The Simons Simplex Collection: A Resource for Identification of Autism Genetic Risk Factors

Gerald D. Fischbach[1,*] and Catherine Lord[2]
[1]Simons Foundation Autism Research Initiative, 160 Fifth Avenue, New York, NY 10010, USA
[2]Department of Psychology, Pediatrics and Psychiatry, University of Michigan, 1111 East Catherine Street, Ann Arbor, MI 48109, USA
*Correspondence: gf@simonsfoundation.org
DOI 10.1016/j.neuron.2010.10.006

In an effort to identify de novo genetic variants that contribute to the overall risk of autism, the Simons Foundation Autism Research Initiative (SFARI) has gathered a unique sample called the Simons Simplex Collection (SSC). More than 2000 families have been evaluated to date. On average, probands in the current sample exhibit moderate to severe autistic symptoms with relatively little intellectual disability. An interactive database has been created to facilitate correlations between clinical, genetic, and neurobiological data.

Introduction

Autism is a developmental disorder characterized primarily by alterations in reciprocal social interaction and restricted interests. Abnormal repetitive behaviors and verbal communication, and various comorbid conditions, also contribute to the broad autism phenotype. Using increasingly expansive definitions, the reported prevalence of autism has increased in recent decades and now approaches the astonishing figure of 1% of the American population. Despite heightened public awareness and concern, causes of autism remain obscure. But there can be no doubt that genetics plays a major role.

Driven by remarkable advances in genome science, the search for genes that enhance the risk of autism has escalated over the past 5 years. However, during the same time period, it has become clear that the genetic landscape of autism is complex, with many genes contributing to the broad autism phenotype (Abrahams and Geschwind, 2008; Cook and Scherer, 2008).

Most cases of autism are sporadic, so one promising approach to the genetic complexity is to search for genetic variants in autistic individuals that are not present in their unaffected parents and siblings. Recent data indicate that probands in such "simplex" families may exhibit submicroscopic deletions and duplications and that the same copy number variants (CNVs) are not present in parents or siblings. Unaffected siblings in simplex families offer ideal controls for identification of de novo CNVs that are truly associated with autism. Case-control studies are less compelling in this regard.

The de novo CNVs discovered to date are rare, occurring in 1% or less of autistic individuals. But there may be many relevant CNVs, and all together, they may account for a significant fraction of idiopathic autism (Christian et al., 2008; Sebat et al., 2007; Weiss et al., 2008). A systematic search should pay dividends but more simplex families are needed.

The Simons Foundation Autism Research Initiative (SFARI), urged on by Michael Wigler at the Cold Spring Harbor Laboratory and by others, embarked on an effort to recruit and carefully evaluate more than 2000 simplex families. We emphasize careful evaluation because the autism phenotype is often not precisely defined at the extremes even though meaningful phenotype/genotype correlations depend on accurate clinical data. Here we describe how the SSC was organized, our methods for ensuring data quality, and our plans for the future.

The SSC extends other autism genetic repositories such as the Autism Genetic Resource Exchange (Geschwind et al., 2001), the Autism Genome Project, and the NIMH repository. These important data sets differ from the SSC in how the data were collected and curated, and in the exclusion criteria employed.

Recruitment of Families

It seemed reasonable that evaluation of families at clinics already serving children with autism and their families would optimize recruitment, dissemination of best practices, and, ultimately, the quality of clinical data. Alliance with providers at a clinic would also optimize chances for follow up studies. Therefore, a coalition was formed of clinics located at Michigan, Yale, Emory, Columbia, Vanderbilt, McGill Washington, and Harvard Universities (Children's Hospital of Boston), and at the Universities of Washington, Illinois (Chicago), Missouri, UCLA, and the Baylor College of Medicine.

A summary to date of the SSC is shown in Table 1. The great majority (>80%) of families include at least one unaffected sibling. The simplex architecture is less certain in trios in which the proband is the only child than it is in quads.

Criteria were defined for case validation and each site was asked to provide at least 20 validated families per quarter. To ensure that the SSC population was independent of earlier cohorts, efforts were made to recruit new families. It was not easy to find new families who met all of our stringent inclusion criteria (see below). Therefore, recruitment tools were developed that included partnerships with local service providers and parent and advocacy groups, and web postings and radio and television ads.

Children are exposed to risks that are greater than risks incurred in activities of daily living when they participate in an evaluation study like the SSC. Therefore parents consented and children assented as required by each local institutional review board. To protect the privacy of participants, Global Unique Identifiers (GUIDs) were constructed from personal information using an algorithm devised in

Table 1. SSC Sample by Family Type and Gender

		Quads		
Proband	Trios	Male Sibling	Female Sibling	Total
Male	306	617	708	1631
Female	52	92	112	256
Total	358	709	820	1887

collaboration with scientists at the NIH (Johnson et al., 2010). Each clinic retained personal identifiers on site and transmitted deidentified GUIDs to a central database (see below).

With GUIDs assigned, individuals can be linked anonymously to other databases including the National Database for Autism Research (NDAR) maintained by the NIMH. They can also be enrolled in follow up studies including new projects conducted by different investigators without concern about duplication or overlap.

The clinic design maximizes the success of longitudinal studies. They are important because a single slice in time, no matter how accurate the data, cannot give an accurate prediction of autism over the lifespan. Follow up studies will become even more important as clinical, genetic, and neurobiological data accumulate.

Infrastructure

A geneticist and a clinical psychologist were appointed as coprincipal investigators at each site. The principal psychologist evaluated families and supervised other clinical psychologists on the site team. Each site also included a project coordinator and a data manager. Best practices, rates of accrual, and recent advances in autism science were shared in monthly phone calls that included members of each site team and members of the SFARI staff. Information was also shared in site visits and at biannual group meetings.

All together, these efforts created a sense of shared purpose and they are, in large part, responsible for our success to date. As shown in Figure 1, we achieved a high rate of accrual that has been maintained throughout the study.

Evaluation of Probands and Families

Probands were evaluated with a battery of diagnostic measures, including the Autism Diagnostic Interview – Revised (ADI-R) (Lord et al., 1994) and the Autism Diagnostic Observation Schedule (ADOS) (Lord et al., 2000). Other instruments provided additional measures of the core features of autism, as well as of intellectual ability (verbal and nonverbal), adaptive behavior, emotional and behavior problems, motor function, and language. A description of instruments employed can be found at https://sfari.org/ssc-instruments. A comprehensive family medical history was obtained that included the proband's prenatal and perinatal history, developmental milestones, immunizations, medications, dietary supplements, and common behavioral treatments. Emphasis was placed on common "comorbidities" including gastrointestinal complaints, sleep irregularities, and seizures. In addition, questions were asked about genetic, autoimmune, and psychiatric disorders in members of the extended family.

Probands were excluded who were younger than 4 years of age or older than 18. Probands were also excluded for conditions that might compromise the validity of diagnostic instruments, such as nonverbal mental age below 18 months, severe neurological deficits, birth trauma, perinatal complications, or genetic evidence of fragile X or Down syndromes. A complete description of exclusion/inclusion criteria can be found at http://sfari.org. The four most common reasons for exclusion were as follows: candidate did not meet criteria for Autism Spectrum Disorder (ASD); primary relatives were on the autism spectrum; medically significant perinatal incidents; and low mental age.

Table 2 shows descriptive data for the first 1887 probands. Measures of adaptive function, behavior-emotional problems, and symptoms of autism were examined in parents and siblings as well as probands.

Thus, the SSC represents a unique, well-described sample of able children and adolescents with relatively severe ASD, as indicated by ADI-R and ADOS Calibrated Severity Scores (Gotham et al., 2009).

Reliability of Data

To maximize the consistency of clinical observations across sites, each clinician was trained in administration of the ADOS and ADI-R to achieve research reliability as judged by expert clinicians. Most clinicians who had not previously received research training required 4–6 months of practice. Videotapes of interviews were exchanged to ensure that reliability requirements were met and maintained throughout the study.

A team of consultants composed of four clinical psychologists from the University of Michigan assisted sites with training and case validation using logical checks of variables across measures that would signal inconsistencies in coding or data entry. Error rates were very low, averaging less than 0.50 errors/1000 data points. Most errors could be corrected immediately, resulting in an unusually clean data set for a multisite study of this size.

Our emphasis on rigorous training and careful data curation paid off: scores based on direct observation or on questionnaires obtained at the various sites fell within a narrow range. One site deviated significantly from the overall mean, so families at that site will be reevaluated. Despite agreement in scores on standardized measures, consistent differences between sites in overall clinical impression were reported. With the same scores on standardized tests, some observers preferred the label Autism, while others preferred Asperger syndrome or Pervasive Developmental Disorder-Not Otherwise Specified. Thus, traditional ASD diagnostic subcategories were not useful in describing SSC probands. Our findings are reflected in the new classification, Autism Spectrum Disorders, proposed for the next version of the Diagnostic and Statistical Manual V of the American Psychiatric Association.

Resources Available to the Research Community

A blood sample was collected from each study participant. DNA was extracted from blood cells, plasma was stored, and cell lines were established from transformed lymphoblasts at the Rutgers University Cell and DNA Repository (RUCDR). The RUCDR also analyzed each sample to confirm parentage and gender and to rule out the fragile X syndrome.

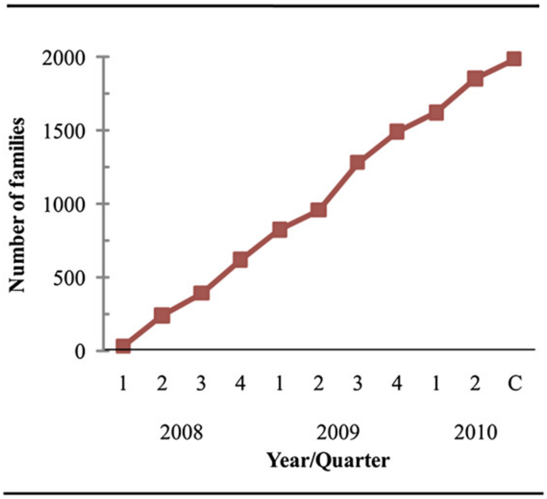

Figure 1. Accrual Rates by Quarter
Validated families with whole-blood DNA. The accumulation of lymphoblastoid cell lines paralleled this curve with a short delay. C = current as of submission.

In collaboration with Prometheus Research LLC, SFARI investigators developed a software platform, SFARI Base, to support data acquisition, validation, and distribution. Notable innovations include a query tool for selecting a defined cohort, measures of interest, and biological specimens. A Data Navigator displays all data from all individuals as well as correlations among the variables. In addition, summary variables that aggregate multiple raw variables into a single value are available. They are intended to help create links between clinical, genetic, and neurobiological data, and they will be updated regularly depending on user demand. Summary variables are particularly useful when dealing with multiple informants and overlapping constructs.

Data on CNVs from the first whole-genome scans of SSC participants based on NimbleGen chips and on Illumina chips are available at https://sfari.org/sfari-base. Investigators can also obtain the SSC clinical data set and relevant DNA samples by registering at https://sfari.org/sfari-base.

Looking Ahead

Our initial aim in creating the SSC was to facilitate chip-based searches for rare, highly penetrant CNVs. Additional families will be evaluated in the coming year, and high-throughput DNA sequencing will be pursued to identify small de novo variants below the resolution of existing chips, and single nucleotide polymorphisms that are candidate autism risk factors. As the SSC grows, inherited variants and more common variants of smaller effect may be discovered as well.

To determine which de novo variants are truly autism risk factors, several questions must be asked. Is the variant recurrent? Does it disrupt gene function, directly or indirectly, perhaps via regulatory elements in noncoding regions? How many genes are involved? Does it produce a relevant behavioral, physiological, or anatomical phenotype in mouse models or in other species? Does genetic imprinting in critical brain regions impact the manifestation of candidate genetic variants?

These questions have been answered satisfactorily for only a handful of genes. A list of about 200 candidates culled from the literature by S. Basu of Mindspec Inc. appears at http://gene.sfari.org/. A curated list in which the genes are rank ordered by stringent criteria will appear at that web site in the near future. Comments will be invited in the hopes that this can be made a collaborative, dynamic effort.

Despite the complexity of autism genetics, there is reason for optimism even in the short term. The multiplicity of genetic variants may lead to one or a few common signaling pathways. The system of synaptic membrane proteins

Table 2. Proband Age and Standardized Test Scores by Gender

		Males (n = 1631)		Females (n = 256)	
		Mean (SD)	Range	Mean (SD)	Range
	Chronological age (years)	8.91 (3.48)	4–18	9.11 (3.69)	4–18
	Verbal IQ*	80.44 (30.03)	5–153	76.03 (32.40)	11–167
	Nonverbal IQ	87.58 (24.75)	9–161	79.16 (26.15)	23–148
ADI-R	Social interaction	19.96 (5.68)	8–30	20.10 (6.13)	8–30
	Communication – verbal[a]	16.30 (4.21)	6–26	16.35 (4.36)	6–26
	Communication – nonverbal	9.07 (3.39)	0–14	8.96 (3.77)	1–14
	Restricted/repetitive behaviors	6.55 (2.53)	0–12	6.20 (2.48)	0–12
ADOS	Calibrated severity scores[b]	7.39 (1.71)	4–10	7.34 (1.74)	4–10
	Social + communication	13.12 (4.19)	5–24	13.59 (4.63)	4–24
	Social affect[2]	10.89 (4.02)	3–20	11.37 (4.27)	3–20
	Restricted/repetitive behaviors	3.94 (2.06)	0–8	3.86 (2.13)	0–8
	Vineland	74.49 (11.59)	27–115	71.36 (12.13)	32–101

n = 1877.

[a] The verbal domain is not calculated for participants who receive a 1 or a 2 on ADI-R item 30; therefore, n = 1470 males, 230 females.

[b] Calibrated Severity Scores and Social Affect totals are not calculated for Module 4; therefore, n = 1580 males, 247 females.

and intracellular proteins that regulate their distribution and function is a promising example. This pathway must regulate the delicate balance between synaptic excitation and inhibition in crucial neural circuits that are active in the social brain. The same pathway is likely involved in forms of synaptic plasticity that underlie the ability to learn and alter behavior. Thus, beyond the level of molecular neurobiology, genetic variants may provide clues about autistic behaviors.

As signaling pathways are explored, the genetic variants may lead to biomarkers that will greatly facilitate detailed phenotype/genotype correlations. This is particularly the case as longitudinal studies progress, as detailed medical information is gathered, and as advances in neuroimaging and cognitive neuroscience suggest new assays of social cognition.

Convincing genetic variants will also allow the converse approach: more detailed genotype/phenotype correlation. A new effort called the Simons Variation in Individuals Project (Simons VIP) is underway under the guidance of W. Chung at Columbia and J. Spiro at the Simons Foundation. Deletions and duplications at chromosome 16p 11.2, one of the most recurrent and, hence convincing, autism-associated CNVs, will be the first variant studied. Individuals identified by clinical laboratories throughout the country and by other sources will be invited for detailed clinical evaluation at one of three sites. Lessons learned in assembling the SSC will be valuable in operationalizing the Simons VIP.

In sum, the SSC illustrates how planning and cooperation can lead to the rapid accumulation of high-quality data from a large number of families gathered at geographically dispersed sites. It has already triggered an enormous amount of research. We anticipate that the SSC will be an extraordinary resource for autism researchers, for those interested in other neuropsychiatric disorders that share features with autism, and, indeed, for all neuroscientists.

ACKNOWLEDGMENTS

We are grateful to the families at the participating Simons Simplex Collection (SSC) sites; the principal investigators at each site (A. Beaudet, R. Bernier, T. Cicero, J. Constantino, E. Cook, E. Fombonne, D. Geschwind, D. Grice, A. Klin, R. Kochel, D. Ledbetter, C. Martin, D. Martin, R. Maxim, J. Miles, O. Ousley, B. Peterson, J. Piggot, C. Saulnier, M. State, W. Stone, J. Sutcliffe, C. Walsh, and E. Wijsman); the coordinators and staff at the SSC sites; the SFARI staff (M. Greenup, M. Benedetti, J. Spiro, and S. Johnson); the UMACC staff (J. Tjernagel, L. Green-Snyder, and the consultants); and E. Petkova and V. Hus for help in preparation of this manuscript. Funds for the SSC were provided by the Simons Foundation.

REFERENCES

Abrahams, B.S., and Geschwind, D.H. (2008). Nat. Rev. Genet. *9*, 341–355.

Christian, S.L., Brune, C.W., Sudi, J., Kumar, R.A., Liu, S., Karamohamed, S., Badner, J.A., Matsui, S., Conroy, J., McQuaid, D., et al. (2008). Biol. Psychiatry *63*, 1111–1117.

Cook, E.H., Jr., and Scherer, S.W. (2008). Nature *455*, 919–923.

Geschwind, D.H., Sowinski, J., Lord, C., Iversen, P., Shestack, J., Jones, P., Ducat, L., and Spence, S.J.; AGRE Steering Committee. (2001). Am. J. Hum. Genet. *69*, 463–466.

Gotham, K., Pickles, A., and Lord, C. (2009). J. Autism Dev. Disord. *39*, 693–705.

Johnson, S.B., Whitney, G., McAuliffe, M., Wang, H., McGreedy, E., Rozenblit, L., and Evans, C.C. (2010). J. Am. Med. Inform. Assoc., in press. Published online July 17, 2010. 10.1136/jamia.2009.002063.

Lord, C., Rutter, M., and Le Couteur, A. (1994). J. Autism Dev. Disord. *24*, 659–685.

Lord, C., Risi, S., Lambrecht, L., Cook, E.H., Jr., Leventhal, B.L., DiLavore, P.C., Pickles, A., and Rutter, M. (2000). J. Autism Dev. Disord. *30*, 205–223.

Sebat, J., Lakshmi, B., Malhotra, D., Troge, J., Lese-Martin, C., Walsh, T., Yamrom, B., Yoon, S., Krasnitz, A., Kendall, J., et al. (2007). Science *316*, 445–449.

Weiss, L.A., Shen, Y., Korn, J.M., Arking, D.E., Miller, D.T., Fossdal, R., Saemundsen, E., Stefansson, H., Ferreira, M.A., Green, T., et al; Autism Consortium. (2008). N. Engl. J. Med. *358*, 667–675.