Concept Clearance for RFA

The Whole Chip: Toward More Comprehensive Analysis of Genome-wide Association Data

NHGRI Advisory Council, May 2011

Purpose

The National Human Genome Research Institute (NHGRI) proposes an RFA to support broader utilization of genome-wide association (GWA) data in studies of human disease. The goals of this initiative are to enhance discovery and further understanding of the genetics of complex human diseases by: 1) facilitating more comprehensive analysis of existing GWA data, including currently under-utilized information such as X chromosome variants as well as Y, mitochondrial (MT), and copy number variant (CNV) data; 2) developing and validating new quality control and genotype calling procedures for these regions, as needed; and 3) developing and validating new statistical/bioinformatics methods, analytical strategies, and study designs for incorporating information from these regions into GWA analyses.

Background

GWA studies have identified over 1,200 associations with $p \leq 5 \times 10^{-8}$ for over 200 traits, yet only 7 such associations have been reported on the X chromosome (chr). This is largely due to exclusion of X chr variants (as well as Y, MT, and CNV data) from analyses even though these regions are assayed on many current microarray platforms. A recent review of the NHGRI Catalog of Published Genome-Wide Association Studies (www.genome.gov/gwastudies) showed that only 121 of 374 (32%) of GWA studies published from Jan 2010 through Mar 2011 reported analyzing the X chr in their Methods sections. Even fewer reported analyzing Y or MT data, although these chromosomal regions are typically more challenging to genotype and analyze. Nor does accounting for the number of genes on the X chr close the gap in GWA associations; with its 1,669 known genes, the X chr has only 7 reported associations at 5 distinct loci, while chr 7, with 1,880 known genes, has 48 reported associations at more than 20 distinct loci. This raises the question: why is the X chr so often excluded from analysis?

For example, in a pre-compute performed by NCBI's dbGaP staff for a GWA study of diabetic nephropathy, an X chr association was found with $p = 5 \times 10^{-11}$. This was the top association in the study pre-compute, yet the first step of quality control in the related paper was to exclude all X chr SNPs, without any stated rationale. Interestingly, *ACE2,* a candidate gene for diabetic nephropathy, is located in this same region of the X chr. Excluding these data from analysis would thus seem to limit investigation of the role of genetic variation in human disease. Mendelian conditions also provide ample evidence for the importance of the X chr in human diseases such as autoimmune, cognitive, and behavioral conditions. Similarly, MT variants are

of interest in aging, cancer, and oxidative stress, while the Y chr is the focus of much work on male infertility, and CNVs have been associated with autoimmune disorders, schizophrenia, and autism. Thus, it is not for lack of biologic relevance that these regions are excluded.

Removal of non-autosomal data is common in GWA QC procedures, but few reasons are given for it. An informal poll of leading GWA geneticists has reinforced the perception that X chr data are under-utilized but shows little consensus as to why. While some are optimistic that improved imputation methods including the X chr will lead to a natural uptake of X chr analysis in meta-analyses, many available datasets may not lend themselves to meta-analyses or re-analysis without stimulation. Despite availability of X chr imputation methods since 2008, and improved algorithms in 2010, little increase in uptake has been noted over the last 15 months.

Concerns were also expressed that current GWA arrays are still poorly designed for these regions, a problem felt unlikely to be overcome by technologic development and requiring sequence data for definitive analyses. Others felt that array quality was not the problem; rather, the special handling needed for sex-specific analyses and their reduced power represented a significant barrier. In addition, the first few GWA reports set expectations that analyzing the autosomes was sufficient for publishing findings. Overall, however, most felt the X chr deserved more attention despite being more difficult to analyze.

Challenges in analyzing and interpreting X chr data, combined with the plethora of findings obtainable from the autosomes alone, may therefore lead many investigators to under-utilize X chr data since important associations can often be found without it. A review of X chr literature and over 300 GWA studies suggests that the genotyping accuracy on the X chr is often lower than other chromosomes due to difficulties with clustering algorithms, higher levels of chr anomalies, and more missing data on the X chr. Using the Gene Environment Association Studies (GENEVA) consortium as an example, 12 out of 14 studies detected more chromosomal anomalies (roughly 14-fold), and 13 out of 14 studies detected more individuals with $\geq$ 5% missing call rate (roughly 4-fold) on the X chr as compared to the autosomes. In addition, random X-inactivation in women could potentially obscure important association signals, though up to 15% of X chr genes may escape X-inactivation. Genotyping of the pseudo-autosomal region shared with the Y chr can also be problematic. These analytic complexities may further reduce power for X chr analyses and make detecting associations even more difficult.

With such diversity of opinions and wide range of analytical issues, improvements in genotype calling accuracy and methods developed specifically for the X chr may facilitate improvements in power to enhance the detection of important associations. Moreover, SNP data from the X chr along with Y, MT, and structural variants already exist on many of today's GWA arrays. Though such data may not be perfect, stimulating the analysis of such existing under-utilized data could enhance discovery and further understanding of the genetics of human disease at modest additional cost. Comparison to targeted sequencing data may also reveal important information about improvements necessary to capture these under-utilized regions of the genome better in future analyses.

Research Scope and Objectives

This RFA would support 4-8 study investigators to obtain and analyze existing GWA data for phenotype associations with X chr variants, as well as Y, MT, and structural variants, as available. Investigators will be encouraged to focus on datasets where these data have not been analyzed, though some comparisons with prior analyses may be useful. They will also be encouraged to develop and validate new quality control and analytic methods as open-access software and disseminate them to the scientific community. Inclusion of diverse populations will also be encouraged, including but not limited to children, minority, and disparity populations.

Shortly after award, investigators will meet to share proposed analytic methods and needs and identify potential common analyses that might be undertaken. They will meet again halfway through the two-year grant period to assess progress and will hold a final meeting as a broader workshop to report on experiences and discuss lessons learned.

Applicants will be asked to present a plan for analyzing under-utilized GWA data (including but not limited to X chr variants, as well as Y, MT, and structural variants, such as CNVs) into human GWA studies within a 2 year timeframe. The focus of this plan should be on analysis of the X chr, though investigation of other areas including Y, MT, and structural variants may also be supported. Investigators must have access to existing human GWA data, either through their own sources or accessible sources such as dbGaP, as these awards will not support large-scale genotyping or sample collection. Genotyping and/or sequencing of DNA in a limited, high-priority subset of subjects could be supported as funding permits for well-justified confirmatory studies such as replicating GWA findings and/or testing the generalizability of analysis methods. Deposition of individual level data in dbGaP will be expected in keeping with NHGRI and NIH policies if the data used have not already been deposited. Simulation studies would be allowed as part of the proposal, but all proposals must include analysis of existing human GWA data. Selection criteria for funding would include studies that provide a broad range of diseases and traits (especially those of high public health importance), ethnically diverse populations, development and dissemination of user-friendly and open-access methods, and at least 2,000 participants with existing high-quality GWA data attempting to assay at least 550,000 variants including X chr variants, as well as Y and MT variants.

Mechanism of Support

This initiative would use the NIH R01 (Research Project Grant) award mechanism. Four to eight awards would be made.

Funds Available

NHGRI will commit roughly $2-3M over 2 years to support these awards.