Responses to a Request for Information: Design and Implementation of a Large-Scale Prospective Cohort Study of Genetic and Environmental Influences on Common Disease (NOT-OD-04-041)

This request for information (released in May 2004) sought advice on approaches to developing a large-scale U.S. study of genetic and environmental influences on common diseases. Advice could include recommendations on optimal characteristics of such a study; recommendations on combining existing cohorts for such efforts; and characteristics of existing studies that might lend themselves to inclusion in such efforts. Respondents were asked to comment on one or more of the issues listed below. The responses received are compiled on pages 2-34 of this document.

INFORMATION REQUESTED:

1. If appropriate funding were available to support a large cohort study (N ~ 500,000) of genetic and environmental determinants of major complex diseases, please comment on what you would see as advantages of recruiting and examining a new cohort vs. building upon existing cohorts.

2. Please describe the characteristics of a large US cohort study that you view as most important to include in any such effort that might be undertaken.

3. Please suggest the family structures, and proportion of related individuals, that you would recommend for inclusion in the proposed study.

4. Please identify the most relevant issues concerning the power to detect genetic effects (such as environmental risk factors, heterogeneity, prevalence) for diseases or traits that you would be interested in studying in a large US cohort study.

5. Other information not specifically addressed by the comments above, but considered important and relevant to the development and implementation of a large-scale study of genetic and environmental influences on common diseases, would also be of considerable interest and value.

SUMMARY OF COHORT STUDY RFI RESPONSES, ITEMS 1-5

1. If appropriate funding were available to support a large cohort study (N ~ 500,000) of genetic and environmental determinants of major complex diseases, please comment on what you would see as advantages of recruiting and examining a new cohort vs. building upon existing cohorts (93 responses; number of respondents per item shown in column to left).

ADVANTAGES OF NEW COHORT (DISADVANTAGES OF EXISTING COHORTS)

15 **design**: selection of population and measures based on needs of study rather than convenience

- science needs to be defined carefully and ppts characterized as no others have been, necessitates new cohort
- random sample of US population would maximize generalizability
- best to plan large cohort study from the beginning; massive undertaking deserves careful, coherent plan
- new cohort enables one to pick age group, ascertainment criteria, and assessment package now desirable
- main advantage to starting new cohort is ability to collect same data, consistently and prospectively across sites; get it right from the start

13 **state of art**: measures in older cohorts may have used obsolete technology, not reflective of current health concerns, particularly for environmental exposures

- older cohorts often quite distant from exposures and recall bias could limit their use
- new cohorts needed for most environmental studies as existing large cohorts unlikely to have power and detail for gene-exposure interactions
- need objective measures of environmental exposure during critical windows of vulnerability; careful measures of exposures during fetal development and early childhood critical
- principal advantage would be standardized, sophisticated measurements of personal and environmental factors (diet, exercise, pollution exposure), physical characteristics, ancestry, and other potential effect modifiers for genetic traits

30 **consistent, standardized protocol**: uniform exposure information, endpoint ascertainment, biospecimen collection

` avoids confounding this study with parameters set for another study

` available data not uniform, might be forced to Alowest common denominator@ approach though might be ways to handle different well-validated instruments

` standardized approaches for data collection, especially environmental, behavioral and dietary data, would make gene-environment studies more feasible

` if environmental exposure assessment not standardized, eventually one will include so much misclassification that quality seriously endangered and risk estimates biased towards null

` new cohort could be systematically ascertained and assessed; major advantage

` broader range of phenotypic data can be collected

11 **poolability/survivorship**: difficult to combine information from questionnaires, assays etc. from studies not initially designed to be "pooled"

` review panels question re-use of data and combination of existing studies due to incompatibility

` one can to some extent "pool" studies for work on genetic risk factors alone but very difficult to pool data on environmental factors

` difficulty pooling due to ascertainment bias, differing sampling strategies, missing data on critical variables, survivor bias

` data already collected in individual studies would not be complete for all of the variables of interest

` will be very difficult to combine data from existing cohorts in valid ways

` given unintegrated state of knowledge on social/environmental factors, unimaginable that there is any kind of comparability across studies, so cobbling together of existing studies would not work and ongoing studies would need substantial re-vamping of their designs to participate

6 **consent**: more straightforward consent, broader and fully informed

` consent needs to be ever green so can do new things with samples and go back for additional information

` many samples obtained without explicit consent to research utilization; consent must be clear on purpose of study, likelihood of return of information, potential misuses; must be sensitive to cultural context

` existing cohorts would require reconsent; current procedures for informed consent could be implemented in new study without complexity of many changes occurring in consent procedures over time

8 **multiple outcomes**: rigorously designed to address multiple health outcomes from start, tailored to optimize questions answerable

` existing studies focus on narrow spectrum of confirmed disease endpoints and specific research goals of PIs who established them, almost always lack data crucial for new use or other conditions

` data often missing on co-morbid medical conditions; studies of mental health seldom obtain data on co-existing medical conditions and vice versa

` existing cohorts inadequate for health services research in critical illnesses requiring hospitalization and utilization of intensive care resources

10 **free and open access/resource**: for entire biomedical community, establish up front as shared and open to all

` open source model would be key, could redefine field, will force issue on privacy and genetic protection as well as HIPAA and IRB approval

` many collaborative efforts would fail if all outside investigators had access to all data as it was measured (perceived autonomy would be crushed; confidentiality would be breached; competition for control)

` need annual infusion of established outside investigators to learn and explore the database; consider NIH-sponsored sabbatical as price of admission to outside investigators, to learn database

` favor prospective approach for generation of large-scale cohort to reduce variability and enhance the wide and standard access to specimens

` new cohort can be built as open design from start where sense of ownership felt by particular investigator diminished

` consider separating functions that store and distribute genetic resources from those concerned with its analysis

` logistic/political factors can be overwhelming; Framingham viewed as grossly undershared resource, legacy costs in using existing cohorts can be substantial

` cooperation of existing administrative structures of various studies would be mixed with some being quite cooperative and others the opposite

` Governance of new study would be very complicated because of need for representation from leadership of all existing studies

` existing cohorts likely to have encumbrances such as commercial and licensing commitments

` assembling large cohorts expensive and not easily supportable with current NIH funding mechanisms; NIH should take lead and establish as central resource for use by individual investigators for independent or collaborative research; Human Genome Project and HapMap great models

` initial large cost of establishing cohort offset by multiple uses as central resource, prevalence data will be useful for studying genetic associations

9 **biologic specimens**: can obtain fresh, high quality specimens under standard protocols

` can establish new procedures for storing samples, as samples for proteomic or RNA analysis require different handling from procedures implemented solely for genomic analysis

` many exciting research questions based on new technologies in genomics and proteomics will not be able to be addressed within existing cohorts; sample availability from existing cohorts limited, RNA essentially absent

` standards for DNA collection and preservation would be uniform

` we do not have the most useful or identical specimens collected in all cohorts and poolability will be questionable from that standpoint

14 **diversity, representativeness**: new cohort could be more representative with minority and low income individuals as opposed to existing cohorts largely limited to high SES and Whites

` Multi-Ethnic Cohort important exception and could serve as model

` specifically selecting minorities and other underrepresented groups would insure that important research questions could be adequately addressed

` sampling unit could be based on census tracts to check for representativeness, also to allow GIS mapping and spatial analysis

` participants should represent nation as a whole, with oversampling for minority groups

` clinical samples are not nationally representative and their representativeness cannot be known

5 **younger ages**: need new cohorts to assess early development of risk factors

` median age of most existing US cohorts high

` studies should include data from pregnancy and first two decades of life

` should admit only very young individuals, as older subjects add unmeasured noise to prospective design

- major advantage of new cohort is to include children and their families and follow through lifecourse

17 **favor new cohort:**

- most cohorts already scoured for effects, most investigators want to collect their own data
- need to recruit new cohorts because nobody has cohorts that address the points described in the RFI
- if all existing cohorts combined would amount to only small fraction of 500K
- use of multiple cohorts necessary for adequate representation but avoid simple pooling as can=t directly compare one cohort to another
- best to have single cohort from one geographic area but not practical, not adequate representation
- advantages of new cohort (younger members, diverse ethnicity, inclusion of relatives) also involve their own considerable disadvantages in difficulties of enrollment and long-term follow-up
- absence of broad range of phenotype and environmental information in existing cohorts necessary for the study as outlined
- new cohort would allow design to handle loss-to-follow up and other issues linked with longitudinal work but also to collect the data that exactly fits the goals of this effort
- availability of premorbid measures of individual gene x environment interaction metrics not clear in existing cohorts
- despite enormous logistical challenges for creation of new cohort, would eventually be of great value
- this project requires too radical a departure in design from most existing studies to permit the false economy of integrating existing studies; very few existing studies that have surveyed environmental factors with rigor and comprehensiveness needed for goals of this project

ADVANTAGES OF EXISTING COHORTS (DISADVANTAGES OF NEW COHORT)

33 **saves time and/or money**: take advantage of extensive follow-up time already accumulated; leverage existing investment

- could usefully supplement in cost-effective way to generate short- and medium-term data
- since genetic markers do not change with time, use of existing DNA repositories would also reduce costs
- allows genetic discovery projects to begin almost immediately
- could recruit cohorts that lacked funding for genetic research as with NHANES III where blood collected from 19,553 individuals but not analyzed
- by using existing databases, resources could be dedicated to the genotyping, data set creation and statistical analyses rather than cohort assembly
- existing studies provide long-term relationship with participants, particularly for conditions such as alcoholism, drug abuse, ADHD, antisocial behavioral for which trust is critical
- years of follow-up already accrued, in children=s studies now ppts have children of their own
- approaching databases with historical information (previous examinations or electronic databases) might provide added utility

8 **experience and expertise**: existing studies and repositories have experience in recruitment, informatics development and laboratory management

- have shown they can collect high quality genotype and phenotype data
- investigators have detailed knowledge and invaluable experience with particular cohorts, could be shared and built upon
- infrastructure for data collection already in place and functioning

4 **recruitment**: should be easier, participants in previous studies may have higher response rate than new subjects; no new recruitment will be necessary

2 **ethics and community responsiveness**: study will need mechanisms for review and accountability such as patient advisory committee, ethics committee, internal review board, and a scientific advisory board

- existing studies have already created relationships with community and its health care and higher education institutions for improving accountability and responsiveness
- much work has already been done in community preparation, individual participants' and families' consent and acceptance of such research

1 **IRB and instituion specific issues**: existing studies have already worked out many privacy and confidentiality issues with their IRBs, is iterative process; already have IRB with understanding of genetic banking

institution-specific requirements: collaborative effort across many research centers will be required, but each institution will have its own procedures; existing studies will have successfully navigated institution-specific nuances, should capitalize on existing institutional relationships

4 **valuable ongoing work**: many studies ongoing in existing cohorts, should not be discontinued

` new, large cohort may be most useful but existing cohorts should not be lost; should be made more readily available to research community

` prior NHLBI effort to bring together large cohort of families 10 years ago, only now coming to stage of fine mapping and gene discovery

` don=t be too quick to abandon existing studies in which extensive effort and money invested; study of this size would divert funds from existing studies, if focus so much resources on new studies when do we take advantage of those we have

` have proponents reviewed what would be added beyond existing studies such as EPIC, cancer consortium or combining other cohorts

explore collaborating with NCS or being partner

2 **desired characteristics of existing cohorts:** should meet certain requirements in terms of consent, privacy protections, type of data collected and construction of its informatics platform

` if use existing cohorts need to be able to re-contact subjects for additional data, specimens, or consent

` ability to standardize specimen storage, retrieval and distribution for future studies is critical

11 **favor existing cohorts**:

` not good idea to build new cohort; fund consortia of current studies, pay for genotyping on samples already collected

` pooling existing studies most cost-sensible way to do this, already have NHANES, Framingham, CARDIA, Strong Heart, Nurses= and Physicians= studies but making single coherent representative project probably prohibitive

` expense hardly justified and how much would it add to what is already available

` given tight budgets and failure to adequately continue support for existing cohorts unclear what advantage to begin cohort study of this size

` proposed study doesn=t seem driven by strong hypotheses

` pooling feasible for studies of gene-environment interactions; by using existing data bases, resources could be dedicated to genotyping, data set creation and statistical analyses rather than cohort assembly

14 **combination of approaches**: should not be mutually exclusive options

` need new cohorts to supplement and extend existing cohorts but some existing cohorts could be supplemented to generate timely short- and medium-term data

` build from existing repositories but supplement as needed to meet goals of proposed project, such as supporting research to develop mechanisms to share data and samples among existing cohorts

` existing cohorts would form valuable adjunct sources of data and undoubtedly provide more detailed data in certain areas

` existing cohorts would contribute breadth and sophistication of phenotyping not attainable in large cohort, extensive characterization of social and environmental exposures to validate (and extend) those in large cohort

- synthetic cohort assembled by pooling selected existing cohorts would strengthen informativeness if incorporated into overall design of a U.S.-representative cohort study
- findings in overall 500,000 cohort could be analyzed in much more depth with the history existing studies provide
- existing cohorts can be used to validate and extend findings from large cohort and vice versa; new cohort can add new cases and increase power
- additional sites could be established and adopt protocols and informatics systems from existing cohorts for rapid start-up of remote collections to ensure desired cohort size and diversity, major new cohort could be enriched by including some existing ones
- may be wise to use existing cohorts to refine and pilot process of standardizing exposure, phenotype determination and patient recruitment
- reliable environmental data (on diet, exercise, alcohol intake, etc.) often very difficult to obtain so existing datasets may be very important, especially if they cover many years of the relevant exposure periods for certain diseases
- large national surveys such as NHANES may also be able to identify subjects with existing data and arrange follow-up or DNA collection as needed among survivors
- attractive to integrate large subgroups (existing studies of 3,000 -15,000, with long histories of data Ain the bank@), though present ppts who are subset of original group (survivors and more eager participants)
- existing investigator and administrative structures might be helpful if decentralized administration found desirable
- existing clinical and phenotypic data should be entered into fully accessible and searchable database
- existing cohorts could be chosen that were more intensively evaluated in some domains, but would have to be re-assessed to cover other diseases
- entirely new cohort should be recruited with some incorporation of existing cohorts where possible

STUDY QUESTION AND HYPOTHESES

2

- estimation of effect size of known genetic risk factors vs detecting previously unknown genetic RF
- for effect size, random population ascertainment appropriate; for detecting unknown genetic risk factors need completely different study designB cannot detect variant and estimate effect size in same study
- best done with quantitative traits as can measure in everyone
- design needs to accommodate both quantitative traits as well as categorical traits

1

- suggest gathering panel of experts in psychiatric genetics to design study
- new statistical methods have been developed such as recruiting from extreme ends of distribution, increasing power to detect linkage or LD

1

- massive undertaking, will be highly expensive; probably predicated on assumption that whole genome scans underpowered
- undertake more focused project with clear hypothesis as proof of concept; examine NIDDK Type II Diabetes Consortium; results to date disappointing

- question whether massive effort needed to incorporate environmental data into gene-finding study; if no prior hypotheses regarding interactions, results would be suspect, especially with large number of comparisons
- favor more focused project that is clearly hypothesis-driven

1

- focused cohorts can select participants of greater interest and conduct examinations more informative for those conditions
- accommodating all conditions would impose participant burdens that are not feasible or acceptable
- would stifle creativity of groups that imagine and build innovative new cohorts with creative new methods

ISOLATED POPULATIONS

1 since genetic factors haven=t been identified need optimally designed gene-detection study with large pedigrees from populations with minimal genetic and enviro-cultural heterogeneity

1 gather as isolated populations as possible with known genealogies, stable society, high-quality health care, search for all possible complex diseases and quantitative traits

SAMPLE SIZE

4

- 500,000 crazy large number to start talking about; big science has been disaster, need larger number of smaller projects at this point
- 500,000 is exaggeration
- not clear that 500,000 individuals needed
- need for 500,000 participants should be discussed; should not be needed to examine the genetic basis of common diseases, though may be proposed because of mobility of US population, to establish common environmental cohorts that disease-gene interactions might be feasible
- would seem rational to examine smaller stable populations before committing to such a large sample size, perhaps in stable population with electronic medical records where environment adequately characterized

5

- moderate sized population-based cohorts (N ~ 5,000-15,000 subjects) best powered to identify associations of common (minor allele frequency 5-15%) polymorphisms with quantitative traits, but cannot adequately address genes where rarer variants have low frequency (<1%).
- small purpose-specific cohorts of questionable generalizability, unclear if data apply to practicing clinicians= patients; large cohort will remove Amy patient is different@ ambiguity
- not enough variation in coding region of human genome to account for degree of heritability of alcoholism, much less determinants of gene by environment interactions; new cohort of 500,000 may not have enough power to examine all gene/regulation/environment/disease interactions
- large prospective study absolutely necessary for atherosclerosis and its complications; case-control studies lead to identification of alleles with unclear population attributable risk because of biased population selection, survival bias, and poorly documented environmental covariates
- major reason to carry out such large data collection effort is to validate hypotheses

		generated during family studies or other smaller studies
	`	large cohort within pre-specified disease category would give required robustness of phenotypic data to validate hypotheses that disease area
	`	would also permit other exploratory analyses or sub-studies, such as family studies, that can be conducted at individual sites
OTHER STUDIES AS MODELS		
1		Immune Tolerance Network sponsored by NIAID has already developed key infrastructure for clinical trials; may be lessons to be learned with regard to storing and processing samples including cells, DNA and other cores they
1		review findings of 1q consortium of NIDDK massive effort to find susceptibility genes responsible for this linkage so far without success
1		very careful development of questionnaire and interview items is crucial; Women's Health Initiative gives warning and poor example as many questionnaires not wisely developed
ANIMAL STUDIES		
1	`	test approach in natural populations of rats or mice to see if can correlate genes found in inbred crosses with traits; see if observational approach can work in cases where experimental approach has been successful; unclear if random sampling can work at all
	`	critical to design studies animal geneticists believe would work; nothing special about being human that makes these problems easierB in fact, almost everything makes them more complex
1		relationship of genomic information to diseases of importance to human and animal diseases, impact of environmental factors on humans and animals, genetic enhancement of health and survival in domestic animals; USDA interested in studies facilitating comparisons between humans and other mammals, especially domesticated animals
1		consider use of animal models, has been very valuable in Human Genome Project; synthetic populations of mice used at U Michigan for coplex, multi-trait genetic analysis; can be rapidly produced and genotyped
2	`	British already doing large cohort; let them spend their money and see if it works before investing in one world view just to keep up with what others are doing
	`	other countries have already been talking about this kind of study; let them spend the money
1	`	population-based ascertainment in Caucasians for diabetes would not yield multiplex families; optimum study design may differ by ethnic group
	`	diet instruments have to be customized for each population; achieving comparability across diet instruments formidable
	`	finding variants under linkage peaks daunting challenge; probably do harbor genes, perhaps need massive DNA sequencing effort of best candidates
1	`	capturing all health care events and environmental exposures should give true representation of natural history of disease and associations with exposure, but will not necessarily reveal subclinical health information
	`	large cohort study problematic due to near absence of population-based health care

	in US and lack of stable locations for current healthcare ` mobility of the US population portends great difficulty in assigning environmental exposures ` frequency and duration of population observation and data collection needs to be better defined prior to embarking on this study ` implementation of such a cohort may be premature for other reasons (biomarkers to augment phenotyping not universally feasible)
5	` ensure data available to and useful to genealogical studies, meaning ability to connect a surname with the DYS of the YDNA or mtDNA and collection of direct Y-DNA lineage and direct mtDNA lineage of participants ` if proposed new study structured to be compatible with Y-DNA surname studies, might be a mutual benefit ` include SNPs of the Y chromosome and the coding region of the MTDNA sufficient to determine their "phylogeny" ` include comprehensive attempt to establish a reliable tree of haplogroups
1	carefully planned large-scale cohort study could elucidate environmental and genetic risk factors associated second malignancies following non-melanoma skin cancer, would be valuable in establishing effective screening program for second cancers among an at-risk population
1	ability emerging to take laboratory to subject instead of bringing the subject to the laboratory; implantable, disease-specific, biosensor recording systems will revolutionize data collection for prospective study designs, a 30-year prospective study should wait for it
1	rather than initiating another large-scale genotyping effort to pull DNA from representative sample of all US individuals, attempt to find multiplex families with extreme loading for specific disorder and attempt to select groups for "purity" of diagnosis

2. Please describe the characteristics of a large US cohort study that you view as most important to include in any such effort that might be undertaken (74 responses; number of respondents per item shown in column to left).

DIVERSITY	
15	` representative cohort; rigorously population-based or Census-driven; will allow accurate estimation of allele frequencies, penetrance, relative risks, and attributable risks for application in public health settings ` large, adequate baseline participation, significant size to ensure high quality statistical associations can be identified for complex multigenic disease traits ` population diversity vital; make-up of individual cohorts less important than final population, don=t require all cohorts to be diverse ` representative of the US population, but with sufficient oversampling of racial and ethnic subpopulations to allow for adequate control and analysis of population stratification ` emphasis should be on inclusion of groups previously underrepresented in research, rather than targeting of disadvantaged or minority groups ` attention should be paid to heterogeneity of the various "traditional" ethnic groupings, and efforts to sample subgroups adequately (and categorize appropriately) should be made
7	women and men
15	every age group, wide range of age ` age range of 20-80 ` multiple ages so that pre-disease and disease outcomes begin to accrue immediately, and time available for adjustment of accrual and collection procedures ` age range from newborn to 100
3	**older ages**: ` suggest age 50-75 for reasonable cancer incidence, both sexes ` cohort should have moved through period of risk so "unaffected" cases can be considered truly unaffected ` less clear that inclusion of children and young adults feasible and as useful as persons over age 40
4	**younger ages**: ` as young as possible; will not be useful for other than common traits ` should be representative sample of women who enrolled in early pregnancy; child must be index case used to enroll other family members ` obtain cohort young enough for prospective study of onset of common diseases with origins in childhood, such as children 5 years of age, ascertained through their pediatrician=s office or through schools
1	if adult chronic diseases, then adults, but if not only childhood diseases but adolescent exposures as risk factors for adults then age include children but makes study more complicated

34	race and ethnicity
8	socioeconomic status; all levels of SES
13	` region/geographic ` major cities and counties, suburban and rural ` several urban areas of US ` special consideration for rural populations ` critical to include populations at high risk for diseases but that also tend to be mobile; risk related to poverty, poor living conditions and environmental threats should be key parts of study
2	profession, occupation
1	include target population of screened and prospectively followed Ashkenazi women from the San Fernando Valley to provide a database for assessing risk of several cancers, developing a proteomic profile, and testing interventions
1	` build on foundation of NHANES ` do not necessarily exclude institutionalized individuals
1	include invisible groups such as gay, lesbian, bisexual, and transgendered people
4	` wide variety dietary habits ` representative and diverse for other common exposures
11	` multiple diseases; as many key disease phenotypes as possible ` include common psychiatric conditions such as depression, bipolar disorder, schizophrenia, panic disorder, Alzheimer's disease, and alcohol and drug dependence ` do not neglect mental conditions and substance disorder ` include infectious as well as non-communicable diseases ` study resilience as well as susceptibility ` include rheumatoid arthritis ` include sleep disorders such as long sleep and delayed/advanced sleep phase syndromes ` include essential tremor ` include auto-immune diseases and inflammatory responses such as atopy, asthma, atopic dermatitis, allergic rhinitis ` include diabetes, hypertension, atherosclerosis and renal failure ` include cancer of breast, prostate, lung; particularly regarding environmental stimulus that increases susceptibility in genetically predisposed group ` identify patients hospitalized with infections and track progression to sepsis, acute organ dysfunction, other relevant clinical outcomes including death
1	environmental exposures best studied in stable population
HOMOGENEITY	
1	` as ethnically homogeneous as possible, involve largest possible families, emphasize as many quantitative traits as possible in same individuals rather than ascertaining on diseased individuals from ethnically heterogeneous populations, would be horrendously inefficient for gene identification ` spreading sample across variety of ethnic populations would dilute power; no

	reason not to sample most culturally and genetically homogeneous populations available with largest possible pedigree structures
DATA ELEMENTS	
9	**phenotyping** • uniform and very high quality phenotypic characterization including access to electronic medical records, ability to link to existing public records • accurate measures of physiological and pathological traits relevant to disease and disorder including mental health and neurocognitive markers. • all centers participating must be using the same protocols with accurate genotyping methods • crucial for the cohort to be randomly chosen without regard to level of the phenotype of interests, so that the generalizability to the major public health issues is high • geocoding needed to link with array of environmental data now digitized and available; social security numbers to provide linkages with health service and other data
18	**environmental exposures** • need high quality lifestyle, dietary, activity, environmental and occupational/workplace exposure data, include alcohol use • complete and unbiased ascertainment of environmental risk factors • validated exposure questionnaires and biomarkers • development of environmental scan technology on level of genomic scans, using unique aspects of immunologic memory and metabonomic methods • state-of-art measures of multiple outcomes, including measures of subclinical disease, quantitative variables, biological markers, sophisticated assessment of environmental exposures • critical questions include depth of collection of behaviors [e.g. diet and physical activity and smoking and drinking] vs only biomarkers and direct morbidity measures • periodic personal monitoring of environmental exposures • comprehensive suite of disease and health-related phenotypes, environmental factors should be measured/ascertained
6	**specimens** • biological samples, adequate DNA • immortalized cell lines, serum, plasma, timed urine • sufficient serum samples for metabolomic and proteomic analysis • complete and periodically repeated capture of biospecimens • collection of pathologic specimens • post-mortem analyses/accrual
1	ascertainment needs to be in some unbiased way, not through advertisement perhaps through Medicare, large insurance such as Kaiser Permanente study in California
1	ethnicity shouldn=t be merged with race; attempt to identify culture with which ppts feel

	most comfortable (as most likely to adopt lifestyle of that group)
18	` quality of life, people=s beliefs and expectations ` accurately determined ethnicity ` matrices of ancestry at least back to 4 grandparents, including ethnicity, birthplace, and language spoken ` acculturation, immigration ` occupations, past and current places of residence ` as much demographic data as possible; socioeconomic status ` reproductive and family history ` use outside of cancer for health disparities in medical care and other diseases ` health practices, including access to and use of health care services ` adverse reactions to drugs, treatment programs ` birth weight ` perinatal stress associated with asthma; history of traumatic events related to asthma, pelvic inflammatory disease, chronic pain, borderline personality ` DNA adducts ` typing of highly polymorphic human HLA region ` study genetic stratification using STR, SNPS, mitochondrial DNA and Y chromosome ` sun exposure, molecular markers of inflammation ` total T3, free T3; reverse T3 on pregnant women ` in addition to biological specimens and existing health and trend data include a qualitative element at every research level ` questionnaires can determine circadian rhythm disorders (morningness-eveningness) and depression ` shift work is important environmental influence which should be carefully measured; measurement of sleep using wrist actigraphic technology or home oximetric monitoring for sleep apnea would be of interest (but not high priority)
1	simple design, such as MRFIT screenees, with a very brief exam, blood and DNA would be key elements cross-sectionally
1	way too many potential variables even if measured just known ones for one disease; would project attempt to go back and measure every newly identified variable on all 500,000?
FOLLOW-UP INFORMATION	
13	**completeness**: need mechanism for disease ascertainment with documentation ` high follow-up rate, complete ascertainment of mortality ` access to pathology specimens as needed, able to link to health care records ` most important is iron-clad ability to maintain contact and follow-up; cannot just collect DNA and medical records, must have ability to go back for follow-up data from subjects themselves to confirm diagnoses ` minimize burden to ppts to maintain adequate follow-up ` other than mortality and cancer, few good methods of passive follow-up in US; this is why so many large cohorts now being conducted abroad

	` more diverse and representative cohort is at enrollment, less likely to participate in active follow-up; this is why almost all successful cohorts are select populations ` availability of detailed SNOMED information valuable ` follow all ppts at maximum 5-year intervals, more frequently if possible; follow affected individuals to determine course of illness, including remissions ` high follow-up rates for a wide variety of diseases and phenotypes ` measure change over lifecourse and risk in vulnerable windows
1	consider incident diseases as outcomes; painstaking and unbiased ascertainment of diseases of interest, need expert clinicians willing and able to do this
1	maintaining follow-up major problem, only fraction of cohort studies still running after 10 years, Harvard Cohort Studies using highly motivated health professionals is one of few proven methods
1	collect identifiers to be linked to National Death Index (for mortality follow-up)
NOVEL MODELS	
1	` Million Person Quest for Health Diverse Builds engages communities, uses web based self reporting tool and medical exams and records, provides place to store medical records as perk; consider using community MDs and clinics, provide services in course of measuring ` ` utilize mobile medical stations across lifespan
1	real opportunity on informatics side to create model for distributed medical record; model personalized medical record system analogous to bank statements, driving records, criminal records, etc
1	` such study worthwhile and critically important but only if can implement radical change in approach to maintain high levels follow-up and quality data on clinical diagnosis ` previously have used selected groups with high levels education and access to care but need to broaden but avoid low follow-up and poor ascertainment ` consider providing ppts with high quality free health care
1	now more than 50 primary care practice-based research networks, might be ideal for recruiting and following large numbers of patients with common diseases
1	consider APeace Corps@ model, creating volunteer army of dedicated individuals and create sense belonging to larger organization serving greater good
1	consider national genetic census week to test high priority candidate SNPs against particular phenotypes, with 5,000 centers making measurements on 100 individuals each, using a web-based protocol, with bloods sent to a national center for genotyping on a low-cost microarray platform
1	solution to technical problem regarding obtaining DNA starting material: use leukocyte depletion filters used by all US blood banks for donated blood; since 1% of population donates blood in US, available number of samples about 2.5 million, most blood donors would be willing to participate
OTHER	
1	hard to determine characteristics without specific disease in mind; need to define major diseases

	again, what are existing resources addressing these issues, will another be that much value added or should existing be strengthened and broadened to cover gaps
1	normal variation seems not to be same as variation at high-risk end of spectrum
1	willingness and mechanism to participate in longitudinal study; effect anonymization to protect identity but make comprehensive meaningful data available to qualified teams; need incentive for ppts to remain and continue reporting
1	**logistic considerations**: ` longitudinal: track changing health status and environmental influences over long period; track outcomes associated with samples when possible ` standardized protocols: standards agreed upon and strictly enforced regarding how ppts consented, privacy protected, demographics collected, clinical data updated, electronic data stored and protected, biologic specimens collected and stored ` adequate funding: estimate will cost $10-$13M to collect 100,000 clinically annotated DNA samples over six years at our institution, including costs related to project management, recruitment, informatics development and sample processing and storage ` comprehensive, flexible consent form: ensure data and samples can be utilized for broad and unspecified research projects (but with regulatory oversight) by third parties (academic, government, corporate, national or international). ` privacy and security: patient identifiers are removed and replaced with barcode, single link remains between DNA sample and database and patient identifiers; physicians not informed of patient=s participation or results ` compliance with regulatory authorities: data derived from this study should be used to support development and marketing of products to improve diagnosis and treatment, so study must adhere to regulatory guidelines and quality control standards to support FDA submissions by third party users (for example, compliance with GLP and 21 CFR pt 11). ` deposit genotypes into database: enrich study over time by re-depositing measured genotypes into study database for future use ` access to electronic medical record: records in electronic format would facilitate automated creation of phenotypes while minimizing frequency and amount of interaction with each participant ` standardized informatics platform: would enable aggregation/networking of collections across institutions; will need to adopt standards in medical terminology ` system for ethical reflection: need broadest viewpoint possible by both internal and external ethicists; include on site bioethicist, national external ethics board, patient advisory committee
1	most critical factor for success will be flexible but robust infrastructure to accommodate variety of analyses, requires: data standardization, accurate and high throughput retrieval of specimens, fail-safe storage, efficient and ethical usage of data and specimens with highly defined process to access and distribute, facilitation of multisite collection, consideration of centralized vs distributed storage
1	do not allow multiple investigators to approach subjects directly but funnel requests

	through research team that ascertained subject initially
2	` broad consent for testing hypotheses not foreseen at outset ` streamlined and global informed consent for future and unknown genetic and molecular studies
1	characteristics important to include in large genetic epidemiology cohort study ` ability to collect DNA and related biospecimens. ` ability to collect phenotypic data in consistent, reliable, valid manner ` ability to collect risk factor information; as much attention should be paid to collecting state-of-art risk factor information as to collection of DNA ` ability to follow cohort members over substantial period ` close involvement of community (non-researcher-trained) members in design, execution, and communication of study ` informatics support, large sample size, broad array of laboratory and clinical data, and availability of data to broad scientific community.

3. Please suggest the family structures, and proportion of related individuals, that you would recommend for inclusion in the proposed study (66 responses; number of respondents per item shown in column to left).

TYPES OF RELATIVES	
14	**first degree** ` focus on first degree; recruitment more distant difficult ` substantial proportion of parent offspring trios though challenge for late onset diseases ` first degree required though three generation may be optimal ` siblings and spouses useful and feasible ` use only first degree relatives ` first degree relatives critical; second degree relatives useful ` sample households and include all first degree relatives of randomly selected propositus residing within pre-specified radius ` nuclear families with at least two biologic children plus additional members if available ` at least three first degree relatives
4	**minimum family structure** ` include index child in birth cohort, mother and, when present, father; siblings of index child; grandparents of about 50% ` always include index case, parents, and as much as possible affected and unaffected siblings ` ideal would be minimum of four family members (bio-mother, bio-father, and two offspring)
13	**three generations** ` at least: 1-2 sibs plus parents and grandparents, could be prohibitively costly and open-ended ` at least 3 generations deep and include as many individuals as possible across generations ` first degree relatives through first cousins ` all available first and second degree relatives ` sib pair with parents make good general purpose family structure ` include or add pregnant probands for pre-natal exposures ` include third-degree relatives only if affected ` to detect environmental interactions with genes will want large families, at minimum trios, better three+ generations; also want many unrelated by geographically and culturally related individuals ` large African-American families with several generations ` also include cousins and more distant relatives to help minimize shared family environmental effects, important when modeling gene-environment interactions.
9	**extended families** ` large families: maximize power for linkage and LD with more individuals and fewer

	independent chromosomes ` large proportion (70-90%) should be recruited from families and multigenerational pedigrees to produce kindreds for linkage analysis ` large kindreds would serve case control association studies as well as formal linkage studies and other designs such as TDT ` multiple generations would be an asset, as would large pedigrees ` a balance of large and small pedigree sizes will provide the best representation as selecting only members of large families will be biased ` large fraction, perhaps ideally even all, should be members of extended, multigenerational families; large sibships are distinct advantage, recruiting in a Mormon population would be attractive ` most powerful structures for study of gene-environment interactions are large extended families ` power maximized when sibships within extended families contain large sibships but to avoid ascertainment bias would sample families randomly ` strive for complete families in whatever fraction of total devoted to families
2	**twins** ` more twin studies including twins raised apart ` include 1000 MZ twin pairs and 1000 DZ twin pairs
1	**trios** ` trios useful for specifying SNP haplotypes, used in HapMap, correct for population stratification ` family triads would allow excellent control for racial stratification at cost of modest decreases in case-control power; mix of triads and isolated subjects
3	**alternative family structures** ` unusual relationships beyond typical nuclear families, such as twins, half-sibs, adoptees, can be helpful in assessing degree to which genetic and non-genetic factors contribute to disease or trait ` specific targeted effort to recruit few thousand families with 2 or more adopted children that adopted as babies and who did not change families
1	range of options: ` spouse pairs efficient for recruitment and sampling, provide spousal controls ` siblings limited utility for linkage analysis, likely to be replaced by association analysis, could be useful for matched case-control studies with sibling controls ` nuclear families natural way of recruiting young children, can do range of association studies if emphasis is on young children ` three generation pedigrees typically difficult for population-based study but could study how SNP haplotypes transmit across generations and how they associate with health outcomes
1	investigate products of conception including spontaneously and voluntarily aborted fetuses
1	` include range of family structures, address single-parent as well as blended families, extended families, and same-sex union and/or parenting units

- non-DNA-related family units would provide a control in studying DNA-related phenomena or data

FAMILIES VS UNRELATED

10 **advantages of families**

- family studies seem more efficiently done in case-control designs; if outcome of interest has familial component, some power will be gained over unrelated individuals, but probably not greater than 2-fold
- family-based studies have limited power to detect small increases in risk likely for specific polymorphisms in complex diseases
- much benefit in assessing population stratification might gained from sampling only DNA from parents of participants, but would be challenge in older ages
- include families in accordance with scientific questions; in absence of major scientific questions, difficult to speculate on proportion of related individuals
- depends on goals of study; if major goal is to look at familial contributions include them but otherwise don=t make related individuals a large part of population
- ideal ratio related:unrelated individuals depends highly on trait in question
- obtaining good family histories at least as important as obtaining multiple members of families prospectively
- being able to use cohort family histories to identify and subsequently enroll families would be important new resource
- informed consent issues are challenging but not impossible; being used now
- family data would be important as HLA inheritance could be followed with disease occurrence

13 **advantages of unrelated**

- family structures not so crucial for association studies per se since confounding can be addressed through genotyping random SNPs
- population-based seems more desirable than families to avoid bias and reduction of effective sample size
- inclusion of families adds complication of confirming genetic relationships and deciding how to proceed if differ from those reported
- primary structure should be rigorously population-based, select primarily unrelated individuals with secondary identification or subsample of living first and second degree relatives of some ppts
- utility of cohort will be identifying broad differences between populations with and without diseases; family studies may be useful in confirming or validating genotypic and phenotypic associations
- study should look at population genomics to validate hypotheses developed in family studies rather than trying to replicate family studies on large scale
- families would reduce susceptibility to confounding errors but recruiting family members less efficient, particularly in late onset diseases; recommend at most 20% of the cohort to be family structures
- propose use of unrelated individuals as much as possible
- depends on study question, most existing cohorts have few related subjects so choosing family structures would mean mostly de novo recruitment

9 **combination of approaches**

- component should include family based studies; few have expertise in collecting such families and they should be encouraged to participate
- balance between related and unrelated individuals to ensure data most generalizable, help avoid analytical problems from complex pedigrees
- include families with sufficient numbers of pedigrees to screen for polymorphisms affecting traits of interest, control for stratification, reduce signal/noise ratio, provide solution for multiple testing problem
- 3-4 generation pedigrees should account for 10-20% participants
- family structure less important than high follow-up rates; some family recruitment may be appropriate or additional family members in interesting families, but focus on longitudinal population-based studies
- consider 5,000 nuclear families (size 4), additional 50,000 sibpairs without parents, remainder unrelated
- portion of study should be family; even better if portion can be twins
- portion should include trios
- data models and software systems should be able to incorporate family relationship and family medical history data
- recruit related individuals comprising 10% of defined demographic cells

4. Please identify the most relevant issues concerning the power to detect genetic effects (such as environmental risk factors, heterogeneity, prevalence) for diseases or traits that you would be interested in studying in a large US cohort study (68 responses; number of respondents per item shown in column to left).

FACTORS DETERMINING POWER	
1	factors well known; focus on gathering information beginning with conception
1	**large pedigrees**: focused study of largest possible pedigrees with smallest Ne and highest Fi be more powerful for any quantitative trait **quantitative traits**: dichotomous disease is horrible trait in most cases generically; animal geneticists study glucose or related QTs rather than diabetes
6	**incidence/prevalence**: absolute number of persons with incident or prevalent diseases; incidence and prevalence known for most common diseases and traits ` pursue traits with prevalence 5-50% ` environmental factors would largely have to be obtained by questionnaire ` power study to detect differences in disease incidence and for certain diseases, differences in survival ` prevalence of SNP or haplotype also important, data more limited, especially in ethnic subgroups
7	**gene-environment interactions** ` issues determining power include effect size, frequency of underlying variants and their population distribution, accuracy of clinical phenotypes and subtypes ` role of environmental effects in gene finding is controversial; unless interaction deviates significantly from multiplicative model, unlikely to be much gain in power ` duration of follow-up and genetic heterogeneity also influence power ` need to identify confounding phenotypic and environmental variables ` need to incorporate main effects of at least established environmental factors on both children and parents and potential gene by environment interaction into analysis ` heterogeneity in phenotypes (and potential misclassification), heterogeneity in genes (ie, multiple causal genes or gene by gene interaction), and heterogeneity in ethnic background
1	**context-dependence** ` risk alleles produce different effects depending on gender, age, genotypes at other loci, environmental factors; will be ideal to split sample into many different groups ` even with 500,000 subjects, difficult to study thoroughly any but the most common risk allele
1	` ability to characterize environmental and behavioral risk factors, phenotypes of main interest, and their frequencies of occurrence would all be significant challenges for a study of this size and complexity
1	` recognize that there may not be any significant genetic risk for particular disease, which is why we have planned to scan ~100 diseases in parallel

	STUDY DESIGNS TO MAXIMIZE POWER
1	power inexorably weak in absence of idealized ascertainment schemes
1	**cohort design inefficient**: for rare genetic diseases power from even large general population cohort always limited, cohort study never most efficient design
4	**environmental measures** ` enhance power by having maximum variability in exposures and genetic characteristics, so conduct in urban areas where higher probability of exposures to environmental and occupational RF, wider range of racial/ethnic groups ` wide range environmental exposures ` at least as many resources should be dedicated to measures of the environment as for the genotyping ` emphasis should be placed on capturing heterogeneity in exposures, so that we can understand how risk is related to both susceptibility loci and specific exposures, throughout the lifespan
8	**genetic heterogeneity and population stratification** ` proper case-control matching to avoid population stratification is key ` samples could be genotyped for common panel of markers to detect stratification, and those data made available to all researchers ` genetic heterogeneity will be problem within race/ethnic groups ` reduce heterogeneity by collecting detailed ethnicity data on parents and grandparents or haplotype analysis ` splitting sample among different ethnic groups might be politically correct but is scientifically indefensible ` best to restrict study to individuals of northern European ancestry because this is largest group in America, most existing cohorts are of this ancestry, and family structure is generally good; if study includes more than token number of African Americans, for example, it should be nearly 100% African American ` ethnic stratification will affect both genetic and environmental variables and uncertainties in Amixedness@ will drive up number of persons required; suggests either use of very large population samples or use of samples from very large but genetically relatively homogeneous population such as Han Chinese initially ` clustering ethnicity into small number of categories would be useful so we do not parse down 500,000 to a meaningless amount
1	` longitudinal follow-up is most useful; don’t just pick 500,000 cross sectionally, but then subject protection becomes issue ` large government study may not be trusted to be confidential and not be abused
DATA QUALITY TO MAXIMIZE POWER	

18 **environmental exposures**

- rigorously classifying environmental exposures and behaviors crucial for maximizing power
- need high quality repeated measures of environment and lifestyle, combined with high follow-up and high quality assessment of outcomes and phenotypes such as CV structure and function, CV events
- need objective measures of environmental exposure during critical windows of vulnerability
- environmental risk factors, life style factors, nutrition factors, disease prevalence, environmental and genetic heterogeneity and family history are all important factors to consider for powerful analyses
- for environmental factors, most any random subsample with cell sizes > 3000-6000 will meet most needs but for gene-environment interactions these are minimal subsample cell sizes
- by prospectively identifying patients and collecting key social, cultural, ecologic, demographic, educational, economic information, relative environmental and genetics contributions to sepsis may be elucidated
- high throughput, multidimensional, potentially automated data collection systems of risk factors and clinical phenotype will enable detection of etiologic differences based on underlying genetic and environmental factors
- robust neuropsychological measures that can generate continuous measure
- in person psychiatric interviewing as part of a integrated global phenotyping of total behavior
- thorough survey of subjects’ environment

1

- availability and comparability across cohorts of environmental risk factor data will be limiting issue
- other issues: minor allele prevalence, functional importance of polymorphism, locus heterogeneity, frequency of environmentally-induced phenocopies, etc; population heterogeneity

STUDY QUESTIONS

1

- unclear what trying to address: power for interactions is function of main effect, prevalence of factor under study, frequency of outcome of interest
- if main goal to study genetic factors, don=t need expense of cohort study, go with large case-control studies for diseases of interest
- most prevalent potential genetic susceptibility factors and common disorders could be studied
- for other exposures likely to be significant contributors, that do change with diagnosis (unlike genes) case cohort study is better design

3

- list a set of major hypotheses and assess power for answering those; what is minimum OR with SNP haplotype detectable
- gene-environment interaction: try to identify host of environmental factors, perform systematic analysis to identify GxE interactions; we have developed one such method

	` methodologies: sampling, measurement error in environment and genotyping, binary, continuous and censored phenotypes, evaluation of sampling designs ` for CVD gene-environment interactions, particularly for exercise and diet habits, will be particularly important, as will gene-gene interactions
5	` large cohort, prevalent disease: need very large cohort for rare diseases ` for common diseases, little interest in large cohort to detect small relative risks, rare exposure-disease associations or multi-level interactions ` diseases and polymorphisms will need to be of sufficiently high prevalence ` if consider all important variables and thousands of markers, power seems inadequate for all but common diseases ` rank diseases by heritability, prevalence, health care burden; puts CVD, obesity, diabetes and psychiatric diseases at top
1	` past 50 years of epidemiologic work have shown clearest and most reliable and useful results are for very strong risk factors (smoking) or very rare ones (asbestos); did not take huge samples to show this ` major set of exposures will be diet and exercise but notoriously difficult to ascertain in real time let alone by interview or retrospectively ` secular trends will add additional but necessary complexity ` before launching such an expensive effort take serious and fair-minded look at what has actually be achieved by existing cohort studies
1	difficulties of late onset disease; die before develop disease, others develop after assessed in study
1	for CAD most genetic effects likely to be modest and interactive, so need to genotype large number candidate genes and collect high-quality environmental measures
1	small twin component would greatly increase power to differentiate genetic effects
1	heterogeneity of exposure is key determinant of power, also stability of exposure over time, limits value of one-time assessment; need repeated measures updating exposure status every 2-4 years, repeated blood samples
1	` if "capacity" is ability of a person (at cell, organ or organism level) to perform some specific work such as fight infection, metabolize glucose, etc., then concepts of “gene x gene" and "gene x environment" are too limited and "capacity x environment" concept might help ` would then need considerable thinking around various capacities to be studied and information needed to operationalize them
OTHER	
1	` build on state of art surveillance such as CDC developed, take to next level ` extremely large numbers of individuals required to see effects ` consider multi-level participation, baseline everyone would do, more information gathered from smaller willing cohort
1	abandon prejudice against trauma in context of genetic research; at issue is predisposition to traumatic injury, may be connected to risk taking, predisposition to addiction, depression, regulation of inflammatory response
1	particularly interested in autoimmune diseases such as RA and SLE, need extensive

	clinical and laboratory/serologic data	
1	`	calculated that cohort of 10,000 persons for each of 100 diseases in which exons and splice sites of all known genes scanned for single base changes would permit detection of any monogenic condition of genetic risk in which less than 80% of the population carries genetic risk
	`	such a cohort would permit detection of multigenic and polygenic risks; even were eight genes to independently confer risk for a particular disease, comparison of allele frequencies in afflicted cohort to those of the 99 unafflicted (control) cohorts would permit unambiguous identification of risk contributing genes and more frequent mutations (hotspots)
1	minimize genetic heterogeneity; each racial group must be analyzed separately sampling scheme and study design will need to be determined based on primary trait(s) chosen for project	
1	low penetrance, heterogeneity, phenocopies will reduce power	

5. Other information not specifically addressed by the comments above, but considered important and relevant to the development and implementation of a large-scale study of genetic and environmental influences on common diseases, would also be of considerable interest and value (80 responses; number of respondents per item shown in column to left).

IDEAL SETTINGS FOR STUDY	
2	very large (3 million), diverse and representative population and the integrated nature of the health care delivery at Kaiser Permanente of Northern California makes our environment an ideal setting for a cohort study like this; Kaiser colleagues have been working on something like this
1	administrative databases of Saskatchewan provide excellent cost effective setting, 1 million persons in whom 95% electronically tracked for medications, hospitalizations, cause of death since 1976
1	I have four computerized genealogical databases for 50,000 individuals in my own family across 60 generations; 5,000 live within 130 miles of me; propose to identify gene or genes associated with hypertropia
1	` CCAAPS birth cohort genetic database is prospective birth cohort study designed to elucidate environmental and genetic factors important in development of allergy in children ` complete set of samples and clinical profiles from AA Pediatric "Autoimmunity" cohort, has DNA samples from approximately 1000 probands including approximately 175 affected sibling pairs ` our cohort would thus be very useful for validation studies and to answer specific questions in preparation for a large-scale population study
STUDIES INTERESTED IN COLLABORATING	
1	opportunity could be missed by excluding population such as Iceland, with homogeneous population, environment, access to health care, treatment regimes and diet; arguments for and against founder populations but recent years have shown utility, our project overshadowed by deCODE would benefit tremendously by collaboration with NIH-NHGRI
1	currently coordinating large European study, GENOMOS, on genetics of osteoporosis, involving over 20.000 subjects (mostly elderly men and women) from 8 study centers across Europe; would be very interested to learn more about possibility of joining forces with US initiative
1	` currently coordinating pooling project on genetics of cardiovascular disease ` MORGAM began in 1998 and based upon cohorts recruited as part of WHO MONICA Project though membership not limited to MONICA ` have 16 cohorts in 8 countries for total genetic cohort of ~100,000 with estimated 5,000 CVD events ` have overcome many problems in pooling these cohorts, idea of degree of complexity illustrated by MORGAM website at: http://www.ktl.fi/morgam/internal ` MORGAM Management Group interested in RFI, would like to be kept abreast of developments, happy to share experiences, insights into ethical, biological and

	statistical problems of international studies involving DNA ` comparison of Americans with background European populations from which their ancestors emigrated would be fascinating ` hope that at some stage in the future we may be able to collaborate with you
1	` in planning for such a cohort study, critical to include individuals with most experience in this area, namely epidemiologists ` human geneticists have very little experience in conducting population studies ` genetic epidemiologists who focus on analysis of genetic data may not have requisite experience in constructing, maintaining and managing a cohort ` recommend including broad range of epidemiologists from various areas of chronic disease, such as cancer epidemiology, neuroepidemiology, psychiatric epidemiology, cardiovascular epidemiology, reproductive epidemiology ` ` also important to include expert in exposure assessment, such as nutritional epidemiologists or environmental epidemiologists
1	recently initiated surgical collection of thoracoabdominal aortic aneurysm and dissection cases and specimens would form valuable cohort
1	patient collection for our association study of 4000 individuals just beginning; we would be interested in possibly merging our study with your larger study
1	large health care organization in Wisconsin in process of deciding if large cell/serum/DNA database can be created to study genetic and environmental disease in Wisconsin; will consider how to organize possible Wisconsin study to facilitate participation in national effort
APPROACH AND HYPOTHESES	
1	` large scale studies of this type are probably ill-advised, overpriced, and underpowered ` focus on large families from small populations ` look for natural experiments to answer other types of questions, can=t do in factory science mentality seeming to motivate this proposal
1	• we would like to be party to discussion of relative merits of prospective study versus novel retrospective approach we have been developing • not clear from RFI whether our assumption of multiallelic risk for common diseases based on analogy to more than 1400 rare diseases is included in overall consideration of large population studies • we regard question of multi-allelic risk as crucial and look forward to exchange of perspectives on this issue
1	maintaining follow-up is a major challenge; understandable instinct to randomly sample needs to be balanced with diminishing returns for follow-up that can be associated with this approach
2	• include cadre of appropriate social scientists as mandate in: 1) design of internal or external projects or RFAs, 2) reviews of proposals, 3) any funded project • linkage with social scientists is not exercise in politeness, but exercise in good science; leaders of this project must ensure the right personnel are incorporated to

	avoid a wasteful debacle • include demographers who have true "population perspective" not common among geneticists and genetic epidemiologists; a nationally representative sample with nested sample of family members must be carefully designed to insure demographics of a representative sample can be matched
1	nutrigenomics is hot topic, and very important in terms of improving our understanding regarding the effects of dietary modifications in prevention and treatment of chronic diseases, including diabetes
1	ethnic minorities must be included due to their unique population history, generally lower level of social resources, and often high prevalence of common disorders
1	• clear need for NIH to design study to oversample for Hispanics from the whole range of Hispanic sub-groups to make real progress in eliminating health disparities in Hispanics • essential that common chronic oral conditions, such as periodontal disease, be clinically documented in detail as part of the planned NIH study
1	need to better define Acommon disease not currently addressed by existing resources@
1	important to consider ELSI issues, collect information about quality of life, disease worry, health and life insurance status, education, employment, knowledge of genetics, and willingness to undergo genetic tests
1	• many people opt out because of needle phobia, and logistics of working with blood are difficult; if could make immortalized cell lines readily available materials such as cheek cells would remove significant barrier • another barrier is participant time; 6-10 hours of interviews is absolute limit • if there is access to medical records information across multiple sites may not be standardized • if record access limited or unavailable, data will need to collected on all 500,000 people which limits the information that can be reasonably collected • consider training selected group of primary care physicians to do assessment on their own patients, and pay enough for this assessment to make thorough work feasible and worthwhile; this might be ideal design
DIVERSION OF RESOURCES	
1	resources will be diverted away from more efficient means (case-control studies, combining existing cohorts)
2	• we hope planning for this large cohort study will not eclipse plans for implementing the NCS • because new cohort proposes to build on National Children=s Study, using established cohorts would inappropriately compete with the NCS
1	` studies like this never end and become unstoppable resource drain; this study will be so expensive that it will threaten all sorts of other research ` best science done in focused way, even if genomic in scale and hypothesis-generating ` history of large longitudinal studies is there are always reasons they should not be

	ended, yet they become 'tired' data. Once box is open, it can't be shut, but opinion widespread that this box already open and it may be too late
1	putting this much funding into single undertaking by limited group of investigators would suppress initiative and decrease resources for other creative, energetic and smart investigators
NEED FOR REPLICATION, COLLABORATION	
1	` putting all eggs in one basket is serious mistake, need replication from variety of studies in multiple populations than from one study which may have flaws ` opportunity for replications may be lost if resources diverted to this ideal study ` ` main reason for such a resource is to address less common rather than more common onesB define what mean by common
1	the design and implementation of a common and complex protocol across multiple sites is a challenging endeavor that requires collaborative teams of investigators with expertise in several disciplines
1	` prospective intervention studies with large number of persons are needed. What I am also suggesting is the use of all the data available from former large scale intervention trials, e.g. the Finnish Diabetes Prevention Study (DPS) and DPP ` these studies are very expensive, and therefore, collaboration in this field is the only way to have a real success
1	proposed study should link tightly with the NCS; this might be umbrella for existing and new cohorts, with NCS core study being a major (or the major) new cohort
1	might be useful to include a subset of cohort from overseas, to compare some environmental or genetic factors more systematically
RESOURCE	
1	` reiterate importance of cohort as central resource; creating cohort largely beyond reach of R01 investigator ` can perform genotyping once and all genotype data available as central resource ` accessing a common resource facilitates collaborations, eliminates excessive duplication ` make all resources available in timely fashion with no strings attached ` ` could also be used to provide general controls; cost savings could be substantial
1	need careful attention to bioinformatics and biostatistician issues; cohort collaborations tend to underestimate resources required to deal with programming and statistical tasks for complex gene-environment analyses
1	` need high level of collaboration/cooperation ` understanding of how data shared and credit given to those who collaborated ` adequate funding for high quality genotypic and phenotypic characterization
OTHER	
2	**community/social issues**: need mechanism for community participation, particularly in communities impacted by environmental concerns that affect health status ` health disparities will need thoughtful consideration ` attention to issues such racial stereotyping, justice, identity and difference will need

	to be carefully structured in this effort ` how to assure ppts that unfair or unjust profits will not be drawn from the samples ` can we learn from public debates surrounding UK biobank ` ` issues to be addressed include: sponsorship and benefit sharing, neutrality and regulatory power of ethics committees, public engagement, consent, data protection
1	**providing results**: need to balance right to privacy with access to one=s own genetic information ` important to manage expectations regarding types of Agenetic information@ that might result from large cohort studies, such as time needed for research results to become clinically applicable, and how individual participants will gain knowledge of this information
1	Congress must support; epidemiology is critical next step in process of translating this information to useful tests, treatments and services for all Americans
3	` consent will be key issue ` our study (and others) have "layered" consent forms with a variety of restrictions participants may want on the use of their DNA; challenging due to many possible permutations of participant preferences ` carrying these wishes forth as a part of a pooled cohort will require review of original wishes of cohort members, and/or re-consenting; either way, participant education regarding the ultimate "fate" of their samples is mandatory ` some groups may require more reassurance than others about intended uses of their genetic material, such as ethnic minorities who may lack trust in a "federal" study of genetics ` privacy issues are largest obstacle to recruitment and sharing of data and materials, but risk to participants in observational studies are minimal and privacy risks may be overstated. ` current approaches to privacy management involve great cost and effort to avoid largely theoretical risk of harm ` might be useful to sponsor national conference to seek ways to simplify privacy management, including recommendations for robust laws prohibiting genetic discrimination and perhaps modifications of HIPAA as applied to research, before starting a genetic study of this size and scope ` heartening to note from the requests for comment that state of the art research consent process is to be integral to proposed study
1	**informatics** will be critical ` Duke Databank for Cardiovascular Disease is model for longitudinal data collection in clinical practice setting but required 30 years to develop nomenclature and standards for Duke Databank, this time frame suggests caution with regard to this effort ` inclusion of multiple clinical settings would increase cost and complexity of informatics including database development and standardization

1	A universal data structure needs to be created and there needs to be agreement on various data elements if the data are to be pooled from various cohorts. The means to make these data sets available to a geographically dispersed group of investigators is a major challenge and will require sufficient IT support
1	calculations performed in developing Global Cardiomics Network show that recruiting all patients within specific disease category more economical than recruiting patients based on pre-determined clinical protocols of subsets of patients of interest
1	` estimate storage of samples at centralized facility much more efficient than storage at multiple locations ` for proteomic, metabolomic or nucleotide analyses, must store -80 C in multiple small aliquots to prevent degradation through freeze-thaw cycles ` ` for cohort of 500,000, assuming storage of 15ml of blood, cells or plasma per patient in 100pl tubes, need facilities for over 10 million individual sample aliquots
1	consider including technology to very reliably identify study participants such as biometric identification based on iris
1	digestible (available) and indigestible (unavailable) carbohydrates should be segmented more extensively in data collection and analysis
1	considerations for study design: ` pick set of diseases to focus on, determined by clinical relevance, population impact, and suitability for genetic study ` fund collaborative studies to conduct pooled analyses of existing cohorts, both to generate preliminary findings and develop questions for future study ` fund second round of studies for additional primary data collection efforts on existing cohorts and new supplemental cohorts ` contract with one (or more) organizations to coordinate data collection and statistical analyses efforts across diseases ` create mechanism for facilitating sharing of data, analytic approaches, clinical assays, etc.
1	value of very large sample sizes in epidemiologic studies has been illustrated by MRFIT screening study (n = 360,000), Nurses Health Study (n = 100,000), and Million Women study, but giant studies raise other problems: cost, difficult administrative structure, central involvement of wide range of scientists, data management, competition for data analysis topics and stored specimens
1	` need better molecular classification of disease, particularly of cancer, rather than continue to rely extensively on histopathology; need to collect fresh tissue for use of protein profiles or mRNA expression ` best approach to develop early detection profiles is to follow large number of individuals over a prolonged period, sampling blood and other body fluids at multiple timepoints to establish, say, protein profiles with predictive power
1	` concentrate DNA repository in 1 or 2 centers to minimize sample mishandling, lack of comparability ` all phenotypic and DNA testing data from this project should be entered into subjects' electronic medical records

1	`	in long run will learn more about genetic epidemiology from mining data generated for health care purposes than from dedicated studies
	`	understanding of genetic epidemiology will only become satisfactory, when substantial DNA testing is routinely performed on (nearly) every patient; resulting data could be mined for research purposes
	`	patients in such an approach would pay for a large fraction of the costs and would benefit directly from improved health care and indirectly from new research results
1	`	convene one-day meeting of experienced investigators of cohorts, and definitely include inexperienced investigators; avoid posturing of experienced senior people which can stifle open discussion and improvements in a mixed forum
	`	give less experienced investigators, such as those who run case-control studies, equal time; you'll get new ideas