Centers for Common Disease Genomics
Cardiovascular Disease (CCDG CVD) Working Group Plan

V. 08/31/2016

Disease Phenotypes: Early-onset Coronary Artery Disease; Hemorrhagic Stroke

The Cardiovascular Disease working group of the Centers for Common Disease Genomics (CCDG) considered five diseases: early-onset coronary artery disease (EOCAD), stroke, atrial fibrillation, congestive heart failure and type 2 diabetes. There were several key criteria in selecting the diseases to be studied:

(i) a large number of high-quality cases (e.g., ~10,000 - 25,000 cases, with ideally many more available for confirmation studies). Diseases for which we do not have an adequate number of cases available were not prioritized;
(ii) diseases should be of major medical and public health importance; and
(iii) diseases should likely cover a range of potential genetic architectures.

There emerged consensus that EOCAD be considered as appropriate for an initial investment and hemorrhagic stroke as an early second focus. (*Note added: an atrial fibrillation (AFib) study was subsequently funded as a CCDG collaboration with NHLBI. See CCDG CVD AFib addendum.*)

CAD is the leading cause of death in the world and common variant association studies show that our understanding of the underlying molecular mechanisms is incomplete (e.g, >2/3 of the 63 common variants mapped for CAD do not directly relate to known risk factors). CAD is representative of a class of common diseases with onset at middle age. When CAD occurs early in life, there is a stronger degree of familial aggregation. In addition, there is empirical evidence that a burden of rare coding alleles in individual genes contribute to risk for CAD.

Stroke is the second leading cause of death worldwide and the fifth leading cause of death in the US. Stroke is composed of several subtypes including ischemic (larger artery, cardioembolic) and hemorrhagic (small vessel). Common variant association studies suggest that studying stroke subtypes can facilitate gene discovery. The working group decided to focus on documented hemorrhagic stroke because of an established precedent (e.g., *COL4A1*) and opportunity for novel discovery.

Disease definitions
Early-onset CAD. We focused the primary effort on the EOCAD phenotype because the condition is much more heritable when the disease occurs at younger ages. EOCAD is defined as myocardial infarction; coronary artery stenosis >70% in at least one coronary artery; and/or coronary revascularization (coronary angioplasty with or without stent placement or coronary artery bypass grafting) at an early age. When available samples are plentiful, we consider "early" to be men ≤ 50 years and women ≤ 60 years, but understand the need to reconsider (i.e., men ≤ 55 years and women ≤ 65 years) for under-studied or especially informative populations. This phenotype will have a strong component of underlying coronary artery atherosclerosis. These age cut-offs represent the ~ 5% extreme tail of the age-of-onset distribution in the United States.

Stroke. We focused on a specific stroke subtype - spontaneous hemorrhagic stroke as our second effort. Diagnosis of spontaneous intracranial hemorrhage, defined as a spontaneous, nontraumatic, abrupt onset of severe headache, altered level of consciousness, or focal neurological deficit that is associated with a focal collection of blood within the brain parenchyma seen on neuroimaging or at autopsy and is not attributable to hemorrhagic

conversion of a cerebral infarction. Ideally, all cases will have central adjudication and evaluation of their neuroimaging for exclusion of hemorrhagic conversion of an ischemic stroke, diagnosis of lobar and non-lobar location, and measurement of hemorrhage volume.

Comparison group

The comparison group need not be individuals who are carefully chosen to be disease-free or to have a very low probability of converting to disease in the near future; there is little loss of power and there are large gains in efficiency in using a set of near-random individuals from the same or similar population as the comparison group. Because the words “comparison group” are unfamiliar to some and cumbersome, we will often use the word “control” to mean the same thing. Theory dictates that for any single disease and fixed sample size, a 1 : 1 ratio of cases to controls is optimal in terms of power. However, opportunistic inclusion of additional controls increases power and helps control inflation of p-values often experienced in rare variant complex disease studies. Practical experience indicates that a ratio of 3 to 4 : 1 of controls to cases is ideal for controlling inflation of p-values. Across the NHGRI CCDG program, we have the opportunity to leverage the comparison groups for other diseases in our investigation of EOCAD and hemorrhagic stroke. To maximize efficiency, we propose to use a large sample of controls across multiple available studies. This shared comparison group should have the following characteristics:

(i) Sampled from the same or similar population as the case group(s) in order to avoid confounding from population substructure and environmental heterogeneity;
(ii) Additional deep phenotyping in the comparison group will facilitate analyses to identify genes underlying potential risk factor phenotypes, thus serving as an additional path linking genotype to disease.

Design Overview

The Funding Opportunity Announcement (FOA) for the CCDG program (RFA-HG-15-001) asked us to propose a definition and strategy for comprehensive identification of *rare genetic variants* associated with common disease. The CCDG CVD working group proposes:

Definition. Comprehensive identification of rare genetic variants associated with common disease means detection, across major population groups, of the vast majority (>90%) of genes (coding regions) and genetic elements (non-coding regions) harboring disease-associated rare variants with relevant levels of risk or protection (>2- to 20-fold) and relevant selection coefficients (s=0.1%-10% for risk and s=0.1%-1% for protection.[1] A reliable strategy (i) should work across diseases with diverse genetic architectures; and (ii) should be coupled to effective ways to characterize variants discovered with respect to penetrance, associated phenotypes and co-morbidities, and physiology.

Strategy. To achieve comprehensive identification in coding regions of human genes, a well-powered rare-variant association study (RVAS) should aim for at least 25,000 cases, together with a much larger number of controls, for initial gene discovery, with follow-up in replication samples.[2]

---

[1]These ranges cover recently discovered loci. Larger effects will resemble Mendelian disorders; smaller effects will not be of interest for rare variants. Stronger selection approaches lethality; weaker selection results in an allelic spectrum with mostly common variants.

[2] Zuk O, Schaffner SF, Samocha K, Do R, Hechter E, Kathiresan S, Daly MJ, Neale BM, Sunyaev SR, Lander ES. Searching for missing heritability: designing rare variant association studies. Proc Natl Acad Sci U S A. 2014;111(4):E455-64. doi: 10.1073/pnas.1322563111. PubMed PMID: 24443550; PubMed Central PMCID: PMC3910587.

Power considerations

Our proposal is shaped by the following power considerations for rare variants in coding and non-coding portions of the genome.[2]

Detection power in coding regions. If one performs sequencing in an initial discovery set of 25,000 cases (with a large number of random controls) and then follows up by sequencing hits in an extension set of another 25,000 additional cases, one has power to detect ~90% of the following two classes of loci[2]:

(i) risk alleles that increase risk by at least 10-fold provided that the strength of selection does not exceed 10%; 5-fold if $s < 3\%$; 3-fold if $s < 1\%$; and 2-fold if $s < 0.2\%$.

(ii) protective alleles that decrease risk by at least 10-fold if $s < 0.6\%$ and 3-fold if $s < 0.2\%$. With this sample size, there is very little power to detect protective alleles under strong selection ($s > 1\%$).

Detection power in non-coding regions. Rare non-coding variants will surely contribute to common disease. With samples of ~25,000, individual variant analysis will be feasible for rare variants down to frequencies of 0.1%. Below this frequency, aggregate variant analysis (AVA) will be necessary. With 25,000 cases, detection will likely to be feasible for large effect sizes. Successful RVAS in non-coding regions will thus require one or more of the following: (i) focusing on *individual* variants; (ii) identifying sets of non-coding regions (e.g., regulatory controls) for each gene that together contain hundreds of functional nucleotides that can be pinpointed with high specificity; (iii) analyzing non-coding regions with high *a priori* likelihood (e.g., associated with genes implicated by coding mutations or common variant association studies (CVAS)); (iv) searching for large copy-number variations (CNVs); and (v) decreasing costs to enable analysis of massive numbers of samples.

Overview of an analysis plan

This research plan has four main components.

(1) RVAS for coding regions. We will:

(i) test all variants individually for disease association, using well-established approaches and thresholds for individual variant analysis (IVA);

(ii) test various collections of variants for disease association, including loss-of-function (LoF) alleles alone and LoF together with certain missense alleles (defined by frequency thresholds and predicted deleterious effects). For missense variants, we will apply tests that allow optimization of frequency thresholds (such as variable threshold, weighting of variants, and mixed directions of effect (such as SKAT-O and C-alpha).

(iii) test for association with gene sets, with emphasis on genes in regions with CVAS hits and genes in relevant biological pathways enriched in CVAS loci.

(iv) apply a statistical threshold corresponding to genome-wide significance, but also identify association at more lenient thresholds that require follow-up in extension samples.

(v) where an informative disease-related quantitative trait is available (such as low-density lipoprotein cholesterol or triglyceride levels for EOCAD), perform association studies for the quantitative traits and use these results to prioritize genes and variants for further analysis.

(2) RVAS in non-coding regions: Individual variant analysis (IVA). We will test all non-coding variants in the frequency range 0.1% - 1% *individually* for disease association, using the standard IVA with a threshold for genome-wide significance of 5×10^{-8}. With 25,000 cases, there is 90% power to detect association for individual variants with frequency 0.1% for effect sizes of 2-fold.

(3) RVAS in non-coding regions: Aggregate variant analysis (AVA). The challenge for non-

coding regions comes from the very large number of low-frequency and rare variants. As for coding regions, we expect that non-coding variants of strong effect may be under strong purifying selection and may thus have very low frequency. To have reasonable power to detect association of such variants with disease, it will be necessary to aggregate non-coding variants. The challenge is that - unlike coding regions, in which exons can be grouped together to form a large target and LoF variants can be readily recognized - non-coding regions lack robust annotation and known functional elements tend to be short. We will test annotated enhancers and sliding windows across the genome, but power may be low unless variants increase disease risk by >20-fold.

To increase power, we need to identify *collections* of multiple non-coding elements related to each gene, containing many hundreds of bases pinpointed with high specificity - in effect, a non-coding equivalent of the exons of a gene. To do so, we will exploit rich sources of information about non-coding regions.

(4) RVAS across the genome: Structural variation. Structural variants (SVs) - including CNVs, deletions, and inversions - can play an important role in human disease, and they are likely to be some of the strongest effects in non-coding regions. We and others have developed algorithms that produce accurate SV inferences by simultaneously using three different forms of information in WGS data: read depth, split reads, and read-pair separation. We will genotype SVs and analyze them for association with disease.

Whole Genome sequencing (WGS) and whole exome sequencing (WES)

To achieve sample size of 25,000 cases, we propose a combination of WGS and WES. WGS offers advantages of detection of variation in non-coding sequence and structural variation whereas WES is more cost-effective to achieve a given sample size.

Sequencing capacity

We estimate that there will be capacity to sequence ~41,000 genomes as part of this CCDG CVD initiative during the first 36 months of the program**.** This total number is based on the following assumptions: (1) target sequencing depth for WGS at a minimum of 20X; (2) exome sequencing costs budgeted at ~33% of WGS costs; and (3) approximately 50% of the total sequencing budget for the three participating centers (Washington University, St. Louis; Baylor College of Medicine and The Broad Institute) will be devoted to the Cardiovascular Disease working group.

Based on the above considerations, the working group's core strategy is to sequence and analyze (i) a large number of individuals with a particular disease or condition (i.e. cases) and compare them to (ii) an even larger number of individuals in a deeply phenotyped "comparison" group.

Description of samples to be sequenced

We propose sequencing 55,000 cases and controls. This represents 34,000 WGS and 21,000 WES (41,000 WGS equivalents). 83% of total capacity is devoted to WGS. By disease, the breakdown is 25,500 EOCAD cases; 7,500 hemorrhagic stroke cases; and 22,000 controls.

The cases are drawn from a variety of sources including three major consortia: Myocardial Infarction Genetics Consortium (MIGen), CARDIoGRAMPlusC4D, and the International Stroke Genetics Consortium (ISGC). Cases are derived from a combination of the following sources: (i) population-based cohorts; (ii) hospital wards (including cardiac catheterization laboratories) and clinics; and (iii) hospital-based biobanks linked to an electronic health record (HER). We have selected cohorts where participants have been richly phenotyped for measures beyond case-control status, and hospital biobanks where rich phenotypes and EHR phenotypes are available.

Specific sample targets by sequencing center are detailed below:

Wash. U.

2,500 African American EOCAD cases (WGS)
2,500 African American controls (WGS)
3,000 Finnish EOCAD cases (WGS)
5,000 Finnish controls (WGS)
1,500 Costa Rican EOCAD cases (WGS)
1,500 Costa Rican controls (WGS)
Sub-total: 16,000 genomes, 0 exomes

Broad:

2,500 South Asian EOCAD cases (WGS)
2,500 South Asian controls (WGS)
2,000 Multi-ethnic EOCAD cases (WGS)
500 East Asian EOCAD cases (WGS)
500 East Asian controls (WGS)
11,000 European ancestry EOCAD cases (WES)
4,000 European ancestry controls (WES)
Sub-total: 8,000 genomes, 15,000 exomes

Baylor:

1,000 Hispanic American EOCAD cases (WGS)
4,000 Hispanic American deeply-phenotyped cohort members (WGS)
2,000 African American deeply-phenotyped cohort members (WGS)
2,000 African American hemorrhagic stroke cases (WGS)
1,000 Hispanic American hemorrhagic stroke cases (WGS)
1,500 European-descent EOCAD cases (WES)
4,500 European-descent hemorrhagic stroke cases (WES)
Sub-total: 10,000 genomes, 6,000 exomes

Overview of samples proposed for sequencing

Disease	**Number of Cases**	**Number of Controls**	**Ethnicity[1]**	**Data Sharing/Consent[2]**	**Key metrics or phenotypes**
EOCAD	25,500	22,000	See above	1	Disease status and cardiovascular risk factors
Hemorrhagic Stroke	7,500		See above	1	Disease status and cardiovascular risk factors

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

The sequencing plan presented above reflects current estimates of sequencing costs and sample availability. In the event that either of these factors change during the course of the project, it may be possible to include additional samples. Additional available samples will be incorporated according to the following priorities: (1) WGS of EOCAD cases and controls from historically under-represented populations such as African Americans or Hispanics; (2) WGS of EOCAD cases and controls from populations that offer unique genetic advantages for trait mapping, such as Finnish or South Asian; and (3) Exome sequencing of EOCAD and

hemorrhagic stroke cases and controls from European populations. In general, we will incorporate additional samples in a manner that is consistent with our current study design, and aim for a case-control ratio in the range of 1:1 to 1:2 for any given population.

Working Group Membership

The Cardiovascular Disease working group is a collaborative effort across Common Disease Genomics centers at Broad Institute, Baylor College of Medicine, and Washington University in St. Louis. Working group members include Eric Boerwinkle, PhD (Co-Chair), Sekar Kathiresan, MD (Co-Chair), Nathan O. Stitziel, MD, PhD (Co-Chair) and Ira Hall, PhD. Representatives from collaborating studies will also participate in CCDG CVD Working Group activities.