NHGRI Centers for Common Disease Genomics

Autoimmune/Inflammatory Disease Working Group Plan

August 31, 2016

INTRODUCTION

The members of the CCDG Autoimmune/Inflammatory Disease Working Group have proposed projects in three major inflammatory and immune-mediated diseases: inflammatory bowel disease (IBD) to be primarily led at the Broad Institute (PI: Daly), type 1 diabetes (T1D) to be primarily led at Washington University, St. Louis (PI: Hall) and asthma to be primarily led at the New York Genome Center (PI: Darnell). These sub-projects were independently proposed but have been linked into a CCDG working group to focus on autoimmune/inflammatory diseases. From a genetic perspective this clearly makes sense because IBD and T1D share extensive disease associations from GWAS, including common and low frequency loss of function (LoF) variants as well as common regulatory variants. Given the considerable genetic overlap and considerable similarity of the proposed designs, the IBD & T1D studies have been combined into a joint project (described first in this document) while the distinct design of the asthma study is described separately. Ultimately the whole set of projects, all emphasizing whole genome sequencing (WGS) in minority population samples in the first years, will perform joint analyses at the appropriate time point.

Disease Phenotypes: IBD and T1D

Inflammatory bowel disease (IBD) is a chronic gastrointestinal inflammatory disorder that affects millions worldwide. It has two major forms (Crohn's disease and ulcerative colitis) that differ in course and therapy but which share more than 50% of associations documented to date. Therapies are of highly variable efficacy, often involve severe immunosuppression and/or lifetime monoclonal antibody treatment, and ultimately most patients require invasive GI surgery to remove affected areas of the GI tract. Once rare, IBD is on a precipitous rise in all parts of the world paralleling the Westernization of lifestyle – in some areas the prevalence of each form reaches 0.1-0.5%.

Type 1 diabetes (T1D) is an autoimmune disease in which nearly complete destruction of the pancreatic beta cells occurs, resulting in lifetime dependence on exogenous insulin for survival. Although T1D occurs most often in childhood (12-14yrs), there is growing evidence of up to 30%-40% of cases with onset in adulthood (~25yrs). Despite improvements in insulin therapies and glucose control, the morbidity and mortality from T1D remains elevated from the complications of T1D (nephropathy, retinopathy, neuropathy, cardiovascular and cerebrovascular complications). Although T1D is thought to be a "European" disease (with highest prevalence in Finland, Sweden and Sardinia), it occurs in

all populations. Further, between 1990-1999, the world-wide incidence of T1D has increased ~3%, while the age at onset has decreased.

IBD and T1D have enjoyed considerable success in the identification of loci via GWAS approaches – each with scores of loci defined indicating previously unsuspected genes and pathways involved in disease pathogenesis. In some cases, genes have been directly implicated in one or both diseases through lower-frequency stronger acting coding variants discovered in GWAS and follow-up targeted sequencing including: *NOD2, IFIH1, IL23R, CARD9, RNF186, SLC39A8, TYK2, PTPN22, SIRPG.* While these findings reinforce the importance of coding variation, they also suggest we have likely only scratched the surface of what can be discovered with lower-frequency variation not completely assessed via GWAS.

IBD and T1D have a well-documented genetic overlap, with numerous instances derived from GWAS, fine mapping (ImmunoChip) and exome sequencing where the same variant confers risk or protection from these diseases (e.g., *PTPN22*, *IFIH1*, *FUT2*, *IL2RA*, *PTPN2*, *TYK2*), suggesting combined analyses of these in particular will be a particularly important and productive activity.

IBD and T1D comprise a unique segment of genetic architecture that rounds out the NHGRI CCDG program. Like the neuropsychiatric (NP) diseases, they have relatively early onset and, in advance of major medical advances of the 20th century, would have similarly strong forces of direct reproductive selection acting against them. However, two important disease-specific aspects very strongly distinguish inflammatory from NP diseases.

Firstly, despite consistent estimates of high heritability, the incidence of both IBD and T1D is more than 10x increased over 2-3 generations ago in developed countries highlighting the very strong, not yet understood role of the environment.

Secondly, although direct selection seems to strongly cap the relationship between effect size and allele frequency in schizophrenia and autism, asthma, IBD and T1D have well-documented effects for common and low-frequency variants (despite the lethality of T1D before ~100 years ago) that eclipse those seen in both NP disease and late-onset cardiometabolic diseases that have no plausible selection acting against them. Taken together, these observations, in addition to the clear role in the immune system of the genes documented to date, raise the real possibility that genetic risk arises from variants that may have played protective roles against various pathogens during evolution. This is particularly appealing in the case of the association of *IFIH1* variants (an innate immune receptor sensing viral infection) with IBD and T1D.

The public health importance and potential for shared genetics with IBD and T1D suggests the two form a coherent autoimmune disease program.

Design Overview

A case-control design is clearly the most appropriate for IBD and T1D. Very substantial progress has been made with this approach on the heels of extensive family-based studies that by comparison only identified effects at a small handful of loci with larger effect risk factors. As exome and targeted sequencing have discovered many additional contributing variants through this design, it is clear we have not yet exhausted the very significant supply of biological insights that may be harvested from straightforward case-control studies.

Little to no WGS has been performed on IBD or T1D samples. Given that our cases are already assembled for the two diseases at different centers, it likely makes most sense for us to continue in this fashion, coordinating the order of populations, sharing controls (and more broadly sharing controls with the other working groups) to facilitate the most useful resource creation and value for the community.

Although we expect WGS to be the primary genomics platform for this project, it may be advantageous to boost sample size and study power by pursuing a limited amount of exome sequencing in European populations, where samples are plentiful and imputation reference panels are more mature. We therefore propose to devote a small portion of our budget (<20%) to exome sequencing of European samples. This activity will be considered lower priority than WGS of the African American, Hispanic and East Asian samples described below.

Given some reasonable assumptions about center funding, priorities and whole genome sequencing costs, we might estimate that, across the first two years, each center could dedicate 5,000 whole genomes to this subgroup – roughly a quarter of the total program sequencing capacity. This is far from a hard figure but something in this vicinity seems a reasonable estimate. Over two years, this may provide ~10,000 genomes fairly evenly contributed from the Broad and Wash U.

To maximize the potential for discovery, we propose to synchronize the sequencing of cases and controls by ancestry across each center in a 2:1 ratio, such that each disease will ultimately have 2:3 case:control ratio. Additional sharing of controls with the CVD group is expected to yield a 2:8-10 ratio, such that combined case analyses (*e.g.*, IBD+T1D vs. control) will still have a well-powered excess of controls from which to draw.

In the case of T1D and IBD, a number of low frequency variants (0.1% - 5%) have been documented with odds-ratios between 2 and 4 that are likely detectable with the proximal sample sizes available in this effort. All have thus far been protein-coding variants (genes include: *NOD2, IFIH1, IL23R, CARD9, RNF186, SLC39A8, TYK2, PTPN22, SIRPG*) and nearly all are detectable with exome sequencing and GWAS approaches. However, even in the unlikely case

that risk variants continue to be found exclusively in coding regions, non-European populations offer the most cost-effective opportunity for near-term progress because they have not yet been explored in sufficiently large sample sizes to reveal novel gene associations. In particular, we know that many of the key associated variants in genes listed above (e.g., *PTPN22*:R620W, *IFIH1*:I923V, *SLC39A8*:A391T, *NOD2*:R702W,fs1003) defined in populations of European ancestry are much rarer or absent in populations of other ancestries. Quite naturally, there are likely coding variants with similar impact that have yet to be discovered in non-European populations that should provide important additional clues to the biology of these diseases.

The power table below presumes a constant control sample of 10,000 is available and indicates that, for sample sizes at or exceeding 2,000-4,000 cases, this study will be well-positioned to define low-frequency variants with OR=3 in populations that have not yet been studied. We can also ask whether such variants are found in any population outside of protein-coding regions that have been well studied – addressing a critical outstanding question in human genetics.

	N cases vs 10,000 controls			
RAF	**1000**	**2000**	**4000**	**8000**
0.01	0.99	0.99	0.99	0.99
0.005	0.73	0.98	0.99	0.99
0.002	0.08	0.34	0.69	0.89
0.001	0.01	0.04	0.14	0.29

Power will clearly be maximized through a case-control design. The high odds-ratios for common and low-frequency variants compared to nearly every other common disease have established that the overwhelming majority of heritability arises from standing variation and the potential for large-scale control sharing with the CCDG cardiovascular disease (CVD) working group suggests that, in relatively short order, we can achieve the best powered studies of African-American and Hispanic/Latino immune-mediated disease performed to date across the allelic spectrum genome-wide. The diagnostic endpoints of IBD and T1D are clear with very few other conditions commonly mistaken for them, offering advantages in the reliable acquisition of additional case samples from biobanks and EMR systems. Further, their early age at onset and medical severity ensures that older control cohorts can be reliably screened for these diseases.

Description of samples to be sequenced

In collaboration with the T1DGC (led by Steve Rich) and the NIDDK IBD Genetics Consortium (led by Judy Cho), we have assembled ample samples to pursue the WGS study aims described above. Given overall synergy potential

with the rest of the project, and the fact that most studies have been on European-ancestry samples to date, we propose to emphasize understudied minority populations. In particular, we propose initiating work in year 1 on African-American and in year 2 Hispanic-Latino samples synchronously for both diseases. Sample consents are adequate for analysis and data release consistent with the requirements of this program.

The NHGRI has placed emphasis on diversity. We agree that this is a sensible priority, particularly for IBD and T1D with much lower rates of disease in non-European ancestry populations (African- and Hispanic Americans).

In addition to this proposal, the CVD working group has proposed that, long-term, they will seek to distribute samples fairly evenly across five ancestries – African, Hispanic, East Asian, South Asian and European (half Finnish representing a homogenous European ancestry, half non-Finnish). At this point, it is not clear that the Autoimmune Disease subgroups have adequate samples from South Asia (i.e., India and Pakistan) that are consented appropriately for broad data sharing.

Leveraging the CVD working group plan to maximize sharing of controls, we would propose that our long-term goal would be to study:

25% African-American
25% Hispanic
25% Other (including Asian and East Asian)
25% European (half Finnish, half non-Finnish)

While our long-term goal may be one of balance, in the near term we agreed that it makes most sense to begin resource building in minority samples, since these populations have historically been understudied and have less complete imputation reference panels. Thus, we propose to build a resource of 5,000+ African-Americans genomes in year 1, with a similar effort in year 2 that focuses primarily on Hispanic and East Asian samples (in that order of priority). As Finnish samples will be sequenced early during the CVD project and are ready to run for T1D and IBD, we also propose to initiate a smaller component of the European (non-Hispanic white) studies with these samples.

In a case/control design, current case availability is adequate for launching this sequencing effort in any population:

	IBD cases	T1D cases
East Asian	1863	1061
Afr-Am	2167	1464
Hispanic	1826	1362

White-FIN	1500	10505
White-NF	11286	27058

4Collaborative efforts

IBD and T1D have a well-documented genetic overlap, with numerous instances derived from GWAS, fine mapping (ImmunoChip) and exome sequencing where the same variant confers risk or protection from these diseases (e.g., *PTPN22*, *IFIH1*, *FUT2*, *IL2RA*, *PTPN2*, *TYK2*) suggesting combined analyses of these in particular will be a particularly important and productive activity. The parallel collaborative communities in these diseases, with a strong-track record of fruitful large-scale genetic studies, stand well-positioned to validate findings in well-documented clinical samples from throughout the world. At the Broad Institute, an ongoing exome sequencing effort has already completed more than 5,000 IBD cases independent of those that are being proposed in the first years here that can synergize with this study for many analyses. At the University of Virginia, an ongoing JDRF-supported initiative focused on GWAS/exomechip search for variants associated with nephropathy in T1D has completed genotyping 30,000 samples. Further, the University of Virginia is leading the NIDDK-supported TEDDY nested case-control WGS study of European, HLA high-risk (HLA DR 3/4) children, ~500 who progressed to T1D vs. ~600 (at the same initial HLA genotype and autoantibody status) who have not progressed. Control sharing, particularly with the similar diverse ancestry case-control studies proposed for contemporaneous study by the CVD working group, will render a much more powerful analysis for the entire NHGRI program.

1. Disease phenotype: Asthma

Asthma affects 5% of the world population and is the most common chronic disease in U.S. children. Its prevalence is highest among Puerto Ricans (18.4%), followed by African Americans (14.6%), European Americans (8.2%) and Mexicans (4.8%). Asthma is a complex, heterogeneous phenotype that is likely to display features of several underlying genetic architectures. Recent studies have shown that many of the variants discovered in European or European American populations do not translate to other racial/ethnic groups in the United States. In the U.S., the patient population for pediatric asthma/asthma-related phenotypes is largely made up of Latino, specifically Puerto Rican, and African American individuals. The cohort proposed is a US, pediatric population, focused on underrepresented minorities.

There are several environmental and clinical factors known to impact asthma susceptibility, and our previous work in Latino and African American asthma populations has shown that certain effects may be population-specific. For example, we found evidence that certain racial/ethnic groups are more susceptible than others to developing asthma following early-life exposure to air

pollution and tobacco smoke. In Latino children, we found that a 5 part-per-billion increase in average nitrogen dioxide (NO_2) during the first year of life was associated with a 17% increased risk for developing asthma later in life. Among African Americans, the risk increased to 43%, suggesting that African American children are more susceptible to pollution-associated asthma when compared with Latinos. We hypothesize that population-specific differences in asthma susceptibility is due, in part, to group-specific gene-environment interactions.

There are several projects currently underway at UCSF, USC, and the Henry Ford Health System (HFHS) that would be enhanced by WGS data. These projects cover a wide range of asthma-related interests, including but not limited to the classification of pediatric asthma sub-phenotypes, pharmacogenetics of drug response, functional genomics of variants in 3'UTRs, population genetics of admixed populations, RNA sequencing/expression Quantitative Trait Loci studies (eQTL), acute exposure transcriptional studies, and biomarkers of stress. In particular, two projects at HFHS that would benefit from WGS data focus on the discovery of genetic drivers of inhaled corticosteroid response in patients with asthma and the identification of genes differentially expressed according to asthma status. WGS data would strengthen efforts to explain racial/ethnic differences in asthma prevalence, morbidity and drug response.

Design overview

Large-scale WGS analyses in racially diverse populations are necessary to further asthma genetics research because of the high prevalence, complex etiology, and racial/ethnic disparities in susceptibility and mortality. The lack of variance in disease explained by previous GWAS, candidate gene investigations, and WES studies suggests that much of the genetic contribution to asthma is in non-coding regions. In addition, WGS captures both common and rare genetic variants in coding and non-coding regions. The current approach of imputing missing genotypes is particularly problematic in minority population groups because (1) linkage disequilibrium (LD) varies significantly across populations, (2) genomic data from representative ancestral populations are limited or missing, and (3) individuals may have considerable between-group admixture. By focusing on a few underrepresented minorities, this study will allow the identification of variants in and between these populations. For consistency with TOPMed, we expect to sequence at 30x.

Heritability estimates from twin studies range from 48 to 92%, however GWAS have uncovered a small number of loci with small to modest effect sizes that together represent a small proportion of heritability, suggesting that causal variation is rare or not captured by GWAS. We hypothesize that this "missing heritability" is due in part to rare variants. Our hypothesis is supported by previous evidence from small scale re-sequencing and admixture mapping studies that show rare variants being significantly associated with asthma susceptibility.

Based on 1,000 simulated association tests, we anticipate over 80% power to identify genome-wide significantly associated rare variants at frequencies of 0.2% or higher with a minimum odds ratio of 3, and 100% power for variants at frequencies 0.3% or higher.

Table 1. Summary of FEV1 % Predicted by Asthma Status in Proposed Study Cohort

Race/ Ethnicity	Source	Asthma Status	N	Baseline FEV_1† (IQR)
African American n = 7,977	SAGE	Case	1,298	96.9 (88.1, 106.9)
		Control	2,270	102.8 (94.9, 111.7)
	SAPPHIRE	Case	3,689	87.3 (75.2, 100.7)
		Control	720	96.8 (86.5, 107.4)
Mexican/ Other Latino n = 8,241	CHS	Case	916	99.4 (92.1, 106.3)
		Control	2,615	100.1 (93.2, 107.3)
	GALA I	Control	165	--
		Parent	600	--
		Proband	300	89.8 (77.9, 101.2)
	GALA II	Case	1,160	95.4 (87.7, 104.0)
		Control	1,223	98.5 (90.6, 107.5)
	HOLA	Case	597	90.5 (81.6, 101.9)
		Control	429	92.9 (85.8, 100.2)
	SAPPHIRE	Case	194	89.4 (79.0, 102.4)
		Control	42	96.9 (88.7, 110.1)
Puerto Rican n = 3,753	GALA I	Control	213	--
		Parent	814	--
		Proband	407	84.8 (75.3, 94.4)
	GALA II	Case	1,214	85.4 (75.7, 95.6)
		Control	1,105	96.3 (87.6, 105.7)
White n = 4,752	CHS	Case	768	98.6 (91.3, 105.7)
		Control	1,430	100.7 (94.0, 108.0)
	SAPPHIRE	Case	2,182	90.6 (80.1, 102.6)
		Control	372	98.0 (89.1, 107.8)
Study Population Total N= 24,723		**Case**	**12,725**	**--**
		Control	**10,584**	**--**
		Parent	**1,414**	**--**

†Percentage of predicted FEV^1; IQR = inter-quartile range

Our study population was created by collaborating across several independent asthma study populations: the Genetics of Asthma in Latino American's Study (GALA), the Genes-Environment & Admixture in Latino Asthmatics Study (GALA II), the Study of African Americans, Asthma, Genes and Environments (SAGE), the Honduran Latino Asthma Study (HOLA), the Children's Health Study (CHS), and the Study of Asthma Phenotypes and Pharmacogenomic Interactions by Race-Ethnicity (SAPPHIRE) (Table 1). The GALA, GALA II, SAGE, and HOLA studies were conceived, initiated, and are under the control of the Burchard Lab at UCSF. We have received letters of support from our collaborators at USC and HFHS, confirming that biological, phenotypic, and environmental exposure data will be made available to us immediately from the CHS and SAPPHIRE populations, respectively.

We will test for gene-environment interactions using a likelihood ratio test on nested models, and a novel two-step analytical method (EDGxE, Gauderman et al, Genet Epidemiol, 2013, PMID 23873611). First, we will apply an exhaustive genome-wide scan for gene-environment interactions (first with air pollution and then with tobacco smoke) using logistic regression, and use a likelihood ratio test to compare nested models with and without an interaction term. However, because this approach may have low statistical power, we will further apply a novel two-step method (EDGxE). The EDGxE approach provides substantially higher power than a standard exhaustive GxE scan, and generally provides greater power than other previously proposed 2-step approaches. We will also evaluate novel methods as they develop.

Description of samples to be sequenced

All asthma cases enrolled have physician-diagnosed asthma or are currently taking asthma medication. Assuming adequate funding is available, 2,000 asthma samples will be sequenced in year 1, 5,000 will be sequenced year 2,

and 10,000 will be sequenced in year 3. If additional funding were to become available, all of the nearly 25,000 samples currently available could be sequenced within the first three years at NYGC; another option to consider would be to do more WES to increase power, although this would be restricted only to coding regions.

Minority collections, with family cohorts and complete environmental will be prioritized, followed by minority collections with complete environmental data, followed by minority collections.

We will prioritize sequencing families to test whether the stratified method produces results in diseases other than Autism. Sequencing minorities has been prioritized given the overall goal of CCDG and that Asthma's prevalence is highest among Puerto Ricans (18.4%), followed by African Americans (14.6%), European Americans (8.2%) and Mexicans (4.8%).

The minority collections will be fruitful for use as common controls. Environmental exposures including drug response, socioeconomic status, air pollution and tobacco exposure have been well documented in a large portion of the cohorts proposed for sequencing. Additional data has been collected through detailed questionnaires including socioeconomic status and perceived discrimination.

Collaborative efforts
Asthma cases and controls for most cohorts can be shared as controls for other studies.

All asthma PI's involved in this project are also part of the NHLBI-sponsored EVE Asthma Consortium. Additionally, the UCSF team is part of the NHLBI's newly sponsored "Trans-Omics for Precision Medicine" (NHLBI TOPMed WGS). UCSF and the NYGC are currently collaborating to perform whole genome sequencing and analysis of 1,500 children with asthma and drug response. UCSF and the NYGC will contribute all data to dbGAP. We will collaborate with all TOPMed Consortia PI's and pool whole genome sequence data.

We are pursuing co-funding from NHLBI. The asthma project would synergize with the current TOPMed Initiative, along integration of independent groups undertaking whole genome sequencing of Asthma samples. We have developed significant data with the collaborators on our NHGRI proposal, looking at a stratified cohort of individuals showing extreme responders to bronchodilators (beta agonists) that point to important genetic contributors to Asthma, and to the synergistic value of the proposed co-funding mechanism.

The CCDG Autoimmune/Inflammatory Disease Working Group
Mark Daly (Chair); Steve Buyske (Rutgers University); Carlos Bustamante (Stanford University); Esteban Burchard (University of California, San Francisco);

Jill Harris (Harvard/Broad/MGH); Ira Hall (WashU); Judy Cho (Mount Sinai); Karyn Meltz-Steinberg (WashU); Natalie Makow (Rutgers University); Tara Matise (Rutgers University); Robert Darnell (NYGC); Manuel Rivas (Broad Institute); Stephen Rich (University of Virginia Health System)

Available samples for the CCDG IBD and T1D studies					
Disease	**Number of cases**	**Number of controls**	**Ethnicity[1]**	**Data sharing / consent [2]**	**Key metrics for this cohort**
IBD	1864	405	92% C, 2% A, 2% H, 2% E, 2% O	(2) dbGAP with specific use	all collections in this section have diagnosis of Crohn's or UC
IBD	7035	3845	12% A, 11% H, 7% O, 2% E, 68% C	(1) dbGAP GRU	
IBD	8497	1914	3% A, 20% E, 12% H, 63% C, 2% O	(2) dbGAP with specific use	
IBD	1100	500	100% A	(2) dbGAP with specific use	
IBD	6950	8200	29% E, 12% SA, 50% C, 8% O	(checking)	59% CD
IBD	2000	2000	100% C	(checking)	41% UC
IBD	2249	908	100% C	(checking)	
IBS	1500	7000	100% C	(2) dbGAP/similar	
IBD	824	2,463	confirming ethnic breakdown..	dbGaP and general use	Londitudinal medical records since 2003 and extensive health, SEA and ancestry survey at enrollment
T1D	23	77 (~500 unaffected sibs)	~90% C, ~10% EA	(2) dbGaP/autoimmunity and complications	multiple autoantibodies, serum and plasma samples availabble; clinical evaluation, evidence of other autoimmune diseases

T1D	5	2 (~600 unaffected sibs)	~100% C	(2) dbGaP/autoimmunity and complications	multiple autoantibodies, serum and plasma samples availabble; clinical evaluation, evidence of other autoimmune diseases
T1D	802	889 (~600 unaffected sibs)	~80% C, ~10% A, ~10% H	(2) dbGaP/autoimmunity and complications	multiple autoantibodies, serum and plasma samples availabble; clinical evaluation, evidence of other autoimmune diseases
T1D	0	0 (~75 unaffected sibs)	~100% C	(2) dbGaP/autoimmunity and complications	multiple autoantibodies, serum and plasma samples availabble; clinical evaluation, evidence of other autoimmune diseases
T1D	1789	1998	~100% C	(2) dbGaP/autoimmunity and complications	longitudinal data
T1D	252	0	~100% C	(2) dbGaP/autoimmunity and complications	longitudinal data
T1D	1192	0	~100% C	(2) dbGaP/autoimmunity and complications	longitudinal data
T1D	~1500	0	~100% C	(2) dbGaP/autoimmunity and complications	longitudinal data
T1D	1284	0	~100% C	(2) dbGaP/autoimmunity and complications	kidney complication data
T1D	0	5134	~100% C	(1) dbGaP with GRU	descriptive conditions
T1D	0	6240	~40% C, ~60% A	(1) dbGaP with GRU	descriptive conditions
T1D	0	500	~100% A	(2) dbGaP with specific use	othere autoimmune data

T1D	3974	0	~50% C, ~50% A	(2) dbGaP with specific use	other autoimmune data
T1D	0	1610	~50% C, ~50% A	(2) dbGaP with specific use	other disease phenotypes
T1D	0	434	~100% A	(2) dbGaP with specific use	other immune markers
T1D	677	737	~100% C	(2) dbGaP with specific use	other immune markers
T1D	8673	0	~100% C	(2) dbGaP with specific use	other autoimmune disease
T1D	0	5417	~100% C	EGA	other diseases,phenotypes
T1D	525	0	~100% C	(2) dbGaP with specific use	other diseases, phenotypes
T1D	0	3059	~100% C	EGA	limited
T1D	0	0	~100% C	EGA	other autoimmune disease
T1D	5535	0	100% C (Finland)	(2) dbGaP with specific use	diabetic complications
T1D	4	0	100% C (Finland)	(2) dbGaP with specific use	other autoimmune disease
T1D	385	1329	100% C (Finland)	(2) dbGaP with specific use	other autoimmune disease
T1D	249	0	100% C	(2) dbGaP with specific use	other autoimmune disease

1. (please list % White / Caucasian [C], % African / African American [A], % Hispanic [H] , % East Asian [EA]; % South Asian [SA]; % other [O])
2. *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

Overview of samples proposed for sequencing in the CCDG Asthma study (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 omic data available)**
Asthma	4,976	5,405	A: 30%, H: 70%	2 dbGap - Specific Disease - Lung Diseases	physician-diagnosed asthma	Genetics of Asthma in Latino Americans (GALA I), Genes-environments & Admixture in Latino Americans (GALA II), Study of African Americans, Asthma, Genes, & Environments (SAGE), Honduran Latino Asthma Study (HOLA); GWAS, methylation (Illumina 450K), RNAseq, cytokines, whole exome
Asthma	6,065	1,134	C: 35%, A: 61%, H: 4%	2	Physician-diagnosed asthma, longitudinal clinical outcomes for multiple conditions, medication exposure	Study of Asthma Phenotypes and Pharmacogenomic Interactions by Race-Ethnicity (SAPPHIRE); GWAS, RNA seq
Asthma	1,684	4,045	H: 62%, C: 38%	2	Longitudinal air pollution and respiratory health study	Children's Health Study (CHS)
Asthma	7,216	19,199	C: 21%, A: 29%, H: 48%, EA: 2%	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

1. (please list % White / Caucasian [C], % African / African American [A], % Hispanic [H] , % East Asian [EA]; % South Asian [SA]; % other [O])
2. *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