Genome Wide Variant Analysis of families with Autism Spectrum Disorder (ASD) using an Integrative Bioinformatics Pipeline

Laura T Jimenez-Barron[1,2], Han Fang[1], Jason O'Rawe[1], Ivan Iossifov[1], Gholson J Lyon[1,3]

[1]Cold Spring Harbor Laboratory, Stanley Institute for Cognitive Genomics, New York, NY, [2]Universidad Nacional Autonoma de Mexico, Centro de Ciencias Genomicas, Cuernavaca, Mexico, [3]Utah Foundation for Biomedical Research, Salt Lake City, UT.

Introduction. Autism spectrum disorders (ASD) are a group of developmental disabilities that affect social interaction, communication and are characterized by repetitive behaviors. There is now a large body of evidence that suggests a complex role of genetics in ASD, in which many different loci are involved. Although many current population scale studies have been demonstrably useful, these studies generally focus on analyzing a limited part of the genome or use a limited set of bioinformatics tools. These limitations make it difficult to see the complete and panoramic picture of each ASD case. To address this problem, here we describe an integrative bioinformatics pipeline used to get a more complete and reliable set of candidate ASD-variants for validation and further functional analysis.

Methods. We studied three simplex Autism Families, two of which belong to the Simon's Simplex Collection (SSC), and all probands and families were clinically evaluated and extensively phenotyped. The third family, recruited at the Utah Foundation for Biomedical Research, had extensive clinical evaluations performed, along with fragile X and Chromosomal Microarray Analysis (CMA) on the proband and mother, with no obvious disease-contributory mutations found. All family members were genotyped using an Illumina Omni2.5 Array and/or WGS was performed using the Illumina HiSeq 2000 to ~40-75X coverage. WGS reads were aligned to the GRCh37/hg19 human reference genome using BWA-MEM software, with variant calling for SNVs and INDELs using the GATK HaplotypeCaller and FreeBayes. To better support de novo calls, we used Scalpel for INDEL detection and the Multinomial Analyzer. The ERDS software was used to call CNVs from WGS data. Microarray data were used to call CNVs with the software package PennCNV using the joint-calling algorithm.

Results. The resulting set of candidate variants include three small heterozygous CNVs (~22, ~36 and ~50 Kb). All of the CNVs were only found by ERDS, and despite the fact that the K21 pedigree had microarray data, PennCNV did not detect any CNV in those regions. A heterozygous *de novo* nonsense mutation in *MYBBP1A* was found in one of the quads (K21) located within exon 1, and a second *de novo* variant was also among the final results from another quad (SSC_2), this time a missense mutation in *LAMB3*, which also has not yet been observed in any other ASD proband.

Having established a more comprehensive WGS pipeline, we are moving to implement our framework for the analysis and study of families from Utah and from the SSC .

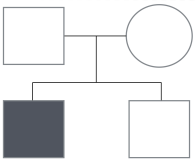

Figure 1. Pedigree structure of a Simplex Autism Family. For a family to be classified as a Simplex Autism Family, it has to be composed of one affected child and at least one unaffected sibling, and both parents do not have obvious autism. Probands and siblings can be either males or females. For this study, two of the families have male probands and unaffected male siblings, whereas the other family has a male proband and a female unaffected sibling.

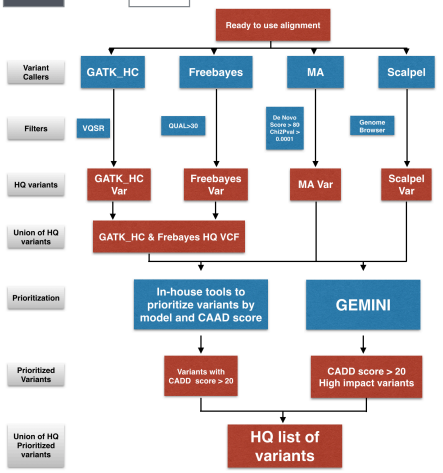

Figure 2. Variant calling pipeline. After aligning the genomes with BWA-MEM 0.7.5a-r405, the resulting alignments were converted to binary format, then sorted and indexed using SAMtools version 0.1.19-44428cd. Duplicated reads were marked and read groups were assigned to each lane using Picard tools **v1.84**. The GATK Indel realigner v3.0-0 was used to correct initial mapping artifacts due to reads aligning to the edges of INDELs, which often map with mismatching bases that may look like evidence for SNPs, while they are not. The GATK Base Quality Score Recalibrator was also used to correct known systematic errors of sequencing technologies. Finally all lanes were merged by sample with Picard tools to generate a ready-to-use alignment. Various algorithms were used to call SNPs and Indels, all resulting variants were filtered and prioritized with different methods.

Table 1. Final set of Small Variants.

Model	Ref->Alt/ Effect	Location	Affected Gene	Algorithms that called the variant	Pedigree ID	ExAC Allele Frequency	CAAD score
De Novo	sub(C->T) missense	chr1: 209823359	LAMB3	Freebayes, Multinomial Analyzer, GATK	SSC_2	0	22.7
De Novo	sub(G-A)nonsense	chr17: 4458481	MYBBP1A	Freebayes, Multinomial Analyzer, GATK	K21	1/74014=0.00001351	40

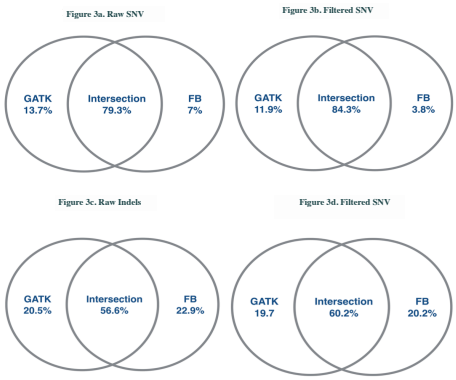

Figure 3. Algorithm concordance. GATK and Freebayes were compared as they are algorithms that call both SNVs and INDELs with a comparable number of calls. The intersection grows when comparing filtered SNVs and Indels. The Multinomial Analyzer and Scalpel were only used to call de novo SNPs and Indels respectively.

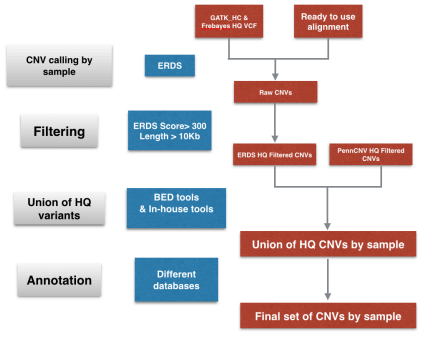

Figure 4. Copy Number Variant calling pipeline. Using the same ready to use alignment described in Figure 2 plus the union of variants called by Freebayes and GATK, the Estimation by Read Depth with Single Nucleotide Variants (ERDS) software was used to call CNVs. PennCNV was used in the samples where Microarray data was available and both calls sets were compared.

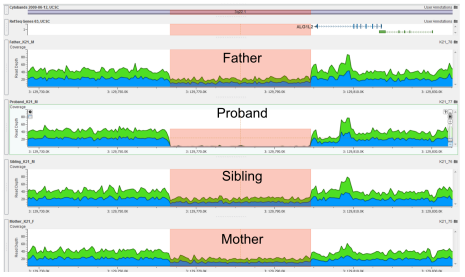

Figure 5c. Genome Browser Screen cut for the Read Depths in the ~ 36 Kb intergenic CNV on 3q22.1 (Pedigree K21).

Figure 5c. Genome Browser Screen cut for the Read Depths in the ~ 22 Kb intergenic CNV on 16p12.3 (Pedigree K21).

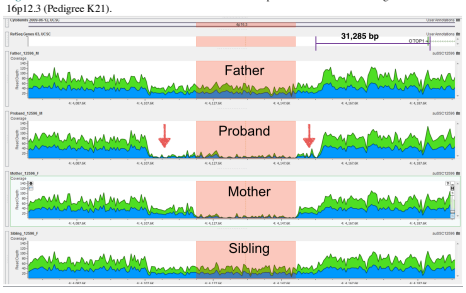

Figure 5c. Genome Browser Screen cut for the Read Depths in the ~ 50 Kb intergenic CNV on 4p16.3 (Pedigree SSC_2).