Identifying Rare Genetic Variants in Autism Spectrum Disorder

A custom-designed SureSelect targeted panel enables massively parallel sequencing to find new variants that may be linked to the disorder.

Customer Success Story

John P. Hussman Institute for Human Genomics, University of Miami Miller School of Medicine

The Solution:

A custom SureSelect DNA targeted panel enabled sequencing analysis of both protein coding and noncoding genomic regions implicated by GWAS to be associated with autism spectrum disorder (ASD), as a first step towards identifying functional sequence variants linked to ASD.

This research study identified more than 27,000 variants in the protein coding regions of nearly 700 genes, including enrichment of loss-of-function variants in ASD candidate genes catalogued in the SFARI Gene database.[1] In addition, more than 450,000 noncoding variants were identified.

The Challenge:

- Genetic variants linked to autism are likely not genotyped in genome-wide association studies (GWAS)
- Whole genome sequencing of large numbers of autism cases is not practical due to cost and time constraints
- A means of targeting regions identified by GWAS for massively parallel sequencing of large sample numbers is needed

Dr. Anthony J. Griswold at the John P. Hussman Institute for Human Genomics is engaged in a search for the genetic alterations that may contribute to risk for autism spectrum disorder (ASD). Twin studies suggest heritability up to 90%, and more than 100 genes have been linked to this disorder[1,2]. However, rare syndromes, de novo copy number variants, and mutations account for only ~20% of cases[3], with no single variant being present in more than 1 to 2% of the total autism population. *De novo* coding variants have also been implicated in a few genes by sequencing studies[4].

Since ASD causing variants are not likely to be the common ones genotyped in GWAS, Dr. Griswold designed and implemented a sequencing approach to attempt to identify un-genotyped common variants and/or rare functional variants that account for the missing genetic contributors to ASD. This study targeted candidate regions identified by GWAS-NR analyses in ASD GWAS datasets, a noise–reduction method for enhancing the power of GWAS to detect true positive associations.[5] It utilized DNA from 2439 ASD cases and 1192 controls, for a total of almost 4,000 samples.

Dr. Griswold chose the SureSelect technology to enrich for the regions identified by GWAS-NR due to its accuracy, reliability, ease of use and amenability to automation to process this large number of samples. The SureSelect baits were designed to capture the exons of 689 genes implicated by GWAS-NR, including their untranslated regions (UTRs), as well as 5 kilobase pairs up and downstream of these genes. Importantly, conserved intronic and intergenic regions implicated by GWAS-NR were also targeted using SureSelect.

Variant Calling

Sequencing provided excellent on-target read alignment, coverage of targeted bases and concordance of genotypes with GWAS (Table 1). This sequence information enabled the identification of nearly 500,000 different variants in the total sample population. Of those, 27,000 (~800 per individual) were protein coding, 77% of these being rare and found in less than 1% of the population.

Less than 100 of these coding mutations per individual are classified as damaging (likely to seriously affect protein function) or nonsense/splice mutations.

The nonsense mutations cause loss of protein function, which was observed in a total of 18 ASD candidate genes present in the SFARI database. Loss-of-function (LOF) variants were observed across 27 cases and at 29 different nucleotide positions in these genes. This study revealed novel LOF mutations in established SFARI gene candidates in ASD cases, including *NRXN1*(3 variants), *SHANK3*, *CNTNAP2*, and *RBFOX1*, which contained a *de novo* premature stop mutation.

No single coding region variant in ASD cases is likely to explain ASD risk, as statistical analysis did not identify a single gene with a significant association. However, rare variants in a combination of genes comprising a pathway might confer increased risk. Statistical analysis of 32 GO terms found a significant association between coding, damaging and missense variants and the "plasma-membrane" GO term.

Non-coding Variants

As expected, non-coding variants greatly outnumbered coding variants, with over 450,000 detected in the total sample population (Table 2). This was due in part to an intentional effort to interrogate non-coding regions that may be implicated in ASD.

Functional non-coding variants likely play important roles in ASD, and a search of functional roles for noncoding DNA elements in the ENCODE database enabled annotation into several categories (Table 3). The discovery of possible functional non-coding variants such as category ½ mutants from the RegulomeDB database is the beginning of an effort to characterize the role of these variants in ASD. For example, one of the non-coding variants occurs in an enchancer in the intron in the AUTS2 gene, which drives gene expression in the midbrain and neural tube.

Table 1. Sequencing Results

Paired-end reads per individual	31.4M ± 13.1M
On-target read alignment	81.9 ± 10.5%
Targeted bases covered ≥ 20X	86.9 ± 6.7%
Average coverage of all targeted bases	84.6 ± 25X
GWAS concordance (5000 markers)	98.6 ± 1.2%

Table 2. Variant Calling

	Cases*	Controls*
Total Variants	**23756 ± 836**	**23749 ± 797**
Coding	765 ± 62	760 ± 64
• Missense	307 ± 30	305 ± 31
• Damaging	85 ± 12	84 ± 13
• Nonsense/splice	3 ± 1.4	2.5 ± 1.8
Non-Coding		
• UTR	871 ± 57	867 ± 58
• Intron	18510 ± 833	18516 ± 818
• Intergenic	3609 ± 220	3604 ± 219

*The numbers shown are averages per individual sample, with standard deviations.

Table 3. Functional Non-coding Variants

	Cases*	Controls*
DNAse I sensitivity sites	70.79 ± 13.4	59.57 ± 11.19
Methylation sites	8.71 ± 3.98	6.1 ± 3.12
Enhancers	10.12 ± 2.15	4.14 ± 1.82
lincRNA	89.67 ± 27.30	59.26 ± 21.2
TX factor binding sites	17.8 ± 4.6	12.6 ± 3.8
Regulome category 1/2	31.12 ± 7.33	23.16 ± 5.18

*The numbers shown are averages per individual sample, with standard deviations.

References

1. A. Bailey et al., *Psychol Med.* **25,** 63-77 **(**1995).
2. C. Betancur, *Brain Res.* **1380**, 42-77 (2011).
3. J.H. Miles, *Genet Med.* **13**, 278-294 (2011).
4. S. J. Sanders et al., *Nature* **485,** 237-241 (2012).
5. J. P.Hussman et al, *Mol. Autism*. 2:1 (2011).

U.S. and Canada
1-800-227-9770
agilent_inquiries@agilent.com

Asia Pacific
inquiry_lsca@agilent.com

Europe
info_agilent@agilent.com

www.agilent.com/genomics/SuccessStories

Published in the USA January 13, 2016
5991-4999EN
PR7000-0332

Agilent Technologies