Simons Simplex Collection whole-genome data

Frequently asked questions

12 August 2016

This list of frequently asked questions (FAQ) aims to cover topics related to the availability of whole-genome sequencing (WGS) data from the Simons Simplex Collection (SSC).

If you have a question that is not answered below, please contact collections@sfari.org or call the Simons Foundation at 646-654-0066.

Description of samples and data

1. **How were the 553 families in the Pilot and Phase 1 batches chosen to be sequenced out of the roughly 2,600 families in the SSC?**
   
   A total of 40 of the families that were sequenced as part of the Pilot study were quad families (defined as both biological parents, the affected child and an unaffected designated sibling) and the other 13 families were trios. The 13 trios and 39 of the 40 quad families were selected on the basis that no *de novo* likely gene-disrupting (LGD) mutations or large copy number variants (CNVs) had been observed in the affected child in previous exome and single nucleotide polymorphism (SNP) microarray analyses\(^1\)\(^-\)\(^7\). One quad family with a known LGD mutation served as a 'control' family for variant detection\(^8\). In addition, among the families chosen for the Pilot study, priority was given to those with older fathers, female probands, and/or male probands who had low intelligence quotient (IQ).

   The subsequent 500 families were randomly selected from among those that were complete quads and for whom no *de novo* LGD mutations or large CNVs had been previously observed in the proband.

   Details about the exact filters that were applied to exclude likely pathogenic mutations and CNVs can be requested from collections@sfari.org.

   Detailed phenotypic information for all individuals is available via SFARI Base. See FAQ #4 for more information.

2. **How were the samples sequenced and what alignments were performed?**

   All WGS data are generated from whole-blood DNA. The genomes of the 40 quad families in the Pilot study were sequenced at the New York Genome Center (NYGC) on an Illumina HiSeq 2500 machine to 30x mean coverage. The 13 trio families in the Pilot study were sequenced at the University of Washington on an Illumina HiSeq 2000 instrument to 51x mean coverage. All subsequent samples are being
sequenced at the NYGC on a HiSeq X Ten machine. All sequencing is via paired-end reads (2 x 150 bp) to 30x mean coverage and will be available in BAM file format containing all passed filter reads and quality scores.

The Pilot and Phase 1 alignment of reads to hg19/NCBI build 37 was done using Burrows-Wheeler Aligner (BWA) software package (BWA-MEM algorithm), and all local realignment and variant calling was done using Genome Analysis Toolkit (GATK) best practices (version 3.4).

Phase 2 and 3 sequences may be analyzed using a different computational pipeline, which will be announced as soon as details are finalized. Pilot and Phase 1 sequences will then be re-analyzed using this new approach to allow for direct comparisons between all families.

In addition to BAM files, single nucleotide variant (SNV), indel and structural variant calls will be available.

3. **Do the BAM files contain unmapped reads?**
   Yes.

4. **How can I find out more about phenotypic data and biospecimens from individuals in the SSC?**
   You can access and request information about the SSC phenotypic data and biospecimens via [SFARI Base](#). If you do not have a Simons Foundation (SF) ID, you must first create one. More information about how to create an ID and access SFARI Base is available here. If you have problems with the login or have questions regarding the SSC, please email collections@sfari.org.

5. **What other genomic and transcriptomic data are available for the SSC?**
   Numerous genomic and transcriptomic datasets (including whole-exome sequencing data, single nucleotide polymorphism genotype data, array-based comparative genomic hybridization data, and gene expression data from lymphoblastoid cell lines) are available for the SSC. More information on these datasets, including citations to the original publications, is available here.

6. **What is the relationship of this effort to MSSNG and the Autism Genetic Resource Exchange (AGRE) WGS efforts?**
   The sequencing of the SSC is completely independent of the MSSNG and AGRE WGS efforts.

**Data access**

7. **Do I need to be a SFARI Investigator to access the WGS data?**
   No, you do not need to have a SFARI grant to be eligible to access and analyze the WGS data. Genetic data, phenotypic data and biospecimens are available to approved researchers via SFARI Base. Details on the application process are available here.

8. **How can I access the WGS data?**
Before you can access the WGS data, you must complete a SFARI Base application. Details on the application process are available here. Additional details will be provided to applicants once their request has been approved.

9. **Do I need to obtain institutional review board (IRB) approval from my institute to access, analyze or store these data in my lab?**
   Yes, all research involving SSC data (including phenotypic and genetic data), biospecimens and recontacting families requires IRB approval or exemption. For some types of requests, you will be asked to upload your approval or exemption letter directly into SFARI Base.

10. **Can researchers perform studies on these data that are not related to autism spectrum and other neurodevelopmental disorders?**
    For the SSC, the original study consents limit research to autism and related neurodevelopmental disorders for all types of requests. We cannot grant exceptions.

11. **How long does it generally take to get approval via SFARI Base?**
    The amount of time varies depending on the type of request. Phenotypic data access for the SSC is granted automatically after the investigator creates an account on SFARI Base (see here for further details) and the Researcher Distribution Agreement and Joinder Agreement are signed. Biospecimen and genetic data requests additionally require SFARI review and, once all documentation is in place, typically take two to three weeks for approval (note that actual receipt of the biospecimens or genetic data may take longer, depending on the type of request).

    Recontacting requests also require both SFARI approval and close collaboration with the SFARI Research Liaison. These requests usually take at least four weeks for approval and typically longer for launch of the project itself. A reasonable expectation would be two to six months from SFARI Base application to project launch.

12. **What is an electronic Researcher Distribution Agreement, and do I need one?**
    The electronic Researcher Distribution Agreement (eRDA) specifies the legal issues pertaining to research data and specimens, and requires approval from the Principal Investigator’s institute. Once the eRDA is approved and executed, researchers must sign the Joinder Agreement to the eRDA. Additional information about this process is available here.

**Managing and analyzing the data**

13. **Can I download the original sequence data to my institute to analyze locally? Or do all computations need to be performed in the cloud?**
    The original sequence files can be downloaded from either the Fermi lab archive or the Simons Foundation Server. The data on Amazon Web Services (AWS) S3 storage are available for computational purposes only and cannot be downloaded.

14. **What analytical tools are available in the cloud computing platform?**
None from SFARI. Investigators are responsible for installing and running their own analysis pipelines.

15. **What is the size of the current data (i.e., 2,174 genomes from 553 families)?**
The estimated size of the data is 300 TB.

16. **Are there fees associated with accessing or analyzing the data in the cloud?**
Yes. Please refer to [AWS pricing information](https://aws.amazon.com) for data analysis charges.

17. **How long will the data be available in the cloud?**
We anticipate that the data will be on AWS S3 for approximately one year. SFARI reserves the right to revise this plan at any time, but we will make every effort to notify current users so that they will have time to plan accordingly.

18. **How can I access whole-exome sequencing (WES) data from the SSC (e.g., to compare with the WGS data)?**
As with WGS data, access to SSC WES data requires agreement with the SFARI genetic data policy. You can request access via [SFARI Base](https://sfari.org). SSC WES data are available for download from our data archive. The data can also be visualized and analyzed via the [WuXi NextCODE SSC portal](https://www.nextcodehealth.com) (a cloud-based database). Web-based training on the use of this portal is available from WuXi NextCODE.

19. **Do I need to provide any final datasets derived from the WGS SSC data to SFARI?**
Yes, you will need to agree to our data-sharing policy before you will be granted access to the data. SFARI reserves the right to determine which datasets to store long term, if at all.

20. **What are SFARI's plans for verifying genetic variants that are identified? Will anyone be able to order DNA (whole-blood or cell-line DNA) to perform experiments to do this, or will it be handled in a centralized way?**
SFARI understands that validation of putative mutations is a critical part of this project, and will coordinate at least some of the confirmation of *de novo* calls with investigators and the NYGC. Requests for DNA to carry out additional validations in individual labs will be assessed on a case-by-case basis. Cell-line DNA requests are likely to be approved on a routine basis, but a higher level of scrutiny will be applied to requests for whole-blood DNA. All requests must be submitted through [SFARI Base](https://sfari.org). See FAQ #4 for more information.

**Publications**

21. **Have any studies been published on the WGS SSC data?**
An initial study of the 53 SSC families included in the Pilot phase has been published.

22. **Is there an embargo period before manuscripts can be published as a result of whole-genome analyses?**
The embargo period has ended for the samples included in the Pilot and Phase I sequencing batches.
A four-month embargo period will be enforced for NIH-funded samples in Phases 2-4 (the embargo period will begin after all samples in a batch have been completed). This embargo will give the principal investigators of those grants the opportunity to publish an initial analysis.

23. How should the SSC WGS data be cited in a publication?
   All investigators must follow SFARI’s Acknowledgement Policy when publishing findings based on SSC data.

24. Do researchers need to send a copy of their manuscript to SFARI in parallel to submitting to a journal?
   Yes, researchers must send a copy of their manuscript to SFARI in parallel to submitting to a journal. Please email a PDF copy of the manuscript to collections@sfari.org.

Future studies and funding requests

25. Can individual samples be resequenced at higher coverage, if requested?
   We do not anticipate routine resequencing of individual samples at higher coverage. If investigators have a compelling scientific argument for systematic resequencing at higher coverage, we encourage them to apply to our Annual Request for Applications (RFA) if they are seeking funds to do so. Investigators with their own funding should apply via SFARI Base to request access to the samples. See FAQ #4 for more information.

26. Can individuals in the SSC be recontacted for additional studies?
   Yes, a subset of participants from the SSC are available for recontacting. Approximately 1,500 families have enrolled in a special cohort at the Interactive Autism Network (IAN) called SSC@IAN. The SSC@IAN is a partnership between SFARI and IAN to create a permanent online research ‘home’ for the SSC families. For more information about recontacting families, please visit the SFARI website or email collections@sfari.org.

27. Can I submit a grant application to SFARI to fund my whole-genome analysis study?
   SFARI announced the Whole-Genome Analysis for Autism Risk Variants RFA in July 2015. Five grants were awarded as a result of this RFA. We do not anticipate funding any additional grants to study the WGS data from the SSC at this stage. If you have a specific hypothesis-driven proposal, SFARI may consider an application submitted as part of our annual Pilot and Research Awards RFA (which is announced each year in the fall) or Explorer Awards RFA (applications are considered on a rolling basis).
References: