Section 2. Specific Aims

The Wisconsin Longitudinal Study (WLS) has followed the life course of 10,317 Wisconsin high school graduates of 1957 and randomly selected siblings through repeated surveys (in 1964, 1975, 1993-94, 2004-06, and 2010-11) and links to such data resources as high school records and yearbooks and tax, disability, Social Security, Medicare, and death records. WLS has thus created a detailed record of educational, social, psychological, economic and mental and physical health characteristics in a relatively homogeneous population that is almost entirely of Northern and Western European ancestry, thus strengthening the power of WLS GWAS studies against population stratification. In 2007-8, the project began collecting DNA samples by mail and more recently samples have been collected in the course of home interviews that began in March 2010. (R01 AG09775, R01 AG033285). It is time to complement the store of WLS phenotypic data with a large array of genetic markers, which takes advantage of recent developments that have vastly increased opportunities for genetic studies of aging: behavior, cognition, personality, mental health, health, disease, and mortality**.**

Importantly, the grant (1R01AG041868-01A1) funding this genotyping, which received a priority score of 13, was solely designed to generate genetic data that would then be disseminated and shared widely. No specific research aims were proposed or funded, as the goal is to create a "public good" that will be used by many other researchers as well as the investigators.
Indeed, since 1983, the WLS has actively shared data with all researchers, with appropriate and careful attention to the privacy and confidentiality of research participants. We will continue this tradition with the addition of these genetic data. Researchers, via varying mechanisms, will be able to access *any of the existing WLS phenotypic data*, in addition to the genetic data. Our procedures to access these data are designed to reduce barriers to access and use, while still ensuring protections for our research participants. Our strategy is to provide a range of ways to access these data, which go beyond the current typical GWAS study in their design to reduce barriers and use, so that appropriate researchers can do what they want as expediently as possible.

Adding genetic data to the WLS would be especially advantageous because it can contribute to data consortia, which while common in medical research are new to the social and behavioral sciences (Benjamin et al. 2008). With GWAS data, the WLS could participate in these efforts because of WLS's rich set of phenotype measures. For example, WLS directors are planning participation in the Social Science Genetic Association Consortium (SSGAC), established by Daniel Benjamin, David Cesarini, and Phillip Koellinger (see letter of support). In addition to the SSGAC, the WLS can be used in a data consortium with the Health and Retirement Study (HRS). In 1992, the HRS began interviewing those born from 1931 to 1941 collecting rich data across domains of health, psychological characteristics, social networks, economic characteristics and behaviors. They have completed genome-wide assays on 16,000 participants in the study. The WLS cohort was born mostly in 1939, and though we have been interviewing our respondents since the late 1950s, starting in the 1990s our surveys are overlapping. The overlap is also in terms of measurement across nearly all domains of health, economics, and varying social and psychological measures. This comparability provides significant potential for GWAS studies based on a data consortium involving the WLS and HRS.

Ultimately, we expect high use of these data because: 1) We are well practiced in promoting and sharing our data. Hundreds of researchers presently utilize the WLS; 2) The WLS is similar to the Health and Retirement Study in terms of design and WLS affords possibilities for replication and augmenting power of HRS analyses in terms of a user-base. HRS released their GWAS data 6 months ago and already has 68 approved individual users listed on their website, in addition to participating in over 20 GWAS consortia; 3) Even though we currently have only a very limited (and outdated) pool of 78 SNPs, we have about 18 approved investigators outside the WLS team using these data; 4) Letters of support include consortia such as GSCAN, CHARGE, and ADGC and individual users including James Heckman, a Nobel prize

The specific aims are:

#1. To genotype WLS salivary DNA samples using the Illumina HumanOmniExpress BeadChip on approximately 9,000 samples.
#2. To create a WLS genotype database with state-of-the-art quality control, data protection, statistical procedures, and bioinformatics and to implement GWAS of WLS phenotypes as part of the quality control process and to demonstrate the possibilities, recommendations, and limitations of analysis of genotypic and phenotypic data from the WLS.
#3. To share raw genotype data and phenotypic data from the WLS with qualified investigators interested in solely utilizing the unique measures available in WLS, but also to have it utilized in the context of larger consortia of data, in a secure form consistent with NIH GWAS policies.

RESEARCH STRATEGY

3a. SIGNIFICANCE

We seek to integrate whole-genome scan data with the extensive phenotypic data collected over the life course of participants in the Wisconsin Longitudinal Study (WLS) who have been followed since 1957. **Importantly, the grant (1R01AG041868-01A1) funding this genotyping, which received a priority score of 13, was solely designed to generate genetic data that would then be disseminated and shared widely.** No specific research aims were proposed or funded, as the goal is to create a "public good" that will be used by many other researchers as well as the investigators. Adding genetic data to WLS would make it unique among nearly all social and biomedical studies, because of the breadth of its phenotypic data, as compared to studies like the Framingham Heart Study, and the length of the follow-up and sibling design, as compared to the few existing social science surveys with genetic data.

Compared to other large population-based longitudinal studies in the United States with broad phenotypic coverage, WLS is unique in its combination of sibling data, age homogeneity (with graduates born 1938-1940), and relative ethnic homogeneity (reflecting the population of Wisconsin high school graduates in the late 1950s, the sample is almost entirely of European ancestry), which strengthen the power of WLS GWAS studies against population stratification and other factors (Ott, Kamatani, and Lathrop 2011). WLS will be a unique, resource available to all qualified researchers that is well suited for genome-wide association studies (GWAS), replication of findings from existing studies in a population-based sibling-design sample, gene×gene interaction studies, and gene×environment interaction studies. The goal is to create a public good: a data resource that can facilitate cutting edge research in genetic studies of aging. Via varying mechanisms, researchers will have access to <u>all</u> of the genetic and phenotypic data WLS has to offer. Further, our data access procedures are designed to reduce barriers to access and use, while still ensuring protections for our research participants. Please see the Data Sharing section for a detailed proposal on how these data will be shared.

3b. INNOVATION

Adding genetic data to the WLS will allow for a range of potentially innovative analyses. While we are only funded to produce and share these genetic data, we do want to provide some specific examples of possible analyses. In addition to more standard health measures (anthropomorphic, self reported measures and data drawn from complete Medicare administrative data for 7,254 of our participants), one important way that WLS can contribute is the wide range of behavioral phenotypes measured in the study—and specifically behavioral phenotypes tightly linked to mental and physical health. While there is extensive evidence of the heritability of behavioral traits (for a review see Bouchard 2004), there is still much to learn regarding the precise genetic variants linked to these traits. Thus, WLS can contribute to large data consortia studies (e.g. Speliotes et al. 2010; Benjamin et al. 2007). While these consortia are common in the biomedical sciences, they have only more recently emerged in the behavioral sciences. WLS directors are already participating in the Social Science Genetic Association Consortium (see letter of support). Their future plans that would involve the WLS include studies that focus on risk aversion and time preference, both of which are heritable and strongly predicts risky behaviors like smoking and drug use (Anokhin 2010; Cesarini et al. 2009; Cherkas et al. 2008; Zhong et al. 2009). We also have letters of support from numerous investigators and consortia, which focus on cognition, adiposity, anthropomorphic measures of functioning (e.g gait speed, lung capacity), nicotine use and alcohol consumption. Our cognitive functioning data, which has been collected since early adolescence to the respondents' early 70s, may be especially valuable in understanding the genetic roots of cognitive decline in later life.

Aside from larger consortia, the longitudinal nature of our data, combined with the extensive and varied phenotypic data, provide ripe opportunity for exploratory innovative analyses. We can contribute to a small—but growing—body of longitudinal GWAS analyses (Furlotte et al. 2012). Longitudinal data has been employed in two ways. First, some studies focus on whether they can identify how genetic markers that have been linked to phenotypes (e.g. obesity) influence phenotypic trajectories across ages/time (Belsky et al. 2012; Imboden et al. 2012; Mei et al. 2012). In the WLS, one could focus on life course (from age 17-72) trajectories (e.g. using growth curve models) for a range of phenotypic characteristics including obesity, mental health, personality, smoking, alcohol use, and cognition among others. Alternately, some genetic epidemiologists argue that in order to understand the complex nature of disease, and the potential for gene-gene and gene-environment interactions, one needs to construct more nuanced life course models—this would consequently require longitudinal data with extensive phenotypic data. Indeed, recent work has begun to explore models to evaluate multiple phenotypic measures over time to distinguish the genetic, environmental, and residual error contributions to the phenotype (Fan et al. 2012; Furlotte et al. 2012; Ko et al. 2013). For these endeavors it would be especially productive to pair our data with the Health and Retirement Study (HRS), which has

comparative phenotypic measures at comparative times/ages for the same cohort. In 1992, HRS began interviewing those born from 1931 to 1941 collecting rich data across domains of health, psychological characteristics, social networks, economic characteristics and behaviors. They have completed genome-wide assays on 16,000 participants in the study. The WLS cohort was born mostly in 1939, and though we have been interviewing our respondents since the late 1950s, starting in the 1990s our surveys are overlapping, both in timing and phenotype measurement across most domains. This comparability allows for GWAS studies based on a data consortium involving WLS and HRS. One specific example would be analyses focused on cognitive change in later life, for which the genetic investigation has been limited by the shorter observation period of many other datasets (Harris and Deary 2011). Both studies have identical measures on a similar population measured at similar years and similar ages.

Finally, the sibling samples in our data, combined with the extensive and longitudinal phenotypic data, also provides the potential for innovative gene×environment analyses. Obvious areas are traits with significant numbers of identified predicted genetic variants, such as obesity, that also have a clear behavioral component. WLS data can be used to evaluate whether genetic risk for obesity is moderated by psychological traits (IQ) or by social environments (SES), and whether this moderation changes as people age. The heritability of obesity is well-documented (h^2 between .4 and .7 in Allison et al. 1996), and hundreds of studies also document the link between SES, IQ, and obesity (e.g., Belsky et al. 2013; Hernandez and Blazer 2006), but a full model of reciprocal relationships between genetic risk, cognitive functioning, and social resources in determining obesity remains to be articulated. Such an analysis could further incorporate sibling data. Sibling models ensure that genetic variation is randomly distributed, in addition to accounting for early environmental experiences. As argued in Fletcher and Lehrer (2011), the inclusion of sibling fixed effects control for both unobserved genetic (e.g. population stratification) and environmental influences, which may bias estimation. Since siblings share roughly 50% of unique genetic variation, sibling fixed effects partially (i.e., 50%) control for unobserved genetic influences that may bias the gene-environment interaction by accounting for a latent gene-gene (GG) interaction. In addition, the use of sibling fixed effects allows the genetic variation in our sample to be considered quasi-exogenous, as differences in genotype of full biological siblings is the outcome of a "genetic lottery" (Fletcher and Lehrer 2011).

*Who Will Use These Data:*As we already noted, the funding for our project was specifically provided so that we could provide a public resource that could be shared widely. As discussed in section 2 above, we expect high use of these data for the following reasons: 1) We are well practiced in promoting and sharing our data, andjundreds of researchers presently are already utilizing WLS; 2) WLS is similar to HRS in terms of design and WLS affords possibilities for replication and augmenting power of HRS analyses. HRS released their GWAS data 6 months ago and already has 68 approved individual users listed on their website, in addition to participating in over 20 GWAS consortia; 3) We have about 18 approved investigators outside the WLS team using the very limited (and outdated) SNP data we have; 4) Letters of support include consortia such as GSCAN, CHARGE, and ADGC, as well as individual users, including Nobel Laureate James Heckman.

3c. APPROACH

Description of Traits

WLS collects a broad range of information about participants and is recognized for the exemplary quality of its survey measurement, as well as its unusually high response rates (Hauser 2005; National Research Council 2001). Given that we have been following participants since 1957 and asking them questions about nearly every dimension of their lives, there are around 29,000 variables so we can only provide details on a very small fraction of these measures. Broadly, the content of WLS surveys has changed to reflect the life course of participants: education motivated the initial data collection, familial and career outcomes was emphasis for respondents in midlife, and later rounds have given increased attention to respondent's health, cognitive status, personality, psychological and other dimensions of well-being. The measures that we include in the Data Dictionary and Data summary include measures of physical health (e.g. chronic conditions, anthropomorphic measures, alcohol use, smoking) mental health (e.g. depression, personality), and key social predictors of health (e.g. education, cognition, social participation). **For a more detailed listing of topics and variables see the Data Dictionary, Data Summary, and Appendix 1. The Data Dictionary and Summary also clarify for what years we have data on each phenotype listed. We do want to emphasize that the sample sizes reported there-in and in Table 1 below are conservative as we just finished the collection of saliva samples and they have not all been processed.**

While we are only funded to produce and share these genetic data, we do want to provide some specific examples of possible analyses with power analyses for just a few of the nearly 29,000 variables. As already noted, one of the unique aspects to what WLS offers is the addition of behavioral phenotypes. From

twin studies, there is considerable evidence of heritability for behavioral phenotypes such as: subjective well-being (h^2≈0.4-0.5, Hamer 1996; Stubbe et al. 2005); educational attainment (h^2=0.4 in Branigan, McCallum, and Freese 2013); general measures of cognitive ability (h^2>0.5, Finkel et al. 1995); and "vigor" in old age, a composite which includes several of the anthropometric measures in WLS (grip strength, gait speed) (h^2=0.33 in Newman et al 2001). All these phenotypes are either aspects of health ("vigor", well-being) or important psychological or social predictors of health (education, cognition). We have 80% power to detect genetic variants that account for as little as 0.53% of the phenotypic variance in WLS alone, and as little as 0.28% when combined with the Health and Retirement Study (HRS) for many traits discussed above (see Table 1). The HRS is a similar age cohort and many survey instruments are overlapping. The exception to this is that we have an early childhood measure of IQ because unlike HRS our sample has been followed since late adolescence. We also want to emphasize that the WLS specific power analyses are lower bound thresholds because we are still collecting saliva samples and we also will have the opportunity to impute phenotypic data—in large part because we have such an extensive array of data on each individual participant.

Variable Name	Variable Description and/or Code	Number of WLS Subjects with Data	WLS			WLS+HRS meta-analysis	
			Mean ± SD	Minimum Detectable R^2_g at 80% power	Minimum Detectable β_g at 80% power	Minimum Detectable R^2_g at 80% power	Minimum Detectable β_g at 80% power
Measures of Subjective Well-being							
auto	Autonomy	8627	22.74 ± 3.84	0.53%	0.43	0.29%	0.32
envrmst	Environmental Mastery	8628	24.56 ± 3.94	0.53%	0.44	0.29%	0.32
pil	Purpose in Life	8620	27.81 ± 4.8	0.53%	0.54	0.29%	0.40
positv	Positive Relations with Others	8629	28.22 ± 5.64	0.53%	0.63	0.29%	0.47
pergrth	Personal Growth	8627	24.53 ± 3.92	0.53%	0.44	0.29%	0.33
Measures of Disability or "Vigor"							
gait	Gait Speet	7921	2.92 ± 5.26	0.58%	0.62	0.31%	0.45
grip	Grip Strength	8000	28.9 ± 10.76	0.57%	1.25	0.31%	0.92
HUI_sum	Health Utility Index Summary Score*	8692	0.79 ± 0.23	0.53%	0.03	N/A	N/A
peakflw	Peak Flow Measure	7919	366.99 ± 132.97	0.58%	15.63	0.31%	11.42
IQ and Educational Attainment							
edu	Years of Education	8940	13.77 ± 2.5	0.51%	0.28	0.28%	0.20
hs_iq	High School IQ	8532	102.83 ± 15.05	0.53%	1.69	N/A	N/A

R2g=proportion of phenotypic variance explained by the genetic variant; βg=The amount of change (in measurement units) per copy of the risk allele at allele frequency=0.30; p-value threshold-Genomewide significance (5

HRS-Addition of HRS into meta-analysis assumes the same phenotypic data completeness as WLS, and an HRS-specified 84% completeness of DNA collection in their 12,507 unrelated individuals

*This is a series of self-reported items that measure mobility, cognition, dexterity, emotion, hearing, and speech.

Study Population

WLS is based on a 1/3 sample of all 1957 Wisconsin high school graduates (N=10,317) and a sibling of these graduates (Sewell et al., 2004; Hauser, 2009). The graduate respondents were originally empaneled with an in-person questionnaire in 1957, which has subsequently been followed with data collection in 1964 (a mail survey of parents), 1975 (telephone survey), 1993 (telephone and mail surveys), and 2004 (telephone and mail surveys, as well as a spouse telephone survey), and 2011 (in-person). The paired sibling was randomly selected from a roster of all siblings unless graduate was a twin, in which case the twin was selected. For siblings, 2000 siblings were empaneled in 1977, and the full sibling sample was implemented in 1994. Once empaneled, a sibling survey has been fielded either subsequently or concurrently to the graduate survey in each round. In 2011, our response rate was 80 percent.

DNA collection. In 2006-7 WLS first collected saliva samples from respondents using Oragene kits and a mailback protocol patterned closely on a previous study (see Rylander-Rudqvist, et al. (2006)). We did an additional sample collection in 2011 during in-person interviews for those who did not submit samples in 2006-7. Oragene kits were selected because of their ability to be used in a mailback protocol (e.g., no need for immediate freezing) and their high average DNA yield relative to widely used alternatives for mailback protocols available at the time. (For protocol and consent forms, see Appendix 2.) We have looked at compliance to the DNA request in relation to variables that have previously been predictive of survey response in the WLS. Response differentials are very similar to those reported by Hauser (2005); the only apparent difference is more favorable response by males than would be expected (male cooperation rates were ~5 percentage points higher; females historically have had slightly higher WLS participation).

Justification of Services Requested The goal of completing a genome-wide scan of this magnitude is to provide a durable, accessible resource for a broad range of investigators working in a swiftly developing field. The Illumina HumanOmniExpress BeadChip provides an excellent platform for this purpose. It is comprised of ~730,000 markers, which include 392,197 markers within 10 kb of a known RefSeq gene, 15,062 non-synonymous SNPs, 19,935 markers mapping to the X and Y chromosomes. We have completed the quality control and analysis of data from this chip for a GWAS of PCOS also using the CIDR genotyping center. Given this, and that we have used similar Illumina genotyping platforms at both CIDR and the Broad Genotyping Center (the Infinium 610-Quad, 660W-Quad, 1M Duo, and Omni1-Quad for GWAS in GENEVA and eMERGE consortia), we can identify and correct any potential problems in the quality (see quality control below).

The Illumina OmniExpress captures 73% of the most common (>5% minor allele frequency) and 58% of the common (>1% minor allele frequency) variants in the 1000 Genomes Project European ancestry dataset. The

ethnic homogeneity of the WLS permits us to use a less dense chip to capture a significant proportion of the known variation in a population like ours, and therefore greatly reduce genotyping costs, without causing concerns about comparability for use of these data in meta-analysis consortia with projects like the Health and Retirement Study (HRS). This is further remedied since we will be imputing untyped genotypes using the 1000 Genomes reference panel (see imputation plan below) which allows us to directly compare data for SNPs even if they were not directly genotyped in both of the data sets to be compared and/or combined.

Data Analyses GWAS experience. Dr. Hayes has considerable experience in conducting genomewide association studies (Hayes, et al., 2007; Hayes, Forthcoming; Cornelis et al. 2010; Fu et al. 2010; Turner et al. 2011; McCarty et al. 2011; Yang et al. 2011; Denny et al. 2011; Zuvich et al. 2011; Crosslin et al. 2011; Kho et al. 2012; Rasumussen et al. 2012; Jeff et al. 2013; Crosslin et al. 2013; Urbanek et al. 2013). Dr. Hayes also served on the steering committees for the Electronic Medical Records and Genomics (eMERGE) NHGRI-funded U01 consortium, and the Gene-Environment Associations (GENEVA) NHGRI-funded U01 consortium, and well as on the genomics and analysis subcommittees for these consortia. The QC procedures used by the CIDR Genetics coordinating center at University of Washington were largely developed the genomics and analysis subcommittee for GENEVA in conjunction with the coordinating center. Through these studies we have gained extensive experience in the implementation, analysis and interpretation GWAS data; we have collaborative QC'd GWAS data with the coordinating center; and we look forward to doing so again with the WLS data.

Genomewide genotyping Quality Control assessments. Upon receipt of the genomic data, we will first perform quality control (QC) analysis on the data. We will follow the eMERGE and GENEVA QC pipelines which we helped to establish during our active participation in these consortia for genomewide genotyping data cleanup (Zuvich et al., 2011; Turner et al., 2011; Laurie et al., 2010). Weekly calls to discuss QC will include investigators from Northwestern University (NU) performing the QC, investigators from NU and Wisconsin-Madison familiar with the cohort and its data collection procedures, the genotyping center (CIDR), the Genetics Coordinating Center (UW), and NCBI which will accept the data for dbGaP.

i. General metrics. Initial analyses focus on a comprehensive assessment of data quality, eliminating only the most egregious markers considered unreliably (or too incompletely) genotyped. Based on our prior experience, we prefer to use quite lenient criteria to set QC thresholds; instead cataloging QC indicators (call rate, departures from Hardy-Weinberg equilibrium, minor allele frequency, etc) for each SNP and considering them during the interpretation of the data. Genotype reproducibility can be assessed by concordance between replicate samples and a comparison of the called and imputed (from the remaining genomic variation) genotype calls for each SNP.

ii. Plate/Batch effects (significant differences in allele frequencies between genotyping plates or batches) should be kept to a minimum, but we will test for statistically significant differences in the metrics described above between the plates nonetheless. With such a large number of primary and secondary phenotypes to be examined for genotype-phenotype associations, the best strategy is to randomize the plating of all participants. Related individuals will be genotyped together on the same plate, and we will balance the sex proportions across the plates, which should eliminate or reduce the possibility of spurious associations due to systematic differences in genotyping conditions between chips.

iii. Relationship and Gender checks. We will compare the pre-specified sex with that determined from the genotype data using mean X and Y chromosome intensities. We will use pair-wise IBD estimates (e.g. with PLINK (Purcell et al., 2007) to identify misspecified relationships, both missing and unanticipated. For the case-control GWAS, we will focus on identifying and removing undetected first-degree related individuals preferentially keeping cases rather than controls, and individuals with higher genotype call rates if the pair was a case-case or control-control. For the family-based GWAS, we will focus on correcting the relationships due to sample swaps, data record errors, nonpaternity, etc. Related samples also provide the opportunity to identify low quality genotypes by examining the high prevalence of Mendelian errors within the families.

iv. Chromosome anomalies. The mean X and Y chromosome intensities also permit the identification of sex chromosome anomalies. These will be compared with the B Allele frequency and LogR ratio to confirm the anomaly. The B allele frequency and LogR ratio will also be used to identify anomalies among the autosomes. Depending on the extent (entire sample, entire chromosome, partial chromosome), the sample with either be removed from the study or the anomalous region will be filtered (removed) from future analyses.

v. Assessing and accommodating population substructure. The WLS participants are almost exclusively of European ancestry (>99% white), so we expect limited genetic diversity. However, given that self-reported racial-ethnic identity is not always completely reflective of genetic identity, population substructure among the WLS participants remains a possibility. Even within a single ethnic cohort such as WLS, subtle population

substructure can occur which leads to spurious associations – for example the false-positives for lactase persistence within a European cohort (Price et al., 2006). The first step in identifying and correcting for substructure is to use principal components analysis (PCA) based approaches of smartpca in EIGENSTRAT (Price et al., 2006) to determine if there population outliers (e.g. >3 s.d. from population center) within WLS using African, European, and Asian HapMap populations to anchor the analysis. Should such extreme outliers exist, they will be removed from the analysis. Next, we will reperform this analysis after the outliers are excluded to model ancestry differences among study subjects along continuous axes of variation to correct for each SNP's variation in frequency across ancestral populations. We will use the first several eigenvectors as covariates in the association testing to adjust for differences in ancestry among the study participants.

vi. Imputation. Although the genome-wide SNP genotyping arrays directly interrogate only a small fraction of human genomic variation, given a reference database like HapMap or 1000 Genomes and the knowledge of the LD patterns in it, it is possible to use the SNPs genotyped on the array to conduct association tests in much of the remaining untyped variation in the human genome. This permits expanding of the phenotype-genotype association tests from 0.7 million markers to >4 million markers. It also permits analysts to directly compare WLS data to other datasets imputed to 1000 Genomes (which is becoming the norm) as well as to combine the data from several studies in the SSGA for joint meta-analyses. We will use IMPUTE2 (Howie et al. 2011) and the cosmopolitan 1000 Genomes v3 reference set from 14 worldwide populations to maximize ease to compare/join with these other datasets.

Tests for association. As part of the QC process, and also to describe the possibilities, recommendation, and limitations of analysis of genotypic and phenotypic data from the WLS, we will perform first-pass analyses (precomputes) of a minimal select number of traits for deposition into dbGaP. For potential qualitative GWAS (e.g. cognitive decline), we will use PLINK to conduct simple allele frequency comparisons for all SNPs generated for the GWAS using logistic regression so that we can include necessary covariates (e.g. first few PCs from EIGENSTRAT). For potential quantitative GWAS (e.g. normalized score on cognitive ability), we will use linear regression between the minor allele dosage and the trait of interest adjusting for confounding variables. For the family-based GWAS (e.g. studies that use sibling data), we will use the DFAM in PLINK for qualitative traits which implements the sib-TDT test and allows for the inclusion of unrelated individuals, or QFAM for quantitative traits which performs a linear regression test but permutes the data to correct for family structure.

Analyses will be conducted on the remaining 1000G SNPs successfully imputed. We will assess the performance of these corrections performed during QC by monitoring the resultant distribution of P-values against those expected under a null distribution (QQ plots) as well as calculating the genomic inflation factor (λ_{GC}) (Devlin et al., 2001). We will also visually inspect the genotype intensity boundaries to ensure quality of the genotype calls for the most significantly associated SNPs.

Meta-analysis. To combine the WLS results with the results from other GWAS (e.g. HRS, SSGA) of the same phenotype(s), these above analyses will be conducted separately in each study, each considered as equivalent strata and combined using standard meta-analysis approaches implemented in METAL (Willer et al. 2010). We will use fixed-effects models (since we are looking for deviation from the null of no effect overall) but will examine estimates of heterogeneity for SNPs showing overall association.

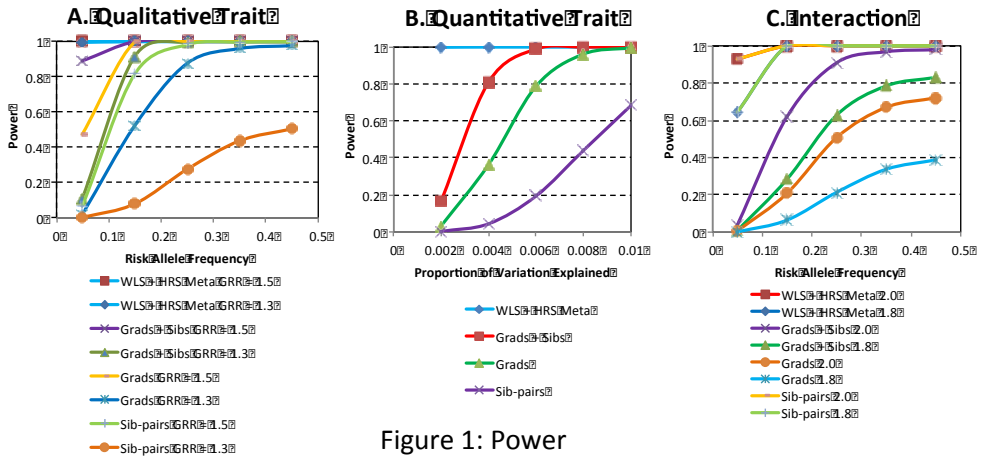

Figure 1: Power

Gene x Gene and Gene x Environment Interactions. We will conduct tests for two-level (SNP by SNP or SNP by environment) interactions. For the GxG tests, we will limit this to the top 10,000 SNPs following GWAS since correcting for *all possible* pairwise SNP interactions would require a nominal significance of $p<5x10^{-14}$ for which we would be very underpowered. In the power analysis below, we examine power at the $5x10^{-5}$ significance threshold which is sufficient for identifying top signals to carry to replication for a joint significance of $5x10^{-8}$. For qualitative traits we will also examine the GxG separately in each of the phenotypic classes (e.g. cases-only and controls-only).

Power and Effect Size We report power for a $P=1x10^{-8}$ genomewide significant threshold for the WLS GWAS as well as the WLS+HLS meta-analysis. Assuming a population prevalence of 0.10 for the trait of interest power is plotted against risk allele frequency for a series of genotype relative risks (GRR) under a log additive model for the ~9,000 grads+sibs (adjusting for relatedness), 6,200 grads-only, the ~2,800 sib-pairs, and the combined WLS + HRS adding another 12,500 (Figure 1A). The study design has good power for detecting susceptibility loci with GRRs in the range of 1.3-1.5 throughout much of the allele frequency spectrum. Figure 1B plots shows that for quantitative traits we have power to detect SNPs that explain as little as four-tenths percent of the total phenotypic variation. Figure 1C shows that we have reasonable power to detect interaction terms of 1.8 or greater assuming marginal effects 1.3 as suggested in Figure 1A.

Data management. We have managed genotype/phenotype data sets of similar size or larger. The GENEVA, eMERGE and PCOS GWAS included genomewide scans for >60,000 participants and for which we currently continue to manage and manipulate data for analyses. All genotype and phenotype data is deidentified and resides in a secure server behind the NU firewall, maintained by a dedicated staff trained in information security. All NU Feinberg School of Medicine (FSM) workforce members are trained to never send sensitive information (e.g. EPHI, user IDs, social security numbers, addresses, phone numbers, credit card numbers, or other personal data) across the Internet unless the information, or the communication itself, is encrypted. Encryption protocols must be approved by the Information Security Office. To ensure that only trained personnel access information for clinical, business or research operations the following mechanisms are in place: An information classification framework (ICF) is used to categorize data by its level of sensitivity; A role based authorization control (RBAC) scheme maps each user role to categories within the ICF; need to know access granted through an RBAC and ICF matrix and permissions are granted on the basis of least privilege. All personnel involved in this study are trained in the appropriate use and secure handling of phenotype and genotype information. Analytic results reports and meta-data created for research will be stored within the existing FSM servers where possible, or on password protected desktop computers, physically secured within a continuously guarded building. Laptops and other types of mobile devices present security challenges and FSM requires the owners of these devices to be responsible for downloading approved FSM encryption technology to protect the data contained on these resources in the event the device is stolen or lost. These devices are stored under lock and key when not required to be outside a physically secured building.

NU uses BC/SNPmax (Biocomputing Platforms Ltd.) software platform to maintain and manage large genotype and phenotype databases required for GWAS. This software consists of: an intranet database and application server with two CPUs, a mirrored disc system (RAID I) and a backup tape device, a RED HAT Enterprise Linux operating system, and IBM DB2 Universal Database relational database software. The database design is highly scalable and has ample capacity to handle the large amounts of data we will generate for our GWAS in WLS. All web browser connections between the system server and user workstations are SSL-encrypted, providing secure access to our genetic data. The Illumina data files can be easily directly uploaded into the database. At any time, users can archive data files to preserve key versions of the datasets, and to "freeze" datasets corresponding to set time points or to particular sets of analyses. It is possible to maintain different active versions of data sets to allow, for example, a comparison of genotype data resulting from the application of different allele-calling algorithms applied to the same intensity data. User access is safeguarded by both the BC/SNPmax system and by the high-security IBM DB2 Universal Database software that is an integral part of the system. Only NU analysis team members have access to the data storage/analysis system.

Plans for Next Use:

Letters of support indicate that numerous individual investigators and consortia including SSGA, GSCAN, CHARGE, and ADGC will use these data immediately. Further, the currently funded R01 devotes considerable resources to facilitate data sharing with a broad range of users. For extensive details on data sharing, which we believe go well beyond traditional GWAS studies, please see the Data Sharing Plan. Finally, this research team is submitting a P01 application to use these data for genomic-relatedness-matrix restricted maximum likelihood (GREML) analyses, which is a new approach to generating heritability estimates (Yang et al. 2011) GREML utilizes all our available genetic data (SNPs) to estimate the proportion of variance in a trait (in our case measures of mental health) that can be jointly explained by the genotyped SNPs. This technique has been employed applied to height, intelligence, personality, and schizophrenia (Yang et al. 2010; Visscher and Goddard 2010; Davies et al. 2011; Vinkhuyzen et al. 2012).