Research in Autism Spectrum Disorders 32 (2016) 80–83

Contents lists available at ScienceDirect

Research in Autism Spectrum Disorders

journal homepage: http://ees.elsevier.com/RASD/default.asp

Brief report

Can we accelerate autism discoveries through crowdsourcing?

Maude M. David[a], Brooke A. Babineau[b], Dennis P. Wall[a,c,*]

[a] Stanford University, School of Medicine, Department of Pediatrics, Stanford, CA 94305, USA
[b] Stanford University, School of Medicine, Department of Neurosurgery, Stanford, CA 94305, USA
[c] Stanford University, School of Medicine, Department of Biomedical Data Science, Stanford, CA 94305, USA

ARTICLE INFO

Article history:
Received 7 March 2016
Received in revised form 30 August 2016
Accepted 1 September 2016
Available online 21 September 2016

Keywords:
Autism
Autism spectrum disorder
Genome-environment interactions
Crowdsourcing

ABSTRACT

Autism is a dramatically expanding public health challenge. The search for genomic variants underlying the disease concomitantly accelerated over the last 5 years, leading to a general consensus that genetics can explain between 40% and 60% of the symptomatic variability seen in autism. This stresses both an urgent need to continue devoting resources to the search for genetic etiologies that define the forms of autism, and an equal need for attention to the interactive roles of the environment. While some environmental factors have been investigated, few studies have attempted to elucidate the combination and interplay between gene and environment to gain clear understanding of the mechanisms by which environmental factors interact with genetic susceptibilities in Autism Spectrum Disorder. Due to financial constraints as well as recruitment protocols limited by geography, such studies have been challenging to implement. We discuss here how crowdsourcing approaches can overcome these limitations.

Despite its rising prevalence, the causes of autism, likely a combination of genetic and environmental factors, remain unclear. Because autism appears to be a highly heritable condition, numerous high-powered efforts have banked genomic data on thousands of carefully phenotyped individuals in the pursuit of genetic markers to explain the molecular pathologies of autism. These large genomic efforts are collectively starting to sharpen the molecular picture of autism, which now includes at least 60 genes harboring variants of high interest (De Rubeis et al., 2014; Iossifov et al., 2014; Robinson et al., 2016). This work has led to the understanding that inherited variation from both common and rare alleles explains approximately 40% of the risk for developing autism (95% confidence interval: 8%–84%, estimated via the squared correlation of outcome with predicted probabilities from logistic regression) (Stein, Parikshak, & Geschwind, 2013).

The environment – e.g., environment of the cell, brain, child, and/or parents – must explain a sizable amount of the remainder, potentially as much as 60%. However, studies of the environmental contributions to autism lag far behind studies of the genome. This is due at least in part to the difficulties in design of well controlled studies, challenges with recruitment, costs due to need for large sample sizes, and choice of environmental factor on which to focus, to name a few. Faced with these challenges, the field has demonstrated statistically significant, but still tenuous links to several factors. For example, systematic exploration of almost 20 years (1984–2003) of medical records from over 400,000 subjects in Sweden has pinpointed advanced parental age as associated with increased risk of autism, particularly for autism with intellectual disability (Idring et al., 2014). Nationwide registered medical records from Denmark revealed links to increased risk of ASD diagnosis in the offspring from maternal infection during pregnancy, specifically viral infection during the first trimester and

* Corresponding author at: Stanford University, School of Medicine, Department of Pediatrics, Stanford, CA 94305, USA.
E-mail address: dpwall@stanford.edu (D.P. Wall).

http://dx.doi.org/10.1016/j.rasd.2016.09.001

bacterial infection during the second trimester, (Atladóttir et al., 2010). Shelton et al. found the proximity of agricultural pesticide during pregnancy associated with autism by linking the physical addresses of almost a thousand participants with the data from commercial pesticide application (Shelton et al., 2014). Several studies have also shown that fetal exposure to the valproic acid (a drug used to prevent epilepsy), which can be teratogenic, may lead to autism (Chomiak, Turner, & Hu, 2013; Williams et al., 2001). Finally, the growing understanding of a role for the gut microbiome in behavioral disorders, i.e. the "gut-brain" axis, cannot be ignored; numerous studies have reported microbial taxa specific to autism (Kang et al., 2013; Macfabe, Cain, Boon, Ossenkopp, & Cain, 2011; Finegold et al., 2010). Mouse models of autism with aytpical microbiota compositions compared to control can revert to typical floral compositions through treatment with the human commensal *Bacteroides fragilis*. This ingestion of probiotics also rectified gut permeability, microbial composition, and ameliorated deficits in communicative, stereotypic, anxiety-like and sensorimotor behaviors (Hsiao et al., 2013), supporting the potential role of environmental factors in the onset and maintenance of autism-related behaviors.

It is clear that neither the genome nor the environment alone will explain the underlying causes of autism, yet the research on environmental and genetic factors remain decoupled. This decoupled state of our autism data presents a substantial barrier to initializing the parameters of the G(genome) + E(environment) = P(phenotype) equation. However, coupling of the two requires: I. a reasonable selection of environmental measures, II. prospective recruitment of large and well-controlled study participants, III. robust recruitment infrastructure, and IV. enormous financial support, particularly if done in prospective clinical trials.

In order to overcome these challenges and unravel specific mechanisms related to the differential impact of the environment with specific genetic variants, several researchers have successfully chosen targeted approaches. For example, using data from dizygotic twins, it has been shown that beta2-adrenergic receptor can be differentially affected by exposure to neuro-active drugs, and is significantly associated with the autism phenotype (Connors et al., 2005). Furthermore, several studies have demonstrated than mutations in the SERT gene (encoding for serotonin transporter) result in increased anxiety in adult life when combined with a stressful environment during development (Gross & Hen, 2004; Lesch, 2001), and that long/short allele polymorphism upstream of the same gene coupled with prenatal stress is associated with a decrease of social interaction in mice, producing autism-like behaviors in offspring (Jones et al., 2010). Other examples from animal models include work investigating the interaction between genetic background of an inbred mouse strain with an environmental insult, which revealed increased severity of the behavioral phenotype in genetically susceptible mice treated with an immune challenge *in utero* (Schwartzer et al., 2013).

While targeted studies have helped to pinpoint interactions between environmental factors and individual genes, there is a pressing need for projects that measure both environmental factors and genetic variations across the whole genome. Standard prospective clinical trials no longer represent a viable avenue for making clinically valuable discoveries at the cross section of the environment and the genome. The sample size to achieve sufficient power is in the tens of thousands of individuals, making time and costs both out of bounds. A standard prospective clinical trial approach would, need to engage multiple clinical sites, substantial clinical coordination across sites, and a coordinated data collection infrastructure. The practical challenges to effective recruitment, phenotyping, and data collection are enormous. Untangling interaction between genome and environment will also require rigorous statistical standards, and the development of methods to determine appropriate correction and models to use, as existing methods for discovery of gene(G) $\times$ environment(E) interaction are known to have low power, particularly when one corrects for testing multiple variations across the genome. One solution already suggested could be to apply a filter step in order to select the variants most likely to be involved in a G $\times$ E interaction (Zhang, Lewinger, Conti, Morrison, & Gauderman, 2016), or aggregating information from multiple loci across a genetic region (Fan & Lo, 2013). In order to apply any of these methods and amalgamate ample data we must look for other, less difficult ways to collect data to couple the environment with the genome in order complete the G x E = P equation: there is promise in crowdsourcing.

Families with autism are overwhelmingly willing to participate in online research efforts, with networks like Interactive Autism Network, Autism Match and MyAutismTeam, representing a potential untapped opportunity for acquisition of big data. The scientific research community may therefore be able to use online approaches, leverage social networks and patient-advocacy groups to enroll thousands of internet-active families with children on the autism spectrum quickly, enhancing our recruitment efforts via completely crowdsourced clinical studies. Crowdsourced approaches to biological data collection through web portalsare starting to show promise of speed, accuracy, and size, including for example, The American Gut Program, which recruited over 4200 participants in $\sim$1.5 years, harvesting both participant metadata and providing open access to the sequencing results. As an ancillary byproduct, social network recruitment has potential for longitudinal retention of subjects, which enables tracking of progress during therapeutic programs that often have enormous but still largely unmeasured benefits. In the same vein, Apple introduced an open-source framework called Apple ResearchKit that allows researchers to create powerful applications for medical research. Similar efforts have emerged in the autism community itself including the Simons Foundation SPARK initiative to recruit 50,000 individuals with autism via the crowd and the Early Life Exposures Assessment Tool developed by Autism Speaks in order to collect environmental exposure data.

Among the numerous advantages afforded by this approach are substantial cost reductions by eliminating the need for physical space and onsite clinical staff, time savings afforded toe participants and their families, and the comfort and simplicity of acquiring samples remotely, i.e. directly from the patients' homes and on their own schedules. Furthermore, this approach reduces the overhead costs per study participant without restricting the information that can be gathered,

enabling recruitment of very large and statistically powerful cohorts. Scientists will be able to generate better patient groups, with less concern that overly stringent study criteria will preclude sufficient recruitment. Finally, this approach allows the analysis of metadata, such as cross-referencing patient addresses with maps of environmental factors such as pollution, arriving faster to sharpened hypothesis that can be tested individually in more traditional ways at a far more manageable pricepoint.

Crowdsourced approaches are not without challenges. One concern in employing these approaches to examine causes of autism is the lack of personal interactions in our model. As we rely on self-reporting of the participants, we need to ensure that the self-reported diagnosis is clinically reliable (Kohane & Eran, 2013). There are several approaches to provide a risk assessment for autism that are compatible with an online interface and remote interaction. The Interactive Autism Network (IAN) for example was able to validate the use of self-reported professional autism diagnosis to the registry's database (Daniels et al., 2012) (Lee et al., 2010). Machine learning classification of parent reported information and features extracted from short home recordings of the child's behaviors have shown promise for remote detecion and screening (Duda et al., 2014; Fusaro et al., 2014; Kosmicki et al., 2015; Wall et al., 2012). Such classifiers can be run quickly to produce a score that indicates both confidence in the classification and phenotypic severity with acceptably high accuracy relative to the clinical best estimate (Duda et al., 2016) and are therefore useful as confirmatory support for the self-reported diagnosis. Such an approach has advantages due to its crowdsourcing capability, allowing online phenotyping to minimize time while maximizing recruitment reach, but also because the home videos provide persistent digital observations of the children in their natural environment, where behaviors can deviate in important ways from those exhibited in the clinic.

Regardless, biases in self-reported information still could easily undermine the benefits of rapid and large scale data acquisition from the crowd to power larger studies that address the complexity of autism as a genetic and environmental disorder. Deviations between self-report and medical records data have been shown, for example, self-report of antibiotic intake differs from the same kind of reporting captured within an electronic medical record (Caverly et al., 2016). Such potential deviation must be controlled through creative platform design. Ideally, the design will engage sensors on ubiquitous devices (e.g., smartphones) that can record objective measures in increasingly more passive ways, and/or by defaulting wherever feasible to the capture of direct observations (e.g. images, videos) rather than surveys alone (Stark, Kumar, Longhurst, & Wall, 2016). This will safeguard against such factors like unpredictable variations in subjective reporting and survey fatigue that often plague the crowd-questionnaire approach. The efforts of researchkit and researchstack from Apple and Google, respectively, highlight the attention now being paid to finding the right platform design to empower the crowd approach and suggest that we will soon have further demonstration that the mobilized crowd can augment the standards in place today to dramatically accelerate discovery.

Another potential drawback of crowd methods for data acquisition is the potential for bias against populations that do not have easy access to mobile technology, especially lower income populations and the elderly. This concern also exists for classical clinical trials, wherein caregivers raising a child with special needs are required to spend hours at a clinic during business hours near the clinics that may be geographically disconnected from important sectors of the study population. Given the increasing ubiquity of access to technology, it is likely that crowdsourced studies will be significantly more accessible to diverse social and economic groups than than classical trials. However, inclusion of such groups must become explicit components of the design and form factor for any mobile, crowd-based study.

Despite the challenges and potential concerns, we believe that a sandbox of highly coupled data constitute a powerful way to parse the complexity of autism into subcomponents that have detectable diagnostic markers, prognostic markers, and that reveal targets for therapeutic intervention.The literature has focused largely on gene-by-environment interaction, where the environmental variable and the gene tested were hypothesized *a priori*. Conclusions were drawn from a candidate gene or G X E association, resulting often in pertinent and useful discovery but sometimes difficult to replicate (Duncan & Keller, 2011). However, large data-driven studies using a combination of GWAS (Genome-wide Association Study) and EWAS (Environment-wide Association Study) were able to spot genes and factors of interest linked to environmental variables, stressing the importance of building larger and more encompassing databases (Nickels et al., 2013; Patel, Chen, & Butte, 2012).

Reaching that global scale and constructing a sandbox of data of the size needed to parse the complexity of autism will require recruitment of thousands of individuals. Arguably the most effective way to get thousands to participate is to reach out to families across the world through a new avenue of recruitment, social media and the autism crowd. To build this sandbox of coupled data we need to look beyond our standard methods of clinical studies and begin to utilize new technologies to engage families across the globe affected by autism. Increasing participation through crowdsourcing will lead to large data sets that can allow the unbiased detection of relevant factors contributing to the autism phenotype. For example, triangulating data and phenotype with environmental factors may unravel any association autism has with potential chemical exposure. These factors can be further tested using targeted approaches, likely leading to robust and replicable findings that can be more efficiently translated into therapeutic approaches.

Acknowledgements

This work was supported by the Hartwell Foundation's Autism Technology and Research Initiative (iHART) and by the Stanford Predictives and Diagnostics Accelerator grant from The Stanford Center for Clinical and Translational Research and Education.

References

Atladóttir, H. O., Thorsen, P., Østergaard, L., Schendel, D. E., Lemcke, S., Abdallah, M., et al. (2010). Maternal infection requiring hospitalization during pregnancy and autism spectrum disorders. *Journal of Autism and Developmental Disorders*, *40*(12), 1423–1430. http://dx.doi.org/10.1007/s10803-010-1006-y.

Caverly, L. J., Caverly, T. J., Kalikin, L. M., Foster, B. K., Simon, R. H., & LiPuma, J. J. (2016). Episodic oral antibiotic use in CF: Discordance between the electronic medical record and self-report. *Journal of Cystic Fibrosis*.

Chomiak, T., Turner, N., & Hu, B. (2013). What we have learned about autism spectrum disorder from valproic acid. *Pathology Research International*, *71275*, 8. http://dx.doi.org/10.1155/2013/712758.

Connors, S. L., Crowell, D. E., Eberhart, C. G., Copeland, J., Newschaffer, C. J., Spence, S. J., et al. (2005). Beta2-adrenergic receptor activation and genetic polymorphisms in autism: Data from dizygotic twins. *Journal of Child Neurology*, *20*(11), 876–884.

Daniels, A. M., Rosenberg, R. E., Anderson, C., Law, J. K., Marvin, A. R., & Law, P. A. (2012). Verification of parent-report of child autism spectrum disorder diagnosis to a web-based autism registry. *Journal of Autism and Developmental Disorders 42*(2), 257–265. http://doi.org/10.1007/s10803-011-1236-7.

De Rubeis, S., He, X., Goldberg, A. P., Poultney, C. S., Samocha, K., Cicek, A. E., et al. (2014). Synaptic, transcriptional and chromatin genes disrupted in autism. *Nature 515*(7526), 209–215. http://doi.org/10.1038/nature13772.

Duda, M., Kosmicki, J. A., & Wall, D. P. (2014). Testing the accuracy of an observation-based classifier for rapid detection of autism risk. *Translational Psychiatry*, *4*, e424. http://dx.doi.org/10.1038/tp.2014.65.

Duda, M., Daniels, J., & Wall, D. P. (2016). Clinical evaluation of a novel and mobile autism risk assessment. *Journal of Autism and Developmental Disorders*, *46* (June (6)), 1953–1961 [PMCID: PM C48 60199].

Duncan, L. E., & Keller, M. C. (2011). A critical review of the first 10 years of candidate gene-by-environment interaction research in psychiatry. *The American Journal of Psychiatry 168*(10), 1041–1049. http://doi.org/10.1176/appi.ajp.2011.11020191.

Fan, R., & Lo, S.-H. (2013). A robust model-free approach for rare variants association studies incorporating gene-gene and gene-environmental interactions. *PLoS One 8*(12), e83057. http://doi.org/10.1371/journal.pone.0083057.

Finegold, S. M., Dowd, S. E., Gontcharova, V., Liu, C., Henley, K. E., Wolcott, R. D., et al. (2010). Pyrosequencing study of fecal microflora of autistic and control children. *Anaerobe*, *16*(4), 444–453. http://dx.doi.org/10.1016/j.anaerobe.2010.06.008.

Fusaro, V. A., Daniels, J., Duda, M., Deluca, T. F., D'Angelo, O., Tamburello, J., et al. (2014). The potential of accelerating early detection of autism through content analysis of YouTube videos. *Public Library of Science 9*(4), e93533. http://doi.org/10.1371/journal.pone.0093533.

Gross, C., & Hen, R. (2004). Genetic and environmental factors interact to influence anxiety. *Neurotoxicity Research*, *6*(6), 493–501.

Hsiao, E. Y., McBride, S. W., Hsien, S., Sharon, G., Hyde, E. R., McCue, T., et al. (2013). Microbiota modulate behavioral and physiological abnormalities associated with neurodevelopmental disorders. *Cell*, *155*(7), 1451–1463. http://dx.doi.org/10.1016/j.cell.2013.11.024.

Idring, S., Magnusson, C., Lundberg, M., Ek, M., Rai, D., Svensson, A. C., et al. (2014). Parental age and the risk of autism spectrum disorders: Findings from a Swedish population-based cohort. *International Journal of Epidemiology*, *43*(1), 107–115. http://dx.doi.org/10.1093/ije/dyt262.

Iossifov, I., O'Roak, B. J., Sanders, S. J., Ronemus, M., Krumm, N., Levy, D., et al. (2014). The contribution of de novo coding mutations to autism spectrum disorder. *Nature*, *515*(7526), 216–221. http://dx.doi.org/10.1038/nature13908.

Jones, K. L., Smith, R. M., Edwards, K. S., Givens, B., Tilley, M. R., & Beversdorf, D. Q. (2010). Combined effect of maternal serotonin transporter genotype and prenatal stress in modulating offspring social interaction in mice. *International Journal of Developmental Neuroscience: the Official Journal of the International Society for Developmental Neuroscience*, *28*(6), 529–536. http://dx.doi.org/10.1016/j.ijdevneu.2010.05.002.

Kang, D.-W., Park, J. G., Ilhan, Z. E., Wallstrom, G., LaBaer, J., Adams, J. B., et al. (2013). Reduced incidence of prevotella and other fermenters in intestinal microflora of autistic children. *PLoS One 8*(7) . http://doi.org/10.1371/journal.pone.0068322.

Kohane, I. S., & Eran, A. (2013). Can we measure autism? *Science Translational Medicine*, *5*(209) . http://dx.doi.org/10.1126/scitranslmed.3007340 [209ed18–209ed18].

Kosmicki, J. A., Sochat, V., Duda, M., & Wall, D. P. (2015). Searching for a minimal set of behaviors for autism detection through feature selection-based machine learning. *Translational Psychiatry 5*, e514. http://doi.org/10.1038/tp.2015.7.

Lee, H., Marvin, A. R., Watson, T., Piggot, J., Law, J. K., Law, P. A., et al. (2010). Accuracy of phenotyping of autistic children based on Internet implemented parent report. *American Journal of Medical Genetics. Part B, Neuropsychiatric Genetics: the Official Publication of the International Society of Psychiatric Genetics 153B*(6), 1119–1126. http://doi.org/10.1002/ajmg.b.31103.

Lesch, K. P. (2001). Molecular foundation of anxiety disorders. *Journal of Neural Transmission*, *108*(6), 717–746 [Vienna, Austria: 1996].

Macfabe, D. F., Cain, N. E., Boon, F., Ossenkopp, K.-P., & Cain, D. P. (2011). Effects of the enteric bacterial metabolic product propionic acid on object-directed behavior, social behavior, cognition, and neuroinflammation in adolescent rats: Relevance to autism spectrum disorder. *Behavioural Brain Research*, *217* (1), 47–54. http://dx.doi.org/10.1016/j.bbr.2010.10.005.

Nickels, S., Truong, T., Hein, R., Stevens, K., Buck, K., Behrens, S., et al. (2013). Evidence of gene-environment interactions between common breast cancer susceptibility loci and established environmental risk factors. *PLoS Genet.*, *9*(3), e1003284 [http://doi.org/10.1371/journal.pgen.1003284].

Patel, C. J., Chen, R., & Butte, A. J. (2012). Data-driven integration of epidemiological and toxicological data to select candidate interacting genes and environmental factors in association with disease. *Bioinformatics 28*(12), i121–i126. . [Oxford, England] http://doi.org/10.1093/bioinformatics/bts229.

Robinson, E. B., St Pourcain, B., Anttila, V., Kosmicki, J. A., Bulik-Sullivan, B., Grove, J., et al. (2016). Genetic risk for autism spectrum disorders and neuropsychiatric variation in the general population. *Nature Genetics*, *48*(5), 552–555. http://dx.doi.org/10.1038/ng.3529.

Schwartzer, J. J., Careaga, M., Onore, C. E., Rushakoff, J. A., Berman, R. F., & Ashwood, P. (2013). Maternal immune activation and strain specific interactions in the development of autism-like behaviors in mice. *Translational Psychiatry*, *3*, e240. http://dx.doi.org/10.1038/tp.2013.16.

Shelton, J. F., Geraghty, E. M., Tancredi, D. J., Delwiche, L. D., Schmidt, R. J., Ritz, B., et al. (2014). Neurodevelopmental disorders and prenatal residential proximity to agricultural pesticides: The charge study. *Environmental Health Perspectives* . http://doi.org/10.1289/ehp.1307044.

Stark, D. E., Kumar, R. B., Longhurst, C. A., & Wall, D. P. (2016). The Quantified Brain: A Frameworkfor Mobile Device-Based Assessment of Behavior and Neurological Function. *Applied Clinical Information*, *7*(May (2)), 290–298. http://dx.doi.org/10.4338/ACI-2015-12-LE-0176 eCollection2016. PubMed PMID: 27437041; PubMed Central PMCID: PMC4941840.

Stein, J. L., Parikshak, N. N., & Geschwind, D. H. (2013). Rare inherited variation in autism: Beginning to see the forest and a few trees. *Neuron*, *77*(2), 209–211. http://dx.doi.org/10.1016/j.neuron.2013.01.010.

Wall, D. P., Kosmicki, J., Deluca, T. F., Harstad, E., & Fusaro, V. A. (2012). Use of machine learning to shorten observation-based screening and diagnosis of autism. *Translational Psychiatry*, *2*, e100. http://dx.doi.org/10.1038/tp.2012.10.

Williams, G., King, J., Cunningham, M., Stephan, M., Kerr, B., & Hersh, J. H. (2001). Fetal valproate syndrome and autism: Additional evidence of an association. *Developmental Medicine and Child Neurology*, *43*(03), 202–206. http://dx.doi.org/10.1017/S001216220100038X.

Zhang, P., Lewinger, J. P., Conti, D., Morrison, J. L., & Gauderman, W. J. (2016). Detecting gene-Environment interactions for a quantitative trait in a genome-Wide association study. *Genetic Epidemiology* . http://doi.org/10.1002/gepi.21977.