Neuropsychiatric Disease Working Group Plan

June 14, 2016

NHGRI Centers for Common Disease Genomics Program

Disease/Phenotype

As its inaugural CCDG project, the Neuropsychiatric Working Group (NWG) proposes to explore two diseases for which there is significant synergy: Autism Spectrum Disorders (ASD) and Epilepsy. ASD and epilepsy show substantial comorbidity (epilepsy rates in people with autism are as high as 20-40%) and the earliest exome sequencing efforts in ASD and epilepsy show significant overlap in genes with *de novo* mutations identified from trio or quad sequencing in each disease, suggesting that studying each disorder in parallel will enrich the other.

The NWG also discussed sequencing Alzheimer's Disease cohorts, given the clinical importance of the disorder, large numbers of available familial and case-control cohorts, and minimal amount of whole genome sequencing efforts completed to date. Due to funding constraints, pursuing this cohort would entail securing additional co-funding.

Autism: The disease criterion is diagnosis by ADI-R and/or ADOS. ASDs are serious neuropsychiatric conditions characterized by impairment in social and communication behavior, as well as restricted and repetitive interests. Symptom onset occurs in early childhood, is generally lifelong but highly variable with diagnosed individuals ranging from severely disabled to professionally successful. ASDs are highly heritable (50-80%) and now have a prevalence of over 1% with a 4:1 male:female ratio. Some genetic syndromes create strong risk for ASD (e.g. Tuberous Sclerosis, Fragile X) but identifiable genetic syndromes are responsible for only a small minority of cases. Commonly used treatments are predominantly behavioral, often targeting language development and behavioral symptoms, and have limited efficacy. There are no drug treatments thought to improve the core symptoms of ASD.

As of the end of 2014, approximately 400 genes had been found to contribute to the autism spectrum of disorders (Iossifov, et al, 2014, Nature) through analysis of CNVs and coding sequence variants of specific genes, and these appear to converge on a subset of pathways. Nonetheless, current genomic strategies leave the vast majority (70%) of heritability in ASD unaccounted for. Through WES of the Simons Simplex Collection, we have recently discovered, for example, a transmission disequilibrium of private gene-disruptive events enriched in probands (Odds Ratio (OR)=1.14, p=0.0002) when compared to unaffected siblings (Krumm et al., Nat. Genet, 2015). The fact this effect was observed in simplex families where variants were transmitted exclusively from mother to sons provides one path for understanding the role of inherited variation with respect to this disease.

In performing WGS on 40 of the SSCs 2500 families, we identified about 30% more de novo coding SNV and small indels in the WGS than were identified in the WES data for the same samples. In addition, we used a combination of approaches based on counts of reads, split-reads, and improper mate-pair mapping to identify rare and de novo intermediate (10bp-1Kb) and large scale CNVs (>1Kb) and genomic re-arrangements such as inversions and translocations. We found two likely pathogenic de novo CNVs in the 40 probands, a 5Kb deletion affecting one the exons of the CANX genes and a 51kb duplication affecting 5 exons of the SAE1 gene, both of which were not identified with the WES data. We also discovered a duplication of a distal regulatory region of the autism-risk gene TRIO; this private variant transmitted from mother to son is a high-impact candidate for disease and was also missed in the WES study (Turner, et al, 2016, AJHG). While many genes exert influence on risk via *de novo* mutation, only a minority have been identified with certainty.

Continued trio-based exome sequencing in parallel with WGS will drive genes from candidates to certain contributors, provide a broader foundation for integration with WGS CNV data as demonstrated in (Turner et al 2016), and will enable the first definitive genes to be identified in which mutations have a more modest impact and which therefore make their contribution solely/largely through inherited rather than *de novo* mutation.

Studies of trios and quads in autism, in which unaffected siblings provide genetically ideal control populations, have led very quickly to the discovery of at least 250 candidate genes in which likely gene disruptive (LGD) mutations are likely to be of large effect, driving the shift in strategy to a complete understanding genetic architecture (Iossifov et al. Nature, 2014, ASC, Nature, 2014). Genome sequence, in principle, allows the contribution of non-coding and coding variants to be compared. We hypothesize that the discovery of genetic lesions of large effect can be used to identify classes of mutation and genes of smaller effect in the context of families. In parallel, significant scaling up of existing efforts in GWAS and exome sequencing are required to garner power for the larger proportion (but weaker individual element), inherited component of the genetic architecture. CNV analysis and whole exome sequencing (WES) of autism trios and quads suggests that ~30% of incidence of simplex autism can be explained by disruptive mutations of coding regions caused by deletion, duplication, nonsense and splice site mutations, small indels causing frameshifts (collectively called 'LGD's, for Likely Gene Disrupting), and missense mutation. Detection of these events has revealed many target genes, with disruptive mutations of strong effect; analysis of WGS will also allow discovery of non-coding variants that contribute to disease, with likely a spectrum of effects.

Based on the pilot described above, we expect that 1) we can detect more coding variants in the WGS than were detected from the published WES data-set on Simons Simplex Collection (SSC), 2) we can accurately detect non-coding variants, 3) read count based methods can easily detect CNVs down to 1kb resolution including variants that affect single exons, 4) the larger size (150bp) of the reads together with split-read methods enable the detection of intermediate size indels, and 5) methods based on split-reads and improper pairs enable the detection of the inversions and translocations. We will put special emphasis on detecting structural variants (SV), such as insertions, inversions, duplications and deletions, because these are more likely than point mutations to be of large effect, including those that alter regulatory regions of genes. Exome sequencing efforts in autism will also be expanded, enabling the efficient conversion of some of the many 'likely-associated' autism genes to be converted into certain, durable genes and will take us more swiftly to the sample sizes required to implicate new genes based largely or solely on inherited rather than *de novo* variation. Taken together, we expect to expand on our current understanding of Autism and classify disease-causing mutations by pathway.

Epilepsy: Epilepsy is a group of heterogeneous disorders characterized by the presence of recurrent seizures, driven by the excess firing in a variety of neuronal networks. The epilepsies are quite diverse in presentation, with variation in the age of onset, brain region affected, pattern of epileptic discharge, frequency and severity of seizures. Electroencephalographic (EEG) and brain imaging criteria are often used to aid in clinical classification. Two main subtypes of epilepsy are recognized: focal and generalized epilepsy. Focal epilepsy (previously known as partial) is characterized by seizures that are restricted to one hemisphere of the brain (~60% of cases), while generalized epilepsy is characterized by seizures that affect both hemispheres (~40% of cases). The motivation for identifying the genetic basis of epilepsy includes: (i) enabling the classification of epilepsy into more homogeneous subtypes, facilitating both physiological studies and clinical trials, (ii) clarifying the nature of the overlap between epilepsy and other neuropsychiatric diseases, particularly ASD, (iii) revealing underlying cellular mechanisms and (iv) suggesting approaches to mechanism-based therapies addressing the core issues of epileptogenesis. According to the CDC, about 1.8% of adult Americans have epilepsy.

Both common and rare variants have been identified as significantly contributing to epilepsy risk. Rare, highly penetrant mutations have been identified in Mendelian epilepsy families, identifying a series of "channelopathies," including mutations in sodium and potassium channels, which are hypothesized to lead

to hyperexcitability. However, these genes collectively explain only a *tiny fraction* of patients with epilepsy. CVAS of epilepsy (8,696 cases and 26,157 controls) identified two genome-wide loci for combined (generalized and focal) epilepsy and a third specific to generalized epilepsy. These findings are striking, because as no loci were identified for focal epilepsy despite analysis of twice as many cases. In epileptic encephalopathies, rare and particularly severe forms of epilepsy, recent sequencing studies have revealed novel risk genes and *de novo* SNV and indel coding mutations in 12% of cases.

Both common and rare variants have been identified as significantly contributing to epilepsy risk. In severe cases (e.g., epileptic encephalopathies), a much stronger contribution of *de novo* mutation is seen than in autism (as well as overlapping genes between the two) so the similar use of exome sequencing as has been done in autism should be very effective in articulating this component of the genetic architecture. Large-scale GWAS has successfully identified several loci for generalized and combined generalized and focal epilepsies and these more common forms will likely be amenable to similar strategies as used for other common diseases throughout the CCDG.

The early success in gene identification through exome sequencing in epilepsy, in particular, epileptic encephalopathies motivated us to dramatically expand the number of cases undergoing exome sequencing. These initial findings suggest that by increasing the sample size will yield novel gene findings, deepening our understanding of epilepsy. In particular, the wide phenotypic diversity of presentation in epilepsy presents an opportunity to learn about how such phenotypic variability relates to genetic risk.

Given that there is significant overlap between the two disorders (epilepsy rates in autism are as high as 20-40%), studying each disorder in parallel will enrich the other. The convergence of epilepsy and autism suggest the possibilities that there is overlap in genes, pathways, and a shared genetic etiology. Therefore there is significant interest in comparing the genetic contribution to pathophysiology and phenotype across diseases.

Design Overview

Autism: Based on the evidence above, we believe that WGS in the context of a family study design will yield the most rapid discoveries given the resources available. We will prioritize sequencing quads to leverage unaffected siblings as controls. WGS enables the identification for the first time of smaller CNVs that will extend our understanding of the genetic architecture of autism, in part by providing initial insights into the interpretation of non-coding genome. The WGS data will also contribute to the pool of CCDG and other data that can be used as a basis for development of tools to analyze noncoding sequence. In parallel, about 40% of samples will be studied with WES. Expanded exome efforts will be the basis for identifying additional definitive genes and provide a stronger base for integration with the novel components contributed by WGS as demonstrated in (Turner et al 2016). The expanded exome component will enable many genes currently in the intermediate 'good false-discovery rate' category to become unambiguously linked to autism by efficiently compiling additional mutations. In addition, genes in which mutations confer a more modest effect, and which therefore make their contribution primarily through inherited variation are currently not detected – power calculations (Zuk et al, 2014) suggest upwards of 20,000 cases are required to achieve power to find such genes and this component will move us more swiftly to that endpoint.

So far in the CCDG program, the familial study design is currently being pursued only in the autism project. Samples to be sequenced include:

> NYGC Year 1, 5,300 total - 2,300 (NHGRI-funded, 500 quads and 100 trios) plus up to 2,000 (Simons Foundation co-funded, 500 quads); NYGC Year 2, 4,800 total - 3,300 (NHGRI-funded, 1,000 trios and 300 samples from a large pedigree) plus up to 1,500 (Simons Foundation co-funded, 367 quads); NYGC Year 3 – 4,000 (NHGRI-funded, 1,333 trios). All WGS.

Broad – Year 1 (3000 trios), Year 2 (3000 trios), all WES – further years evaluate switch to WGS based on costs and relative insights gained from assessment of NYGC & Broad results in Y1/Y2

If additional capacity were to become available, we could expedite complete sequencing of the autism collections for which we have samples in hand including SSC, AGRE, SAGE, CAG-Autism, TSAC, Autism CRC, Collaborative Autism, Mount Sinai Biobank and Homozygosity Mapping Consortium for Autism together totaling over 35,000. The collection of additional samples is ongoing including the Simons Foundation's SPARK initiative and Dan Geschwind's Autism Center of Excellence Network, which will enroll over 2,000 African-American subjects, composed of sibling-pairs. Independent of these, more than 9,000 cases (most with parents) are already collected by TASC, ASC, USCF and Miami Children's collaborators.

NYGC will pursue 30x WGS for consistency with samples from the same cohorts that have been previously sequenced. WGS at Broad is conducted PCR-free at targeted depth of ~22x to insure optimal sensitivity:cost ratio for singleton and de novo variation (though no Broad WGS is proposed here for the present). WES target depth is targeted at the same sensitivity – consistently evaluated and adjusted by attenuation of number of lanes used to run barcoded pooled DNA samples.

For the autism data produced at NYGC, we will run NYGC's standard pipeline for SNVs, indels, SVs and CNVs along with familial analysis, mitochondrial analysis, additional structural variant analysis, joint genotyping, and analyses in development to prioritize and interpret variants. The methods in development will pull from the strengths of previous algorithms developed by our team and collaborators including fitCons, RVIS, fGWAS, dCGH and lobSTR. By developing methods to identify and prioritize non-coding variants, this study will enable discovery of missing heritability in autism. These methods will be shared throughout the program to enable discoveries in other common and Mendelian diseases. Integration of all autism and epilepsy exomes sequenced to date with CCDG-generated data will enable a comprehensive picture of sharing between the disorders and clarify when the plateau of gene discovery suggests the trio exome approach has been exhausted.

The Broad Institute standard pipeline of Picard (implementing BWA and processing tools originally developed in GATK), generation of gVCF and joint calling of all previous autism and epilepsy exome samples will be performed in order to generate an exome data set suitable for all individual and cross-disease analyses. De novo mutations and constraint/intolerance analyses will be performed by tools developed by Goldstein and Daly labs and QC and statistical analyses performed within the HAIL framework developed by the Neale group.

The GSP Data Working Group has and will continue to harmonize processes between the centers.

Power will be assessed using two different approaches. The first being the burden of *de novo* variants within genes and the second using case-control analyses. Combining our data with existing WGS data from other collections we have the capability of assessing ~10,000 cases with autism. In addition, we can take advantage of other datasets as controls including 35,000 general research use (GRU) consented WGS samples from TOPMED (https://www.nhlbi.nih.gov/research/resources/nhlbi-precision-medicine-initiative/topmed) and data on 45,000 WES individuals from ExAC (http://exac.broadinstitute.org/).

Examining *de novo* mutations from 10,000 familial autism genomes, for example, we should be powered (80% detection) to reach genome-wide significance for genes with eight or more *de novo* mutations (assuming 1.8 *de novo* mutations per generation for ~20,000 RefSeq genes). This analysis is focused on coding events and will be useful for data generated by either WES or WGS. Importantly, WGS will afford us the opportunity to assess genes not previously captured and will provide insight into new etiological factors for autism.

By comparing cases and controls, we will be able to assess various types of variants, including both coding and noncoding, at a genic level. Considering WGS of available autism samples (10,000 cases and 5,000 controls), we estimate, based on 1,000 simulations, that we have 80% power to detect a proportion difference of variants of 0.0036 between cases and controls at the genic level (p-value cutoff < 2.5×10^{-6} for

n=20,000 genes). When utilizing 35,000 whole-genome sequenced TOPMed samples as additional controls, we will have 80% power at a proportion difference of 0.0011. Finally, if only considering WES regions we can also use data from ExAC (n=45,000) and get 80% power at a proportion difference of 8.00×10^{-4}.

Epilepsy: The significant scaling-up of efforts using exome sequencing is far and away the fastest route to gene discovery in epilepsy. A case/control design will be used (2-3:1 cases:controls). As the excess of rare protein-disruptive variants documented in limited studies to date is greater than in autism and comparable to developmental delay/intellectual disability studies, it is highly likely that the significant discovery trajectory in those other phenotypes will be matched in epilepsy as we move from hundreds to thousands of exomes examined. Moreover, the massive exome control samples currently available (>100,000 in upcoming release of ExAC) allow the straightforward recognition of genes under strong selective pressure without the need for *de novo* status to flag them as likely damaging – thus a design primarily focused on case sequencing can more efficiently get to these discoveries. Severe epilepsy constitutes the phenotype with the strongest contribution from rare, disruptive protein-coding mutations compared with the other current CCDG disease choices. In a not yet published case-control study using an analysis framework that collapses rare functional variants together in 356 epileptic encephalopathy cases compared to ~7000 ethnically-matched controls, genome-wide significant signals in three genes were identified. All three genes are accepted epileptic encephalopathy genes, *KCNT1*, *SCN2A*, and *STXBP1.* Assuming a similar frequency of disease-causing mutations in additional cases, adding 2000 cases would yield genome-wide significant signals for nine known epilepsy genes, and likely reveal significant associations of novel genes. These preliminary data strongly support the study design pursued here.

Samples to be sequence at Broad include Year 1 (7000 samples – roughly 2-3:1 case:control), same in year 2. Specifically, an additional ~1000 individuals with epileptic encephalopathies and ~5000 individuals with more common subtypes of epilepsy, including genetic generalized epilepsies and non-lesional focal epilepsies, will be sequenced each year. Recent analyses in these more common forms of epilepsy reveal that, with respect to genes so far implicated in risk , the majority of the risk single clearly arises from very rare variants (MAF<0.05%, Epi4K Consortium). Simulations reveal 80% power to detect a genome-wide significant signal of these very rare variants that explain the disease in 0.2% of cases in cohorts exceeding 5000 cases. When the return of WES begins to be diminished in expanded sample sizes, an indicator that the majority of epilepsy genes have been identified, a decision will be made to switch to WGS to identify regulatory risk alleles. If additional sequencing capacity were to become available, 37,000 epilepsy samples are available through the Epi25K consortium (led by David Goldstein, Sam Berkovic, Dan Lowenstein and Holger Lerche), encompassing a wide range of epilepsy phenotypes. Decisions about phenotypes to concentrate on will be driven by the knowledge of the genetic architecture of the epilepsies at the time.

The Broad Institute standard pipeline of Picard (implementing BWA and processing tools originally developed in GATK), generation of gVCF and joint calling of all previous autism and epilepsy exome samples will be performed in order to generate an exome data set suitable for all individual and cross-disease analyses. De novo mutations and constraint/intolerance analyses will be performed by tools developed by Goldstein and Daly labs and QC and statistical analyses performed within the HAIL framework developed by the Neale group. We recognize that WES will not detect all genetic variants in familial epilepsy cases. To move toward a comprehensive approach to understanding epilepsy, we will continue to attend to the possibility of re-analyzing cases by WGS, which is technically different from WES and therefore has different thresholds for sensitivity and specificity.

Description of samples to be sequenced

Autism: We expect that SSC, Autism Genetic Resource Exchange (AGRE), and the Homozygosity Mapping Consortium for Autism will be sequenced first. Initial priority for samples collected by The Autism Simplex Collection (ASC/ TASC) sources that are previously validated, have data use letters for dbGAP deposition in hand, and are available for sequencing now at Broad. SSC is our highest priority cohort and we are prioritizing individuals who have been consented for recontact. We will also prioritize the Homozygosity Mapping Consortium for Autism, of which we expect 300-400 will have sufficient material to be sequenced using a PCR-free protocol. We plan to sequence the rest of the AGRE. As soon as Dan Geschwind's collection of African-American sibling-pairs are available, we will prioritize them to include significant diversity.

We have prioritized familial collections and will prioritize collections containing underrepresented minorities as soon as possible; the only collection in progress that we are aware of is Dan Geschwind's work to collect African-American sibling-pairs.

We are pursuing the possibility of reconsenting individuals in these collections to allow the data to be used outside the current restriction of "autism and related diseases".

Epilepsy: Epi25K samples are currently being selected and shipped with same requirements as for autism samples. Many of the samples have been ideally consented, are recontactable (EpiGen), and for some of the collections the DNA is verified to be of high quality (and in all cases is collected from a blood sample can be sequenced PCR-free). The prioritization of samples from a phenotypic standpoint is decide by the Epi25K Consortium, in consultation with the CCDG at the Broad Institute.

We will similarly prioritize collections containing underrepresented minorities when possible.

Controls, particularly for samples from ancestries not well covered by available WES data, will be prioritized for sequencing if complete general use sharing is available and these will be contributed to shared control resources.

Collaborative efforts

Autism: The Simons Foundation has committed to funding sequencing for 3,500 whole genomes. We will explore genomic connections between cardiovascular disease and autism; with improvements in pediatric neonatal surgery and consequent increased survival among children with previously fatal congenital anomalies, we and others have recently recognized the possibility that such children have an increased incidence of autism. We will explore common features of genomic architecture, genotype and phenotype between the CCDG autism and CVD cohorts, as well as CMG CVD efforts.

Epilepsy: Other data sets that have the potential to synergize with the proposed epilepsy study are mentioned above.

Working Group Members

Bob Darnell (Chair), Carlos Bustamanate, Steve Buyske, Mark Daly, Evan Eichler, David Goldstein, Ira Hall, Ivan Iossifov, Adam Locke, Tara Matise, Ben Neale, Joe Pickrell, Aniko Sabo, Tychele Turner, Julia Moore Vogel, Mike Wigler, Mike Zody.

Overview of samples available for sequencing (note, more details are included in the CCDG Project Cohort Details table)

Disease	Number of Cases	Number of Controls	Ethnicity[1]	Data Sharing/ Consent[2]	Key metrics or phenotypes	Other details (Cohort name and other omics data)
Autism	2874	8467	C=72, A=4, H=11, O=13	2 dbGaP deposition and National Database for Autism Research (NDAR)	Autism diagnosis by ADI-R and ADOS; All simplex families	Simons Simplex Collection (SSC); WES and SNP microarray
Autism	3230	6105	C=69, A=8, H=11, O=12	2 dbGaP deposition and NDAR	Autism diagnosis by ADI-R and ADOS; All multiplex families	Autism Genetic Resource Exchange (AGRE); SNP microarray
Autism	448	634	C=86, A=3, H=1, O=10	2 dbGaP deposition and NDAR	Autism diagnosis by ADI-R and/or ADOS, Clinical reports, FSIQ, SRS; Familial cohort	Study of Autism Genetics Exploration (SAGE), High Functioning Autism; WES
Autism	2815	5407	C=54, A=23, H= 3, O=20	1 dbGaP and general release	ADI-R and/or ADOS, EMR/development evaluations/meds	CAG-Autism; SNP microarray, EMR and clinical evaluations
Autism	748	1648	C=70, A=3, H=0, O=27	2 dbGaP, NDAR, NIMH repository	Autism diagnosis by ADI-R and/or ADOS	The Autism Simplex Collection (TASC); SNP microarray
Autism	300 (n = 1200 by December 2017)	n = 1000 by December 2017	C = 85, O = 15	3 Ethical approval obtained for deposition of data on all open access platforms	Autism diagnosed clinically, and confirmed via extensive behavioural phenotyping (ADOS and 3Di)	Autism CRC; note: 100 families are immediately available.
Autism	139	220	C=98, O=2	1	Male: 82% , Age: 3-30 years (Mean = 14 years) , IQ Range: <70	Collaborative Autism Study; microarray

					= 27%; 70-79 = 17%; 80-89 =19%; 90-109=27%; 110-119=5%;120-129=4%; >130=1%	
Autism	22	n/a	C: 32%, A: 27%, H: 41%	1 dbGaP and general use	Longitudinal medical records since 2003 and extensive health, SEA and ancestry survey at enrollment	Mount Sinai Biobank; GWAS chip, exome chip, exome sequencing
Autism	263	535	C: 96%, A: 1%, EA&SA: 3%	2 – Samples can be used to study neurological disorders and conditions in the family.	ASD diagnosed by clinical evaluation, DSMIV-R, or ADOS;IQ available for many; Familial cohort	Homozygosity Mapping Consortium for Autism; WES, SNP chips

(% White / Caucasian [C], % African / African American [A], % Hispanic [H] , % East Asian [EA]; % South Asian [SA]; % other [O])

(1) dbGaP or similar deposition with General Research Use, (2) dbGaP or similar deposition with specific research use; (3) general release (i.e. SRA) (4) Other - with description

Disease	*Number of cases*	*Number of controls (either population control or unaffected family member)*	*Ethnicity (please list % White / Caucasian [C], % African / African American [A], % Hispanic [H] , % East Asian [EA]; % South Asian [SA]; % other [O])*	*Data sharing / consent (*Please choose from (1) dbGaP or similar deposition with General Research Use, (2) dbGaP or similar deposition with specific research use; (3) general release (i.e. SRA) (4) Other - with description*	*Key metrics for this cohort*	*Cohort*
Epilepsy	392		100% African / AA	Working with IRB, expect (1)		African Cohorts: SEEDS Program
Epilepsy	2300	1000	89% White / 2% East Asian, 2% South Asian, 6% Other	1 dbGaP or similar as supervised by NIH data advisory committee with opt out letter	Deep phenotyping, epilepsy diagnosed clinically	Australia: Melbourne; some GWAS, 800 WES

Epilepsy	456	0	91% White, 1% African / African American / 2% East Asian / 3% South Asian / 4% Other	1 dbGaP or similar as supervised by NIH data advisory committee	Epilepsy diagnosed clinically	Australia: Royal Melbourne; 495 GWAS, 192 RVAS
Epilepsy	500	0	80% White, 5% African / African American, 3% East Asian, 5% South Asian, 5% Hispanic, 2% Other	Many will require reconsent	Deep phenotyping, many years of longitudinal records (as many as 40 years)	Canada: Andermann
Epilepsy	321	308	Nearly 100% White / Caucasian	Working with IRB toward (1)		Canada: Andrade
Epilepsy	200	28	Nearly 100% White / Caucasian	1 dbGaP or similar as supervised by NIH data advisory committee	Epilepsy diagnosed clinically	Cyprus

Epilepsy	10,000	10,000	Nearly 100% White / Caucasian	Working with IRB toward (1)	Epilepsy diagnosed clinically	Denmark; GWAS ongoing, WES available on controls
Epilepsy	10,000		Nearly 100% White / Caucasian	1 dbGaP or similar as supervised by NIH data advisory committee	Epilepsy diagnosed clinically, extensive medical records and pharmacoresponse data	EpiPGX, EpiCURE, and beyond; ~4000 GWAS, ~300 RVAS
Epilepsy	460	0	Nearly 100% White / Caucasian	1 dbGaP or similar as supervised by NIH data advisory committee	Epilepsy diagnosed clinically	Finland: Kalviainen
Epilepsy	1,000	0	Nearly 100% White / Caucasian		Epilepsy diagnosed clinically	France: Lyon
Epilepsy	400	100	Nearly 100% White / Caucasian	1 dbGaP or similar as supervised by NIH data advisory committee	Epilepsy diagnosed clinically	France: Paris, Auvin

Epilepsy	900	0	Asian	1 dbGaP or similar as supervised by NIH data advisory committee		Hong Kong
Epilepsy	1,045	70	Nearly 100% White / Caucasian	Working with IRB, expect (1)		Ireland: Dublin 450 GWAS, some WES
Epilepsy	300	100	Nearly 100% White / Caucasian	1 dbGaP or similar as supervised by NIH data advisory committee	Epilepsy diagnosed clinically	Italy: Bologna
Epilepsy	300	0	Nearly 100% White / Caucasian	1 dbGaP or similar as supervised by NIH data advisory committee	Epilepsy diagnosed clinically	Italy: Catanzaro

Epilepsy	114	0	Nearly 100% White / Caucasian	1 dbGaP or similar as supervised by NIH data advisory committee	Epilepsy diagnosed clinically	Italy: Milan
Epilepsy	208	274	50% White / 50% East Asian	Working with IRB, expect (1)		Japan: Fukuoka
Epilepsy	80	0	East Asian			Japan: RIKEN Institute
Epilepsy	300	100	Nearly 100% White / Caucasian		Epilepsy diagnosed clinically	Lithuania
Epilepsy	300	0	Nearly 100% White / Caucasian	1 dbGaP or similar as supervised by NIH data advisory committee	Epilepsy diagnosed clinically	Macedonia
Epilepsy	350	0	South Asian	1 dbGaP or similar as supervised by NIH data advisory committee		Malaysia

Epilepsy	2,200	0	Nearly 100% White / Caucasian	Working with IRB, expect (1)		UK: Liverpool, Imperial
Epilepsy	300	0	79% White, 14% African American, 4% Other, 1% Hispanic, 1% East Asian	1 dbGaP or similar as supervised by NIH data advisory committee	Epilepsy diagnosed clinically	USA: Baylor
Epilepsy	100	0	Mixed		Epilepsy diagnosed clinically	USA: Boston
Epilepsy	1,220	40	Mixed	1 dbGaP or similar as supervised by NIH data advisory committee		USA: Columbia; 800 GWAS, 400 exome
Epilepsy	2,613	1580	75% White, 10% Hispanic, 6% African American, 2% East Asian, 7% Other	1 dbGaP or similar as supervised by NIH data advisory committee	Epilepsy diagnosed clinically, up to 10 years of medical records, digital EEG/MRI	USA: EPGP; 400 exome

Epilepsy	250	0	78% White, 13% African American, 4% East Asian, 5% Other	Working with IRB, expect (1)	Prospective clinically diagnosed epilepsy cohort with digital EEG/MR and all medical records from diagnosis	USA: Human Epilepsy Project
Epilepsy	500	300	Mixed	Working with IRB, expect (1)		USA: New York: Mount Sinai
Epilepsy	2,500	0	81% White, 14% African American, 2% Hispanic, 1% Other	Working with IRB, expect (1); IRB wants to see full protocol	Epilepsy diagnosed clinically, medical records for many	USA: Philadelphia 2300 GWAS
Epilepsy	365	70	Nearly 100% White / Caucasian			Wales

Autism	9585	15834	15% H, 7% EA, 2% AA, 75% C	93% (2) specific research use, 7% (1) general research use	Autism diagnosis by ADI-R, ADOS	ASC: GWAS, exome	21% 22%
Autism	2856	3252	75% C, 3% AA, 2% H, 20% O	(2) dbGAP with specific use	Autism diagnosis by ADI-R, ADOS	NIMH repository: all GWAS and/or exome chip, 15% exome seq	
Autism	438	1048	81% C, 4% A, 4% EA, 11% O	(2) dbGAP with specific use	Autism diagnosis by ADI-R, ADOS	Boston Autism Consortium: 175 with exome, all with GWAS/exome chip	
Autism	890	1780	22% C, 17% H, 61% EA	(2) dbGAP with specific use		UCSF & Korea	
Autism	1766	0	48% H, 11% A, 16% C, 25% O	(2) dbGAP with specific use		Miami Childrens	