Uncovering genetic components of a previously un-described syndrome

Jason O'Rawe[1], Yiyang Wu[1], David Mittelman[2], Han Fang[1], Gholson J. Lyon[1,3]

[1]*Cold Spring Harbor Laboratory, Human Genetics, Cold Spring Harbor, NY, 11724,* [2]*Virginia Tech, Department of Biological Sciences, Blacksburg, VA, 24061,* [3]*Utah Foundation for Biomedical Research, UFBR, Salt Lake City, UT, 84106*

Abstract

Background: We describe a whole genome sequencing study of one family containing two affected male children, aged 10 and 12, with severe intellectual disability, autism-like behavior, and very distinctive facial features. High accuracy is of paramount importance in the context of detecting or discovering the genetic influencers of human disease, yet each sequencing and analysis pipeline is still imperfect. In this study, we leverage data from two sequencing platforms and many data processing and downstream analysis pipelines to more confidently identify variants that may play an influential role in the disease state of these two children.

Methods: Whole genome sequencing (WGS) was performed on ten members of this family using the Illumina HiSeq2000 platform, with four (the two affected boys and their parents) being additionally sequenced using the Complete Genomics (CG) WGS platform. CG data analysis was performed by CG, using their version 2.0 pipeline. Illumina reads were mapped to the hg19 reference genome using BWA v.0.6.2-r126, and variant detection was performed using the GATK v. 2.4-9. A second analytical pipeline was used to map the Illumina reads and detect variants using Novoalign and the FreeBayes caller. Disease variant discovery procedures included traditional filtering techniques (using ANNOVAR and Golden Helix SVS), and a statistical framework for identifying the likely disease causing variants (using VAAST).

Results: CG WGS covered >85% of the genome and >90% of the exome, both with 20 or more reads. Illumina WGS covered >90% of the genome with 30 reads or more and with >80% of the bases having a quality score of >30. We found a ~2 to 5-fold difference in the number of variants detected as being relevant for various disease models when using different sets of sequencing data and analysis pipelines. In one instance, employing a 'quad' study design reliably identified three putative variants that followed an X-linked disease model, in TAF1, ZNF41 and ASB12 respectively. However, ZNF41 and ASB12 variants were subsequently found to be false positive findings when the study expanded to include more family members and more data from each.

Presentation of the phenotype

The propositi are two affected male brothers (Fig. 1), aged 10 and 12 respectively, with severe intellectual disability, autism-like behavior, attention deficit issues, and very distinctive facial features (Fig. 1), including broad, upturned nose, sagging cheeks, downward sloping palpebral fissures, relative hypertelorism, high-arched palate, and prominent ears. Their parents are nonconsanguineous and are both healthy, and the family history does not demonstrate any members with anything resembling this current syndrome.

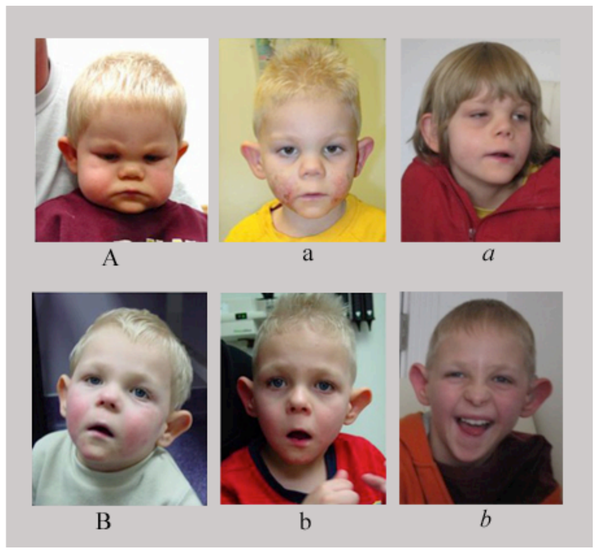


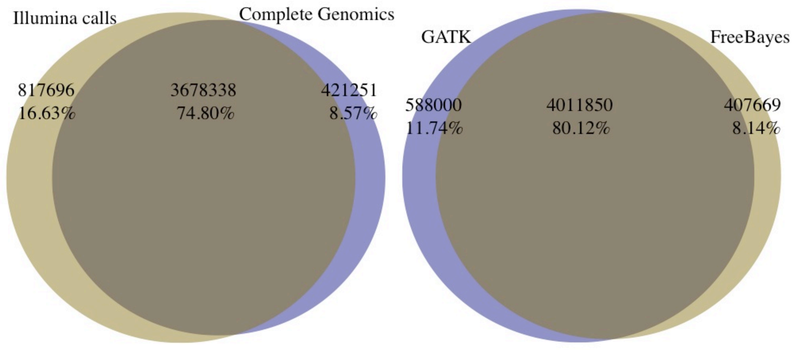
Fig. 1 Facial phenotype of the younger brother at age 19 months (A), 3.5 years (a) and 7 years (a) ; and the elder brother at age 3 years (B), 5 years (b) and 9 years (b). A pedigree displaying intra familial relationships.

Complete Genomics WGS

Whole genomes of the mother, father and both affected boys were sequenced and analyzed with the Complete Genomics WGS sequencing and bioinformatics pipeline. Reads were mapped to the Genome Reference Consortium assembly GRCh37. Due to the proprietary data formats, all the sequencing data QC, alignment and variant calling were performed by CG as part of their sequencing service, using their version 2.0 pipeline. Complete Genomics WGS was optimized to cover 90% of the exome with 20 or more reads and 85% of the genome with 20 or more reads.

Illumina WGS

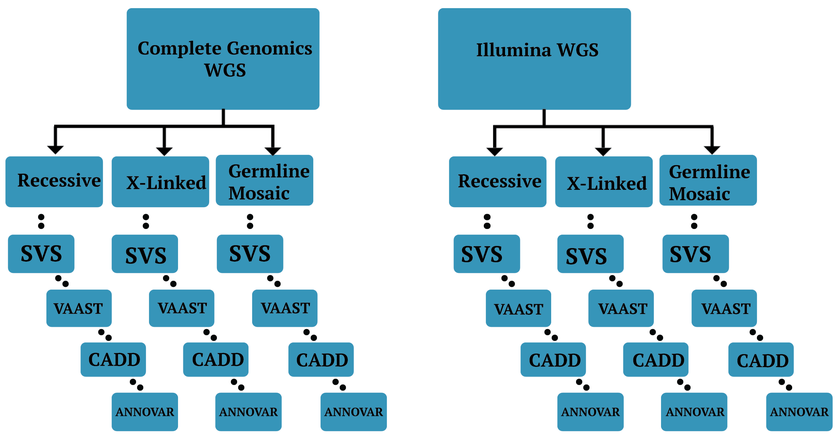
Whole genomes of the entire pedigree were sequenced using the Illumina HiSeq2000. WGS covered >90% of the genome with 30 reads or more and with >80% of the bases having a quality score of >30. Illumina reads were mapped to the hg19 reference genome using BWA v. 0.6.2-r126, and variant detection was performed using the GATK v. 2.4-9. A second analytical pipeline was used to map reads to the hg19 reference genome using Novoalign, and variants were detected using the FreeBayes caller.



Bioinformatics analysis

Bioinformatics analyses of multiple disease model pathways were performed in order to prioritize and identify any putative mutations that might aid in better understanding the pathogenesis of the described syndrome. We performed analyses to interrogate variants conforming to a de-novo, autosomal recessive and x-linked model of disease transmission.

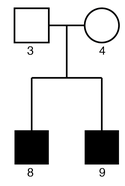
We used several methods to prioritize and identify potentially disease contributing germ-line mutations, including VAAST, Golden Helix SVS, ANNOVAR and CADD, a integrative tool for scoring single nucleotide variants and insertion/deletions. VAAST employs a likelihood-based statistical framework for identifying the most likely disease contributing variants given genomic makeup and population specific genomic information. SVS and ANNOVAR employ more traditional filtering techniques that leverage data stored in public genomic databases.



Using only nuclear family:

55195 Variants were found to be *de-novo* in the two affected boys
122 were coding :

- 107 non-synonymous missense
- 4 splicing
- 3 frame-shift deletions
- 3 frame-shift insertions
- 2 frame-shift substitutions
- 2 stop-gain
- 1 stop-loss



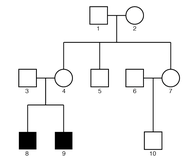
26514 Variants were found to conform to an X-linked disease model
28 were coding:

- 27 non-synonymous missense
- 1 splicing

Using information from a greater portion of the family structure:

17726 Variants were found to be *de-novo* in the two affected boys
40 were coding :

- 32 non-synonymous missense
- 3 splicing
- 2 frame-shift deletions
- 1 stop-loss
- 1 frame-shift insertion
- 1 frame-shift substitution



2824 Variants were found to conform to an X-linked disease model
4 were coding:

- 3 non-synonymous missense
- 1 splicing

Conclusions: Using multiple sequencing and bioinformatics pipelines provides greater power in reducing false positive findings in the context of WGS studies - biological conclusions can shift between sequencing smaller to larger portions of a family.