant is a disease-causing mutation. *Genome Res* **15**, 960-6 (2005).

27. Cohen, J.C. et al. Multiple rare alleles contribute to low plasma levels of HDL cholesterol. *Science* **305**, 869-72 (2004).

28. Oti, M. & Brunner, H.G. The modular nature of genetic diseases. *Clin Genet* **71**, 1-11 (2007).

29. Hennah, W. et al. Families with the risk allele of DISC1 reveal a link between schizophrenia and another component of the same molecular pathway, NDE1. *Hum Mol Genet* **16**, 453-62 (2007).

30. Lesnick, T.G. et al. A Genomic Pathway Approach to a Complex Disease: Axon Guidance and Parkinson Disease. *PLoS Genet* **3**, e98 (2007).

31. Folstein, S. & Rutter, M. Infantile autism: a genetic study of 21 twin pairs. *J Child Psychol Psychiatry* **18**, 297-321 (1977).

32. Bailey, A. et al. Autism as a strongly genetic disorder: evidence from a British twin study. *Psychol Med* **25**, 63-77 (1995).

33. Ritvo, E.R., Freeman, B.J., Mason-Brothers, A., Mo, A. & Ritvo, A.M. Concordance for the syndrome of autism in 40 pairs of afflicted twins. *Am J Psychiatry* **142**, 74-7 (1985).

34. Jorde, L.B. et al. Complex segregation analysis of autism. *Am J Hum Genet* **49**, 932-8 (1991).

35. Risch, N. et al. A genomic screen of autism: evidence for a multilocus etiology. *Am J Hum Genet* **65**, 493-507 (1999).

36. Laumonnier, F., Cuthbert, P.C. & Grant, S.G. The role of neuronal complexes in human x-linked brain diseases. *Am J Hum Genet* **80**, 205-20 (2007).

37. Skuse, D. X-linked genes and the neural basis of social cognition. *Novartis Found Symp* **251**, 84-98; discussion 98-108; 109-11, 281-97 (2003).

38. Zechner, U. et al. A high density of X-linked genes for general cognitive ability: a run-away process shaping human evolution? *Trends Genet* **17**, 697-701 (2001).

39. Ross, M.T. et al. The DNA sequence of the human X chromosome. *Nature* **434**, 325-37 (2005).

40. Kitano, T., Schwarz, C., Nickel, B. & Paabo, S. Gene diversity patterns at 10 X-chromosomal loci in humans and chimpanzees. *Mol Biol Evol* **20**, 1281-9 (2003).

41. Nielsen, R. et al. A scan for positively selected genes in the genomes of humans and chimpanzees. *PLoS Biol* **3**, e170 (2005).

42. Arbiza, L., Dopazo, J. & Dopazo, H. Positive selection, relaxation, and acceleration in the evolution of the human and chimp genome. *PLoS Comput Biol* **2**, e38 (2006).

43. Khaitovich, P. et al. Parallel patterns of evolution in the genomes and transcriptomes of humans and chimpanzees. *Science* **309**, 1850-4 (2005).

44. King, M.C. & Wilson, A.C. Evolution at two levels in humans and chimpanzees. *Science* **188**, 107-16 (1975).

45. Caceres, M. et al. Elevated gene expression levels distinguish human from non-human primate brains. *Proc Natl Acad Sci U S A* **100**, 13030-5 (2003).

46. Gilad, Y., Oshlack, A. & Rifkin, S.A. Natural selection on gene expression. *Trends Genet* **22**, 456-61 (2006).

47. Keightley, P.D., Lercher, M.J. & Eyre-Walker, A. Evidence for widespread degradation of gene control regions in hominid genomes. *PLoS Biol* **3**, e42 (2005).

48. Collins, F.S., Morgan, M. & Patrinos, A. The Human Genome Project: lessons from large-scale biology. *Science* **300**, 286-90 (2003).

49. Shendure, J., Mitra, R.D., Varma, C. & Church, G.M. Advanced sequencing technologies: methods and goals. *Nat Rev Genet* **5**, 335-44 (2004).

50. Sanger, F., Nicklen, S. & Coulson, A.R. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A* **74**, 5463-7 (1977).

51. Smith, L.M. et al. Fluorescence detection in automated DNA sequence analysis. *Nature* **321**, 674-9 (1986).

52. Paegel, B.M., Emrich, C.A., Wedemayer, G.J., Scherer, J.R. & Mathies, R.A. High throughput DNA sequencing with a microfabricated 96-lane capillary array electrophoresis bioprocessor. *Proc Natl Acad Sci U S A* **99**, 574-9 (2002).

53. Ewing, B., Hillier, L., Wendl, M.C. & Green, P. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res* **8**, 175-85 (1998).

54. Ewing, B. & Green, P. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res* **8**, 186-94 (1998).

55. Chan, E.Y. Advances in sequencing technology. *Mutat Res* **573**, 13-40 (2005).

56. Cutler, D.J. et al. High-throughput variation detection and genotyping using microarrays. *Genome Res* **11**, 1913-25 (2001).

57. Zwick, M.E. et al. Microarray-based resequencing of multiple Bacillus anthracis isolates. *Genome Biol* **6**, R10 (2005).

58. Steinberg, K.M., Okou, D.T. & Zwick, M.E. Applying rapid genome sequencing technologies to characterize pathogen genomes. *Anal Chem* **80**, 520-8 (2008).

59. Margulies, M. et al. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**, 376-80 (2005).

60. Bentley, D. et al. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**, 53-59 (2008).

61. Craig, D.W. et al. Identification of genetic variants using bar-coded multiplexed sequencing. *Nat Methods* **5**, 887-93 (2008).

62. Harismendy, O. et al. Evaluation of next generation sequencing platforms for population targeted sequencing studies. *Genome Biol* **10**, R32 (2009).

63. Harris, T.D. et al. Single-molecule DNA sequencing of a viral genome. *Science* **320**, 106-9 (2008).

64. Kasianowicz, J.J., Brandin, E., Branton, D. & Deamer, D.W. Characterization of individual polynucleotide molecules using a membrane channel. *Proc Natl Acad Sci U S A* **93**, 13770-3 (1996).

65. Rhee, M. & Burns, M.A. Nanopore sequencing technology: research trends and applications. *Trends Biotechnol* **24**, 580-6 (2006).

66. Okou, D.T. et al. Microarray-based genomic selection for high-throughput resequencing. *Nat Methods* **4**, 907-9 (2007).

67. Shendure, J. et al. Accurate multiplex polony sequencing of an evolved bacterial genome. *Science* **309**, 1728-32 (2005).

68. Hinds, D.A. et al. Whole-genome patterns of common DNA variation in three human populations. *Science* **307**, 1072-9 (2005).

69. Sjoblom, T. et al. The consensus coding sequences of human breast and colorectal cancers. *Science* **314**, 268-74 (2006).

70. Raymond, C.K. et al. Targeted, haplotype-resolved resequencing of long segments of the human genome. *Genomics* **86**, 759-66 (2005).

71. Raymond, C.K., Sims, E.H. & Olson, M.V. Linker-mediated recombinational subcloning of large DNA fragments using yeast. *Genome Res* **12**, 190-7 (2002).

72. Kouprina, N., Noskov, V.N. & Larionov, V. Selective isolation of large chromosomal regions by transformation-associated recombination cloning for structural and functional analysis of mammalian genomes. *Methods Mol Biol* **349**, 85-101 (2006).

73. Dahl, F. et al. Multigene amplification and massively parallel sequencing for cancer mutation discovery. *Proc Natl Acad Sci U S A* **104**, 9387-92 (2007).

74. Bashiardes, S. et al. Direct genomic selection. *Nat Methods* **2**, 63-9 (2005).

75. De Boulle, K. et al. A point mutation in the FMR-1 gene associated with fragile X mental retardation. *Nat Genet* **3**, 31-5 (1993).

76. Gu, Y., Lugenbeel, K.A., Vockley, J.G., Grody, W.W. & Nelson, D.L. A de novo deletion in FMR1 in a patient with developmental delay. *Hum Mol Genet* **3**, 1705-6 (1994).

77. Bentley, D.R. Whole-genome re-sequencing. *Curr Opin Genet Dev* **16**, 545-52 (2006).

78. Karolchik, D. et al. The UCSC Genome Browser Database. *Nucleic Acids Res* **31**, 51-4 (2003).

79. Kleinjan, D.A. & van Heyningen, V. Long-range control of gene expression: emerging mechanisms and disruption in disease. *Am J Hum Genet* **76**, 8-32 (2005).

80. Engels, B. Amplify 3. *http://engels.genetics.wisc.edu/amplify/* (2005).

81. Mardis, E.R. The impact of next-generation sequencing technology on genetics. *Trends Genet* **24**, 133-41 (2008).

82. Ng, S.B. et al. Targeted capture and massively parallel sequencing of 12 human exomes. *Nature* (2009).

83. Kryukov, G.V., Pennacchio, L.A. & Sunyaev, S.R. Most rare missense alleles are deleterious in humans: implications for complex disease and association studies. *Am J Hum Genet* **80**, 727-39 (2007).

84. Porreca, G. et al. Multiplex amplification of large sets of human exons. *Nat Meth* **4**, 931-936 (2007).

85. Turner, E.H., Lee, C., Ng, S.B., Nickerson, D.A. & Shendure, J. Massively parallel exon capture and library-free resequencing across 16 genomes. *Nat Methods* **6**, 315-6 (2009).

86. Shah, S.P. et al. Mutational evolution in a lobular breast tumour profiled at single nucleotide resolution. *Nature* **461**, 809-13 (2009).

87. Olson, M. Enrichment of super-sized resequencing targets from the human genome. *Nat Methods* **4**, 891-2 (2007).

88. Sharp, A.J. et al. Optimal design of oligonucleotide microarrays for measurement of DNA copy-number. *Hum Mol Genet* **16**, 2770-9 (2007).

89. Chou, C.C., Chen, C.H., Lee, T.T. & Peck, K. Optimization of probe length and the number of probes per gene for optimal microarray analysis of gene expression. *Nucleic Acids Res* **32**, e99 (2004).

90. Hodges, E., Xuan, Z., Balija, V., Kramer, M. & Molla, M.N. Genome-wide in situ exon capture for selective resequencing. *Nat. Genet.* (2007).

91. Philippe, A. et al. Genome-wide scan for autism susceptibility genes. Paris Autism Research International Sibpair Study. *Hum Mol Genet* **8**, 805-12 (1999).

92. Thomas, N.S. et al. Xp deletions associated with autism in three females. *Hum Genet* **104**, 43-8 (1999).

93. Liu, J. et al. A genomewide screen for autism susceptibility loci. *Am J Hum Genet* **69**, 327-40 (2001).

94. Jamain, S. et al. Mutations of the X-linked genes encoding neuroligins NLGN3 and NLGN4 are associated with autism. *Nat Genet* **34**, 27-9 (2003).

95. Yonan, A.L. et al. A genomewide screen of 345 families for autism-susceptibility loci. *Am J Hum Genet* **73**, 886-97 (2003).

96. Laumonnier, F. et al. X-linked mental retardation and autism are associated with a mutation in the NLGN4 gene, a member of the neuroligin family. *Am J Hum Genet* **74**, 552-7 (2004).

97. Yan, J. et al. Analysis of the neuroligin 3 and 4 genes in autism and other neuropsychiatric patients. *Mol Psychiatry* **10**, 329-32 (2005).

98. Chih, B., Afridi, S.K., Clark, L. & Scheiffele, P. Disorder-associated mutations lead to functional inactivation of neuroligins. *Hum Mol Genet* **13**, 1471-7 (2004).

99. Comoletti, D. et al. The Arg451Cys-neuroligin-3 mutation associated with autism reveals a defect in protein processing. *J Neurosci* **24**, 4889-93 (2004).

100. Yamakawa, H. et al. Neuroligins 3 and 4X interact with syntrophin-gamma2, and the interactions are affected by autism-related mutations. *Biochem Biophys Res Commun* **355**, 41-6 (2007).

101. Daoud, H. et al. Autism and Nonsyndromic Mental Retardation Associated with a De Novo Mutation in the NLGN4X Gene Promoter Causing an Increased Expression Level. *Biol Psychiatry* (2009).

102. Feng, J. et al. High frequency of neurexin 1beta signal peptide structural variants in patients with autism. *Neurosci Lett* **409**, 10-3 (2006).

103. Chubykin, A.A. et al. Dissection of synapse induction by neuroligins: effect of a neuroligin mutation associated with autism. *J Biol Chem* **280**, 22365-74 (2005).

104. Kim, H.G. et al. Disruption of neurexin 1 associated with autism spectrum disorder. *Am J Hum Genet* **82**, 199-207 (2008).

105. Szatmari, P. et al. Mapping autism risk loci using genetic linkage and chromosomal rearrangements. *Nat Genet* **39**, 319-28 (2007).

106. Bucan, M. et al. Genome-wide analyses of exonic copy number variants in a family-based study point to novel autism susceptibility genes. *PLoS Genet* **5**, e1000536 (2009).

107. Rujescu, D. et al. Disruption of the neurexin 1 gene is associated with schizophrenia. *Hum Mol Genet* **18**, 988-96 (2009).

108. Marshall, C.R. et al. Structural variation of chromosomes in autism spectrum disorder. *Am J Hum Genet* **82**, 477-88 (2008).

109. Alarcon, M. et al. Linkage, association, and gene-expression analyses identify CNTNAP2 as an autism-susceptibility gene. *Am J Hum Genet* **82**, 150-9 (2008).

110. Bonaglia, M.C. et al. Identification of a recurrent breakpoint within the SHANK3 gene in the 22q13.3 deletion syndrome. *J Med Genet* **43**, 822-8 (2006).

111. Durand, C.M. et al. Mutations in the gene encoding the synaptic scaffolding protein SHANK3 are associated with autism spectrum disorders. *Nat Genet* **39**, 25-7 (2007).

112. Boeckers, T.M., Bockmann, J., Kreutz, M.R. & Gundelfinger, E.D. ProSAP/Shank proteins - a family of higher order organizing molecules of the postsynaptic density with an emerging role in human neurological disease. *J Neurochem* **81**, 903-10 (2002).

113. Skaletsky, H. et al. The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. *Nature* **423**, 825-37 (2003).

114. Philibert, R.A., Winfield, S.L., Sandhu, H.K., Martin, B.M. & Ginns, E.I. The structure and expression of the human neuroligin-3 gene. *Gene* **246**, 303-10 (2000).

115. Irie, M. et al. Binding of neuroligins to PSD-95. *Science* **277**, 1511-5 (1997).

116. Levinson, J.N. et al. Neuroligins mediate excitatory and inhibitory synapse formation: involvement of PSD-95 and neurexin-1beta in neuroligin-induced synaptic specificity. *J Biol Chem* **280**, 17312-9 (2005).

117. Nam, C.I. & Chen, L. Postsynaptic assembly induced by neurexin-neuroligin interaction and neurotransmitter. *Proc Natl Acad Sci U S A* **102**, 6137-42 (2005).

118. Sheng, M. & Kim, E. The Shank family of scaffold proteins. *J Cell Sci* **113 (Pt 11)**, 1851-6 (2000).

119. Varoqueaux, F. et al. Neuroligins determine synapse maturation and function. *Neuron* **51**, 741-54 (2006).

120. Ichtchenko, K., Nguyen, T. & Sudhof, T.C. Structures, alternative splicing, and neurexin binding of multiple neuroligins. *J Biol Chem* **271**, 2676-82 (1996).

121. Bolliger, M.F., Frei, K., Winterhalter, K.H. & Gloor, S.M. Identification of a novel neuroligin in humans which binds to PSD-95 and has a widespread expression. *Biochem J* **356**, 581-8 (2001).

122. Chih, B., Engelman, H. & Scheiffele, P. Control of excitatory and inhibitory synapse formation by neuroligins. *Science* **307**, 1324-8 (2005).

123. Klauck, S.M. Genetics of autism spectrum disorder. *Eur J Hum Genet* **14**, 714-20 (2006).

124. Jamain, S. et al. Reduced social interaction and ultrasonic communication in a mouse model of monogenic heritable autism. *Proc Natl Acad Sci U S A* **105**, 1710-5 (2008).

125. Tabuchi, K. et al. A neuroligin-3 mutation implicated in autism increases inhibitory synaptic transmission in mice. *Science* **318**, 71-6 (2007).

126. Chocholska, S., Rossier, E., Barbi, G. & Kehrer-Sawatzki, H. Molecular cytogenetic analysis of a familial interstitial deletion Xp22.2-22.3 with a highly variable phenotype in female carriers. *Am J Med Genet A* **140**, 604-10 (2006).

127. Talebizadeh, Z., Bittel, D.C., Veatch, O.J., Kibiryeva, N. & Butler, M.G. Brief report: non-random X chromosome inactivation in females with autism. *J Autism Dev Disord* **35**, 675-81 (2005).

128. Talebizadeh, Z. et al. Novel splice isoforms for NLGN3 and NLGN4 with possible implications in autism. *J Med Genet* **43**, e21 (2006).

129. Skuse, D.H. et al. Evidence from Turner's syndrome of an imprinted X-linked locus affecting cognitive function. *Nature* **387**, 705-8 (1997).

130. Wermter, A.K., Kamp-Becker, I., Strauch, K., Schulte-Korne, G. & Remschmidt, H. No evidence for involvement of genetic variants in the X-linked neuroligin genes NLGN3 and NLGN4X in probands with autism spectrum disorder on high functioning level. *Am J Med Genet B Neuropsychiatr Genet* **147B**, 535-7 (2008).

131. Blasi, F. et al. Absence of coding mutations in the X-linked genes neuroligin 3 and neuroligin 4 in individuals with autism from the IMGSAC collection. *Am J Med Genet B Neuropsychiatr Genet* **141**, 220-1 (2006).

132. Gauthier, J. et al. NLGN3/NLGN4 gene mutations are not responsible for autism in the Quebec population. *Am J Med Genet B Neuropsychiatr Genet* **132**, 74-5 (2005).

133. Vincent, J.B. et al. Mutation screening of X-chromosomal neuroligin genes: no mutations in 196 autism probands. *Am J Med Genet B Neuropsychiatr Genet* **129**, 82-4 (2004).

134. Ylisaukko-oja, T. et al. Analysis of four neuroligin genes as candidates for autism. *Eur J Hum Genet* **13**, 1285-92 (2005).

135. Siepel, A. et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* **15**, 1034-50 (2005).

136. Thomas, P.D. et al. PANTHER: a library of protein families and subfamilies indexed by function. *Genome Res* **13**, 2129-41 (2003).

137. Thomas, P.D. et al. PANTHER: a browsable database of gene products organized by biological function, using curated protein family and subfamily classification. *Nucleic Acids Res* **31**, 334-41 (2003).

138. Zwick, M.E. et al. Patterns of Human Genetic Variation Ascertaned by High-throughput Microarray-based Resequencing. (in prep).

139. Kimura, M. & Crow, J.F. Number of Alleles That Can Be Maintained in Finite Population. *Genetics* **49**, 725-& (1964).

140. Alkan, C. et al. Personalized copy number and segmental duplication maps using next-generation sequencing. *Nat Genet* **41**, 1061-7 (2009).

141. Karolchik, D. et al. The UCSC Table Browser data retrieval tool. *Nucleic Acids Res* **32**, D493-6 (2004).

142. Lawson-Yuen, A., Saldivar, J.S., Sommer, S. & Picker, J. Familial deletion within NLGN4 associated with autism and Tourette syndrome. *Eur J Hum Genet* (2008).

143. Wermter, A.K., Kamp-Becker, I., Strauch, K., Schulte-Korne, G. & Remschmidt, H. No evidence for involvement of genetic variants in the X-linked neuroligin genes NLGN3 and NLGN4X in probands with autism spectrum disorder on high functioning level. *Am J Med Genet B Neuropsychiatr Genet* (2008).

144. Shinawi, M. et al. The Xp contiguous deletion syndrome and autism. *Am J Med Genet A* **149A**, 1138-48 (2009).

145. Pampanos, A. et al. A substitution involving the NLGN4 gene associated with autistic behavior in the Greek population. *Genet Test Mol Biomarkers* **13**, 611-5 (2009).

146. Zhang, C. et al. A neuroligin-4 missense mutation associated with autism impairs neuroligin-4 folding and endoplasmic reticulum export. *J Neurosci* **29**, 10843-54 (2009).

147. Watterson, G.A. On the number of segregating sites in genetical models without recombination. *Theor Popul Biol* **7**, 256-76 (1975).

148. Li, H., Ruan, J. & Durbin, R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res* **18**, 1851-8 (2008).

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

150. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-60 (2009).

151. Burrows, M. & Wheeler, D.A. A block-sorting lossless data compression algorithm. (Digital Equipment Corporation, Palo Alto, 1994).

152. Kumar, P., Henikoff, S. & Ng, P.C. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat Protoc* **4**, 1073-81 (2009).

153. Zerbino, D.R. & Birney, E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res* **18**, 821-9 (2008).

154. Delcher, A.L. et al. Alignment of whole genomes. *Nucleic Acids Res* **27**, 2369-76 (1999).

155. Delcher, A.L., Phillippy, A., Carlton, J. & Salzberg, S.L. Fast algorithms for large-scale genome alignment and comparison. *Nucleic Acids Res* **30**, 2478-83 (2002).

156. Larkin, M.A. et al. Clustal W and Clustal X version 2.0. *Bioinformatics* **23**, 2947-8 (2007).

157. Mills, R.E. et al. An initial map of insertion and deletion (INDEL) variation in the human genome. *Genome Res* **16**, 1182-90 (2006).

158. John, B. et al. Human MicroRNA targets. *PLoS Biol* **2**, e363 (2004).

159. Betel, D., Wilson, M., Gabow, A., Marks, D.S. & Sander, C. The microRNA.org resource: targets and expression. *Nucleic Acids Res* **36**, D149-53 (2008).

160. Maragkakis, M. et al. DIANA-microT web server: elucidating microRNA functions through target prediction. *Nucleic Acids Res* **37**, W273-6 (2009).

161. Rusinov, V., Baev, V., Minkov, I.N. & Tabler, M. MicroInspector: a web tool for detection of miRNA binding sites in an RNA sequence. *Nucleic Acids Res* **33**, W696-700 (2005).

162. Wigginton, J.E., Cutler, D.J. & Abecasis, G.R. A note on exact tests of Hardy-Weinberg equilibrium. *Am J Hum Genet* **76**, 887-93 (2005).

163. Okou, D.T. et al. Combining microarray-based genomic selection (MGS) with the Illumina Genome Analyzer platform to sequence diploid target regions. *Ann Hum Genet* **73**, 502-13 (2009).