DESCRIPTION: State the application's broad, long-term objectives and specific aims, making reference to the health relatedness of the project. Describe concisely the research design and methods for achieving these goals. Avoid summaries of past accomplishments and the use of the first person. This abstract is meant to serve as a succinct and accurate description of the proposed work when separated from the application. If the application is funded, this description, as is, will become public information. Therefore, do not include proprietary/confidential information. **DO NOT EXCEED THE SPACE PROVIDED.**

Modeling human genetic variation is critical to understanding the genetic basis of complex disease. The Human Genome Project has discovered millions of DNA sequence variants (single nucleotide polymorphisms or SNPs), and millions more may exist. As the coding for proteins takes place along chromosomes, SNP organization along each chromosome, the haplotype structure, will be most useful for discovering genetic variants associated with disease. Haplotype-based association studies are powerful procedures for detecting genetic influences on complex diseases. However, association tests of haplotype effects with unphased genotype data can be sensitive to estimates of haplotype frequencies even with family-based study designs and complete genotype information.

The broad objectives of this proposal focus on enhancing the arsenal of statistical methods researchers use to dissect genetic factors in complex diseases. Specifically, we propose to apply results from coarsened-data semiparametric efficient model theory to derive optimal tests and estimates of haplotype and haplotype interaction effects that are robust to haplotype frequencies using unphased, and possibly missing, genotype data. The data structures we will consider are motivated by those found in the _____ study and the _____ Study. In addition, we propose to apply these newly developed techniques to the _____ samples in fine mapping and candidate gene studies.

This research will form the core of a 5-year career development plan for Dr. _____ under the mentorship of three exceptional researchers, each with expertise that complements one another and represent the three areas addressed in this proposal: cardiology, genetics, and statistics. They propose a career development plan that combines didactic and practical training in genetics, cardiology, and genetic epidemiology with an ongoing research program within the unique research environment of _____ University. This career development plan will foster Dr. _____ development into an established independent quantitative research scientist with expertise in both methodology for dissecting genetic factors in complex disease and cardiovascular genetics.

PERFORMANCE SITE(S) *(organization, city, state)*

_____ University _____, _____

KEY PERSONNEL. See instructions. *Use continuation pages as needed* to provide the required information in the format shown below. Start with Principal Investigator. List all other key personnel in alphabetical order, last name first.

Name	Organization	Role on Project
_____	_____ University	Principal Investigator
_____	_____ University	Mentor
_____	_____ University	Co-Mentor
_____	_____ University	Co-Mentor
Advisory Committee		
_____	_____ University	Advisory Committee
_____	_____ University	Advisory Committee
_____	_____ University	Advisory Committee
_____	_____ University	Advisory Committee

Disclosure Permission Statement. Applicable to SBIR/STTR Only. See instructions. ☐ **Yes** ☐ **No**

RCA TOC Substitute Page Candidate *(Last, first, middle):* _____

Use this substitute page for the Table of Contents of Research Career Awards. The name of the candidate must be provided at the top of each printed page and each continuation page.

RESEARCH CAREER AWARD
TABLE OF CONTENTS
(Substitute Page)

Page Numbers

Note: Type density and size must conform to limits provided in the Specific Instructions.

*Include these items only when applicable.

CITIZENSHIIP

☒ U.S. citizen or noncitizen national ☐ Permanent resident of U.S. (If a permanent resident of the U.S., a notarized statement must be provided by the time of award.

RESOURCES - _____

Laboratory:
N/A

Clinical:
The _____ is a multidisciplinary group of faculty and staff from the Departments of _____. In addition, there are more than _____ research professionals who support the _____ (masters-level biostatisticians; masters-level programmers; analysts; data managers; nurses; technicians; and financial, editorial, and administrative staff). The _____ is located _____ with bus service to and from, allowing for quick and easy access to the clinical domain while providing a separate and distinct research environment, a combination that ideally supports the pursuit of academic medicine.

Animal:
N/A

Computer:
Computer resources currently available to serve the extensive computing and information-sharing needs in the _____ include over 1000 Pentium-class PC desktops linked by a fully-switched network to over 60 Microsoft Windows and Sun Solaris servers, which collectively provide about 4 TB of disk space. An enterprise backup system is implemented with a capacity over 6 TB and throughput of 40 MB/s. Databases are maintained with capabilities for point-in-time recovery. A 1-Gbps backbone Ethernet network with 100-Mbps to the desktop provides quick response and efficient data transfer. A secure FTP file server is used for transfer of data to and from remote sites. Internet access is restricted with a computer firewall to the _____. The computing resources also include networked laser printers, scanners for black-and-white and color scanning, and FAX services for delivery of documents from the desktop. SAS for Windows is installed on the network. _____ also maintains an 8-processor Sun Ultra Enterprise 4500 with the current versions of SAS, S-PLUS, and Oracle software. The Ultra E4500 has 8 GB of shared memory, an extensive array of RAID disk drives (over 1.5 TB) with throughputs of 40-100 MB/s.

Office:
Office space for all staff and support personnel in this project is located in the _____. The day-to-day operational activities of this project will be housed in the _____. This facility includes multiple conference rooms, adequate space for computer resources, and open work areas for ease of collaboration (in addition to individual office space). The combined office area for the _____ at this location is in excess of 160,000 square feet.

Other:
The _____ has numerous conference rooms (large and small), video-conferencing capabilities, a separate IT training room, and a lecture hall with advanced sound and presentation systems that include dual-screen projection. It also includes a library and many open work areas for ease of collaboration (in addition to individual office space). All of these are available as needed.

Major Equipment:
Each of the _____ floors occupied by _____ personnel at the _____ is equipped with several fax machines, many of them specific to ongoing studies and others specific to faculty members of functional groups. In addition, each floor is equipped with at least 3 high-volume, multiple-capability copiers, mailing and overnight service drop-off stations; and other assorted office equipment and supplies (heavy-duty staplers, electronic three-hole punchers, etc.).

Resources - _____ Laboratory and _____

Biomedical Laboratory: _____ laboratory currently includes one 750 ft^2 and one 250 ft^2 lab, two 100 ft^2 cold rooms, and one 100 ft^2 office as well as the additional office space described below. The laboratory is immediately adjacent to that of other investigators pursuing molecular studies in other systems. Also, two 500 ft^2 labs containing a dark room, common equipment items, and heavy equipment (freezers, ultracentrifuges) is shared with 1 other laboratory. No additional space is required for this project.

Clinical: All of the ascertainment will be out of the _____. This is a 58,000 ft^2 preventive medicine clinical and research facility which contains the 28,000 ft^2 _____ facility with a 800ft^2 indoor pool, 1/12 mile indoor and 1/6 mile outdoor running tracks; the 14,000 ft^2 _____ with the 1,000 ft^2 _____ Laboratory; and the 16,000 ft^2 _____ for Nutritional Studies with a cafeteria and feeding kitchen. The facility is staffed with 4 physicians and over 50 staff that include exercise physiologists, nurses, physical therapists, physician assistants and nutritionists. _____sees clinical patients at the _____. The facility supports over 400 patient visits each day for cardiac rehabilitation, aerobic exercise, nutrition classes, patient support groups, clinic visits, education classes and other treatment modalities.

The 14,000 square foot _____ at the _____, has 11 patient examination rooms, a fully equipped nurses' station (including emergency code team and defibrillator, pharmacology cabinet, first aid supplies), X-ray room, blood laboratory, and nuclear cardiology unit, 16-slice ultrafast CT scanner, DXA scanner and – as described above – an exercise physiology testing laboratory with treadmills and metabolic carts (ParvoMedics, MedGraphics and Sensormedics). The clinic is staffed by six physicians (including three cardiologists and a cardiology fellow), 12 exercise physiologists, 7 nurses, 2 physician assistants, and support staff. In addition to _____ staff, the clinic supports an array of multidisciplinary _____ physicians who have clinics in our facility including: cardiology, pulmonology, endocrinology, neurology, rheumatology, and general internal medicine clinics throughout each day. In addition, the clinic is conveniently located within close proximity of the fitness center and offices of both the principal investigator and study coordinator.

_____ controls over 1500 ft^2 of dedicated research space on the _____ campus for the research team.

Animal: Not applicable.

Computer: The _____ research unit contains multiple PC microcomputers and laser printers for word processing, data analysis, and graphics. Through the Internet, the laboratory personnel have access to multiple computer-based databases for nucleic acid and protein sequences. Various other computer facilities are available to the other investigators for the data collection and statistical analyses.

Office: The various investigators associated with this proposal have administrative and secretarial support services available to them for use in this study. Approximately 500 square feet of office space has been allocated for this project. Each office is computer ready with phone lines, work stations, and filing cabinets. Included is a large conference room for meetings with audio-visual capabilities. Office space includes a copy machine, fax machine, and two printers. The offices are located on the _____ campus within a one minute walk of both the clinic (exercise physiology laboratory) and the Fitness Center.

Other:

Major Equipment: In Dr. _____ laboratory, major equipment items include a Perkin-Elmer LS-3 fluorescence spectrometer; Perkin-Elmer Lamda 4B spectrophotometer; Baker B40-112 laminar flow chamber, 2 Queue Systems 2210 tissue culture incubators, 2 Revco ULT 12100 B-L-N Ultralow freezers; Kenmore lift top—20^o freezer; Queue System 4730 shaking incubator; Fisher bacterial culture incubator; Fisher table top oven; Virtis System I lyophilizer and spin-vac system; Dupont-Sorvall RC-80 ultracentrifuge; Beckman J2-21 Superspeed and J-6B Lowspeed centrifuges; assorted centrifuge rotors; Sorval 24S and Eppendorf 5414 microcentrifuges; Perkin Elmer Cetus DNA Thermal Cycler (PCR apparatus); ABI 7000 Taqman sequence detection system; pH meters, analytical scales, shaking water baths, Robbins hybridization chamber, numerous dry blocks and various other minor equipment items; electrophoresis power supplies and chambers by LKB, IBI, Pharmacia,

BRL, Biorad and Raven for DNA, RNA, and protein electrophoresis, transfers and other analytical work. The common resource facility provides common access by the laboratories of two investigators to: 2 New Brunswick floor shaking incubators, Beckman Optima XL-90 ultracentrifuge with 4 optional rotors, Forodyne MP-4 UV and visible light photographic documentation system, LKB 1251 bioluminometer; Beckman 388/166 HPLC; Beckman LS 6000IC scintillation counter; Brinkman PT3000 tissue homogenizer; Zenopure ultrapure distilled water; Kodak automated film processing unit.

RESOURCES – Center _____

Laboratory: In November 2002 the Center _____ moved to a new facility at _____ with over 55,000 square feet dedicated to _____ activities. Included in the new facility is a 5,000 square foot DNA banking facility, as well as an additional 8,000 square foot high-throughput genotyping facility. This facility unifies all of the _____ personnel, laboratories, offices, and equipment. The new facility includes office space, computer labs, and wet laboratories. Adequate clerical staff is available to all researchers. All personnel have access to common instrument resources and equipment through the Genetic Epidemiology Laboratory, the Genomic Screening Laboratory, and the Genomic Resources/DNABank Laboratory.

The _____ laboratories have all the facilities necessary for modern molecular and statistical genetics work, including all apparatus necessary for DNA sequencing and synthesis, pulse-field gel electrophonesis, PCR, standard electrophoresis, mutation analysis and genotyping. Available equipment is listed below.

Clinical: The new _____ facility includes a separate 4,000 square feet devoted to clinical research space, including one fully-equipped examination room and two interview rooms, one of which is equipped with an observation mirror and video capabilities. Patients and their families will have access to free, accessible, nearby parking for their visits.

Animal: N/A

Computer: Housed within the _____ is a dedicated computer local area network. The network consists of approximately 40 Sun Microsystems workstations running the Solaris operating system, approximately 150 personal computers running Windows 2000, and a Network Appliance Filer. Laboratory equipment such as the Hitachi FMBIO IIT Multi-View Scanners and the sequencers are networked to allow seamless transfer of data to the database. The Center also supports external connections via a terminal server. All machines are kept up-to-date with virus protection software and security patches. In addition, our systems reside behind the _____ Center firewall, providing an extra level of security.

_____, the database system for the Center _____, has been in development over the past 2 decades. The database, running Sybase 12.0, currently resides on a Sun Microsystems SunBlade 1000 workstation. A Sun Microsystems Sun Fire V880 4 processor system has been purchased and will soon be used to house the database. This system has a dedicated backup server running NetBackup to a L100 tape drive unit with a 100 tape cartridge capacity. Two T3 arrays are attached running hardware RAID 5 with a capacity of approximately 327 GB fully mirrored. Migration to Oracle database will be complete by _____. The database engine, Oracle 8.1.7, will support demographic, clinical, genotypic, mutation and sample information. An additional database instance will contain sample and aliquot laboratory tracking data. The LAPIS data management program is currently available on all platforms and allows the transfer of data between the _____ laboratory database and the a wide variety of genetic analysis program formats. The genetic analysis programs currently available in the Center _____ include: LINKAGE/FASTLINK, VITESSE, SLINK, SIMLINK, MAPMAKER, CRI-MAP, MENDEL, SAGE, HOMOG, MOMOZ, APM, WPC, SIMIDB, ASPEX, SIMEX, GENEHUNTER, GENEHUNTER-PLUS, CASPER, GASP, TDT, ALLEGRO, RELPAIR, and RELATIVE

Office: Within the 20,000 square feet, ample office space and computer support (see above) is available to the members of the Center _____

Other:

MAJOR EQUIPMENT: List the most important equipment items already available for the project, noting the location and pertinent capabilities of each.

- Two Beckman Instruments CEQ 2000 DNA Analysis System Capillary Electrophoresis Sequencers
- One Li-Cor 4200 LongRead™ DNA Sequencer (Two-Dye) for Automated Sequencing
- Three Transgenomic Wave™ DNA Fragment Analysis Systems (HPLC)
- One ABI 373A for Automated Sequencing

- One Molecular Dynamics FluoroImager SI
- Two Hitachi FMBIO II™ Multi-View Scanners
- One Packard Multiprobe EX Robotic Liquid Handling System with Barcode Capabilities
- One Packard Multiprobe 204DT Robotic Liquid Handling System
- Two Bio-Rad Chef and One LKB Pulsaphor units for pulse-field gel electrophoresis
- Twenty CBS Scientific Custom Electrophoresis Apparatus for Gel Electrophoresis
- Fourteen Techno MWG Primus-96HPL Well Thermal Cyclers
- Two Hybaid Multiblock PCR Systems with three each Gradient 0.2ml blocks per system
- Two Techne MW-2 96 well thermal cyclers
- Four Hybaid OmniGene® with Two Satellite blocks
- Four Hybaid Touchdown® Sub-Ambient Thermocyclers with Satellite Units
- One Beckman Coulter Allegra 25R Centrifuge
- One ABI Prism 7900 HT Sequence Detection System
- One LabRepCo MVE Liquid Nitrogen Freezer with Full Auto Control and Battery Backup
- Six MJ PTC-200 PCR units
- Four Perkin-Elmer Cetus microeppendorf tube DNA thermal cyclers
- One LKB Pulsaphor Unit for Pulse-Field Gel Electrophoresis
- Two BioRad Chef
- Eleven IBI baserunners for DNA sequencing
- One Robbins Scientific Hydra-96 Microdispenser
- Twelve IBI STS 45 sequencers
- Three Robbins Scientific Hybridization Ovens
- One Bellco Autoblot Hybridization Oven
- Two Bio-Rad 583 gel dryers
- Four Sorvall RC2B Centrifuges and One Sorvall RC5B Centrifuge
- One Beckman L8-5M Ultracentrifuge and One L2-65 Ultracentrifuge
- UV Stratalinker 1800
- One Gilford Spectrometer
- Two shaking Incubators
- Two Kodak RP X-omat Processors
- Four CO_2 Incubators
- One vacuum oven
- One Shimadzu UV-1201 Spectrophotometer
- One Lyophilizer
- Balances

- pH meters

Section II: Specialized Information

1. Introduction to Revised Application – N/A

2. Letters of Reference

3. The Candidate

A. Candidate's Background

Like many who receive a PhD in Statistics/Biostatistics, I began my academic career as a mathematics student, receiving both a bachelors and masters degree. Though I followed the "pure" math track, I enrolled in several courses in ordinary and partial differential equations and became interested in modeling physical phenomenon. This interest was further strengthened by my experience as a summer student at _____ where I was able to apply techniques I had learned to solve real world problems. During my first summer I developed a program for modeling heat dissipation in an anisotropic crystal being hit by a laser beam. This program was used to determine the safe operating limits of the crystal. My project for the second and third summers was to develop a mathematical model for a generally astigmatic laser beam (intensity profile is elliptical and twists as it propagates) which was used to correct a drive laser for a particle accelerator. The interaction between mathematical theory and scientific problems was deeply satisfying and made a lasting impression on me.

My interest in using differential equations to model real-world phenomenon also formed the bridge to my current career in statistics. While studying the differential equations that model predator/prey interactions, I came across literature describing this dynamic system using stochastic (i.e., probabilistic) differential equations. I was intrigued, not only by the technical challenges the probabilistic components created, but by the richness of the model and the unique biologic questions that could be addressed by it. I could, for example, quantify extinction probabilities. This quantity is out of the reach of the purely deterministic modeling approach. Because of this, I began studying probability theory, stochastic process theory, and statistical theory, leading to my current career path as a statistician.

A year of working as a statistician on the _____Study reaffirmed my plans of pursuing a career in statistics and I matriculated at _____. My dissertation work focused on multivariate random length data and its relationship to a unique type of missing data problem. One of the motivating examples for my dissertation work was in cardiovascular medicine where both the number of atherosclerotic lesions and the degree of occlusion of each lesion may inform on treatment effect. I advanced models to address problems like this. This also initiated my interest in cardiology.

During the final semester of my PhD work I was exposed to statistical genetics for the first time. During a CDC/ATSDR sponsored symposia on statistical methods Dr. _____ presented an invited lecture on statistical methods for assessing familial aggregation of disease. I was fascinated by the fundamental role probability theory played in the genetic model. The interaction between quantitative and biological models was remarkably similar to what I experienced when studying differential equations. This congruence excited me and I began to read all I could find on the subject.

After completing my dissertation, I joined the faculty at _____ and _____ in July of 2001. My work at _____ has included collaborating as a statistical investigator on several cardiovascular disease clinical trials. I have also become involved in genetic sub studies to cardiovascular clinical trials as well as gene mapping studies such as the _____ study[1]. Concurrently, I maintain an active research program developing statistical methodology for missing data generally[2;3] and missing data in family-based genetic association studies[4;5] in particular.

While conducting these activities I met and began collaborating with my proposed mentors. My primary mentor, Dr. _____, organized a working group oriented towards devising solutions for integrating and analyzing complex datasets composed of clinical, genotypic, gene/protein expression, and imaging data. Dr. _____ was also involved in this working group and has since worked with me on several projects, including a manuscript resulting from our collaboration at the genetics analysis workshop[6] and an editorial on pharmacogenetic analyses[7]. Dr. _____ and I have collaborated on applying coarsened-data semiparametric model theory to

problems in genetics. This work is featured in the "previous studies" segment of the research plan. The same theoretical framework (coarsened-data semiparametric model theory) forms the basis of the proposals found in the research plan.

During the past 2 years my interest in statistical genetics, cardiovascular disease, and cardiovascular genetics has grown and converged and I have made significant progress in understanding some of the issues involved in each. I recognize, however, that I need additional training to compensate for my lack of a biomedical background so that I may become an independent scientist capable of conducting research studies in the genetics of cardiovascular disease and developing new statistical methodology for such studies that is grounded in a clear understanding of the underlying biological system.

B. Career Goals and Objectives: Scientific Biography

My overarching career goal is to become an independent academic research scientist focused on both (1) developing statistical methodology designed to quantify biologically motivated hypotheses concerning the relationship between genetic variants and disease and (2) applying these techniques with the goal of uncovering genetic factors and understanding their role in cardiovascular disease. My goal is to have a balance of these two activities as each will motivate and provide direction for the other. In the near term, my goals are (1) to continue to build the skills that I will need to succeed in a research career, while (2) making significant contributions to quantifying and understanding the genetic basis of coronary artery disease. To accomplish these goals, I have developed a career plan that combines (1) didactic training in genetics, biology, and statistical genetics, (2) a mentored reading program in cardiology and cardiovascular genetics, and (3) ongoing research under the mentorship of successful statistical, cardiovascular, and genetics investigators. My plan is to develop and apply efficient yet robust methods for family-based haplotype association studies of coronary artery disease. This work is a natural extension of my previous work in family-based association studies and cardiovascular clinical trials.

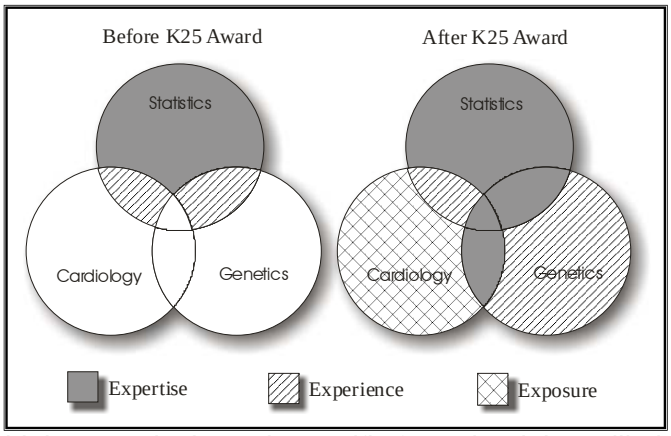

Figure 1: Areas of Expertise, Experience, and Exposure of Candidate in Areas of Research Relevant to this Application.

Importance of the K25 award in attaining career goals

This mentored quantitative research career development award (K25) will provide an ideal mechanism to obtain my career goals. First, the K25 will provide me with the resources for training in genetics and cardiology necessary to conduct independent research in cardiovascular genetics and to ensure that any statistical models developed are well anchored in the underlying biology. Only through specific formal training will I be able to acquire the background knowledge in these areas necessary to become an independent and biologically focused quantitative researcher. Secondly, the K25 will facilitate ongoing collaborations with established cardiovascular, statistical, and genetic epidemiology investigators allowing me to develop expertise in emerging fields of interest (statistical genetics and the genetics of cardiovascular disease). Finally, the K25 will provide the protected research and training time necessary for the development of my research with a focus on progression to independence as a researcher. At the conclusion of this K25 award, I will have (1) developed substantial theoretical and practical understanding of genetics, cardiology, and statistical genetics; (2) acquired expertise in the intersection of these three research areas; (3) gained experience in collaborating with veteran scientists; and (4) completed an in depth program of method development and application under the guidance of qualified mentors.

My previous research experience in statistics and statistical genetics as well as my collaborative experience in cardiovascular clinical trials provide an important foundation and fertile ground from which to

build a successful career. Figure 1 illustrates the expertise I currently possess as well as that which I will gain after successfully executing my K25 career development plan. I have been given full support from the _____, the _____, and the _____, and their commitment of the resources necessary for me to obtain these career goals (see letters section 5.b. from _____).

C. Career Development Activities during Award Period

The development of my academic research career requires that I acquire or solidify my knowledge in several areas. In order to do so, the primary components of my career development plan will involve the following:

1) Acquiring a solid foundation in genetics including molecular genetics, population genetics, and genetic epidemiology
2) Acquiring theoretical understanding and practical experience in cardiovascular medicine
3) Acquiring expertise in the genetics of coronary artery disease
4) Acquiring an understanding of genomic technologies
5) Enhancing expertise in statistical genetics
6) Planning, initiating, managing, completing, and presenting the research outlined in this application under the mentorship of, and in collaboration with, a group of highly successful investigators.

In addition, other experiences, such as honing my writing skills, developing mentoring abilities, and completing training in the responsible conduct of human research, will be necessary to complement my main objectives. Based on these components, I propose the following career development plan over the 5-year period covered by the K25 award.

Table 1: Timeline for K25	**Year 1**	**Year 2**	**Year 3**	**Year 4**	**Year 5**
Didactic Training (see details below)					
Complete core genetics and cell biology coursework					
Supplemental genetics coursework as required					
Complete coursework in Genomic technologies					
Supplemental coursework in statistical genetics					
Workshops/short courses in genetics and genetic epidemiology					
Practical Exposure to Cardiology (see details below)					
Shadowing in Cardiac Care Unit					
Shadowing in Cardiac Catheterization Lab					
Attend Cardiology Grand Rounds					
Mentored readings and Journal Clubs (see details below)					
Mentored readings in Cardiovascular Medicine					
Mentored readings in the genetics of coronary artery disease					
Journal club in coronary artery disease genetics					
Improved Statistical Methods for Haplotype/Disease Association Analyses (as detailed in section 6)					
1. Robust haplotype methods for complete unphased genotype data					
2. Robust haplotype methods with missing pedigree members					
3. Robust methods for haplotype interaction effects					

Table 1: Timeline for K25	**Year 1**	**Year 2**	**Year 3**	**Year 4**	**Year 5**
4. Computer software for novel haplotype methods					
5. Application of novel haplotype methods to _____ data					
Skills Development (see details below)					
Regular mentoring					
University writing course					
Training in responsible conduct of human research					
Collaboration with cardiovascular, statistical, and genetic epidemiology investigators					
Written manuscript production					
Local/national oral presentations					
Transition to mentoring others					
Follow-up Projects					
Project development					
Seek peer reviewed funding					

This career development combines didactic training, real-world exposure to cardiovascular medicine, mentored readings and journal clubs, research conferences/seminars, and national meetings to give me the necessary background in key scientific areas to become an independent and productive quantitative research scientist. My planned mechanism for developing a solid background in genetics, genomic technologies, and genetic epidemiology is mainly through formal coursework supplemented with seminars and national meetings. Designing a plan for acquiring an exposure to the fundamentals of cardiovascular medicine was more challenging as a formal course in cardiology does not exist. Therefore I have devised a multifaceted approach that combines formal coursework in cell biology, practical exposure to the realities of cardiovascular medicine by “shadowing” practicing cardiologists as they round, and a mentored reading schedule. I expound on the various components of this plan below.

Didactic Training:

I will take 25 credit hours of formal coursework in genetics, cell biology, genomic technology, and genetic epidemiology through _____ University. These courses will give me a firm foundation on which to build a successful career as a biologically focused quantitative researcher. All courses are taught each year at _____ . With this course schedule I will exceed the course requirements for the certificate in _____ . I have provided the name of each course and its role in my career development below (a more detailed description of courses is found in Appendix I).

Course	Role in Career Development
_____ . Principles of Genetics and Cell Biology I	This course will provide fundamentals of modern genetic theory, giving me a basis on which build in subsequent courses.
_____ . Principles of Genetics and Cell Biology II	This course and _____ will form the core of my biologic training. The treatment of the regulation and control of cellular processes will be especially useful in understanding current models of arteriosclerosis, inflammation, and cellular signaling processes in cardiology
_____ . Molecular Population Genetics	The source and quantification of genetic variation are necessary considerations when developing methodology to detect certain variations and their relationship to disease. This course will provide a theoretical understanding of how genetic variation arises in a population and changes over time.
_____ . Statistical Genetics	Although I have done substantial reading in this area, this course will provide a formal overview of statistical genetics, filling in gaps and enhancing my current understanding.
_____ . Genome Informatics & Sequence Analysis	This course will provide a solid foundation in the basic principles and methods of genome informatics and sequence analysis.
_____ . Genome Technologies	This course will build upon _____ by focusing on the technology used to generate sequence and expression data. Understanding the origin and limitations of the data is necessary for relevant method development.
_____ . Gene Expression Analysis	Gene expression studies are becoming a more important facet of understanding genetic factors in the development of human diseases. This course will allow me to understand, interpret, and conduct such studies.
_____ . Introduction to Proteomics	Ultimately genes exert their influence on human disease through their role in synthesizing and modifying proteins. This course will allow me to understand and interpret these studies. A solid understanding of proteomic, gene expression, gene association, and gene mapping studies will allow me to successfully integrate many sources of complementary information in understanding genetic influences in cardiovascular disease.
_____ . Responsible Conduct of Research	Training in the responsible conduct of research is an important part of any career development plan. This course is especially relevant to my research focus as it emphasizes issues of particular importance when using genetic information in research.

Workshops/short courses: In addition to the coursework above, I will participate in the educational resources provided by the National Heart Lung and Blood Institutes Program for Genomic Applications (PGA). I have provided a description of each short course and its role in my career development below.

- *Genetic Approaches to Complex Heart, Lung, and Blood Diseases (Jackson Lab PGA educational program)*
  Description: This short course focuses on genetic approaches to complex heart, lung and blood diseases and includes techniques used to analyze human population data and data derived from genetic experiments done in the laboratory mouse.
  Role in Career Development: Provides opportunity to train, via formal lectures, discussion groups, demonstrations and tutorials, in the application of molecular biology and genetics to the analysis of complex diseases such as coronary disease. The level of interaction with nationally recognized experts and fellow students with similar interests provides a unique and varied opportunity to learn cutting edge techniques and form future collaborative relationships.

- *Short Course on Gene Microarray Development and Analysis: Approaches to Heart, Lung, Blood and Sleep Disorders (Jackson Lab PGA educational program)*
  Description: This course will focus primarily on gene expression analysis in a variety of mouse models relevant to the study of heart, lung, blood and associated sleep disorders.
  Role in Career Development: This course will be complementary to _____ with an emphasis on chip development with applications to heart, lung, blood and sleep disorders. It will provide significant insight into developing and executing gene expression studies in cardiovascular medicine.

- *Bioinformatics Tools for Comparative Genomics (Berkeley PGA educational program)*
  Description: The course presents an introduction to basic sequence analysis tools and databases, skeptical usage of annotated genome information, mining and finding additional information on your favorite genes, and using the computational tools, databases, and technologies developed by the NHLBI program in genomic applications.
  Role in Career Development: This course will be complementary to _____ with a focus on using the resources developed by the PGA. This will aid in using these resources in executing future studies of the role of genetic variants in cardiovascular disease.

Practical Exposure to Cardiology:

- *Shadowing in cardiac care unit and catherization lab:* In order to better ground my didactic training in the practical realities of cardiovascular medicine, I propose to shadow cardiologists as they round on the cardiac care unit (CCU) and observe interventionists in the catherization laboratory. I have the full support of Dr. _____, co-director of the CCU, and Dr _____, director of interventional cardiology research (see support letters from Drs. _____ and _____ in section 5.b.). I have "bracketed" this experience around my didactic and mentored reading experiences so that this shadowing can both motivate my coursework/readings as well as make the theoretical knowledge I acquire more concrete.

Mentored reading and Journal Clubs:

- *Mentored reading in cardiology.* Under the direction of Dr. _____ I will survey the literature relevant to developing an understanding of the key current issues in cardiology. This survey will include original research articles, review articles, and book chapters. We will discuss my reading progress during weekly mentoring meetings. It is expected that there will be significant interaction between the direction of these readings and the topics covered during my readings of the genetics of coronary artery disease literature.
  Role in career development: These readings will develop the core of my cardiology knowledge.
- *Mentored reading in the genetics of coronary artery disease.* Under the direction of Drs. _____ and _____, I will review the current and historical literature related to the genetics of coronary artery disease.
  Role in career development: This exercise is a necessary first step in mapping out the landscape of this research area.
- *Journal club on the genetics of coronary artery disease***.** I will initiate a journal club of the genetics of coronary artery disease in which we will read and discuss current developments in the understanding of this area of research. The club will include myself, Drs. _____ and _____, other faculty members and fellows, and others interested in recent work in the genetics of CAD.
  Role in career development: This experience will further my understanding of the direction this field is taking.
- *Statisical Genetics/ Genetic Epidemiology Journal Club.* The _____ has an ongoing journal club focused on reviewing recent papers advancing new methodology for the statistical analysis of complex human diseases such as CAD. I will be an active participant in the club for the duration of the K25 career development award.
  Role in career development: This club will keep me abreast of current developments in statistical genetics.

Seminars, Conferences, and National Meetings:

The educational activities outlined above will be supplemented by a regular schedule of research conferences, seminars, and national meetings at _____ and throughout the world.

Seminars and conferences:

Seminar or Conference	**Description**
Weekly _____ Seminars	Involves scientists from around the world presenting both scientific and methodologic work on the role of genetics in human disease
Monthly Division of Cardiology research conferences	Includes the presentation of basic, translational, and patient oriented cardiovascular research topics.
Weekly Cardiology Grand Rounds	Incorporates core curriculum lectures along with presentations of new results and topics from leading investigators from within _____ and elsewhere (See Appendix II for 2003-2004 lecture sequence).
Weekly _____ research conferences	Involves leading clinical investigators from both _____ and elsewhere in a vigorous discussion of both scientific and methodologic issues in human subjects research.
Weekly seminars in statistics	Numerous groups in the greater _____ area, including: _____, _____, _____, and the _____. Though I will not be able to attend all of these seminars, with so many opportunities I will be able to choose topics of particular relevance or interest.

I will participate regularly in this sequence of conferences, and will present the progress of the research outlined in this proposal at one of the above venues at least once annually.
Role in career development: Attending seminars and conferences adds two important components to my career development: (1) it exposes me to current developments in my field of interest or closely aligned fields, and (2) it allows me to meet and form relationships with other scientists at my institution and beyond engaged in similar research. Presenting at seminars and conferences will also enhance my communication skills. Cardiology Grand Rounds will be particularly important to developing theoretical cardiology knowledge as the “core curriculum” presentations will provide a broad yet detailed overview of modern cardiovascular medicine.

National meetings:

Meeting	**Description**
American Society for Human Genetics (ASHG)	The largest human genetics meeting with sections representing the spectrum of applied disease and methodologic research (Annual).
Genetic Analysis Workshop (GAW)	Participants analyze a common dataset using innovative methodologies and share their discovery experiences (Bi-Annual).
International Genetic Epidemiology Society (IGES)	A smaller meeting focusing on methodology for genetic epidemiology. It is often scheduled immediately proceeding or following another meeting such as ASHG or GAW (Annual).
American Heart Association (AHA)	Large cardiovascular medicine meeting with a variety of sub disciplines represented including cardiovascular genetics (Annual).
American College of Cardiology (ACC)	Cardiology meeting with sections devoted to genetics (Annual).
Joint Statistical Meetings (JSM)	Largest statistics meeting with sections on statistical genetics (Annual).

As I am developing a career encompassing the intersection of three disciplines, my intention is to attend at least one genetics meeting, one cardiology meeting, and one statistics meeting each year.
Role in career development: Attending national meetings is an important component of my career development, both because it exposes me to current developments in my field of interest and because it allows me to meet and form relationships with other scientists throughout the world engaged in similar research.

Training in the responsible conduct of research:

Training in the responsible conduct of research is an important part of my career development plan. I will complete _____ (Responsible Conduct of Research). This course covers issues concerning the ethical conduct of research generally but also focuses particular attention on special issues faced by researchers collecting genetic information. Ongoing mentoring will also focus on the responsible conduct of research as it applies to the specific project outlined in this proposal. In addition, I will complete two of the _____ University Medical Center Research Ethics Program's human subjects training modules each year.

Mentoring:

The development of my academic research career requires that I further my knowledge in genetics, cardiology, and statistics with a focus on the intersection of these three disciplines. During the course of this award, my career development will be significantly enhanced by the involvement of three exceptional mentors, each with expertise that complements one another and represents my three areas of interest: statistics, cardiology, and genetics. My primary mentor, Dr. _____ is a cardiologist focusing on the genetics of cardiovascular disease and the primary investigator of the _____ study. Dr._____, one of my two co-mentors, is a genetic epidemiologist with a PhD in statistical genetics. Dr. _____ primary research area is in the genetics of cardiovascular disease. Dr. _____ also is the primary investigator on an NIH analysis grant for the _____study. Dr. _____ the other of my co-mentors, is a statistician who has done fundamental work in survival analysis, clinical trials, and coarsened data semiparametric estimation. I have illustrated each of my mentor's core areas of expertise and experience relevant to my career development in Figure 2. Note that collectively they possess expertise in each of the three relevant research areas. Each of these mentors have long records of accomplishment in their respective areas of expertise as well as extensive mentoring experience. I will meet with each on a regular basis. The exact frequency of meetings will vary by mentor depending on where I am in my development process (and therefore whose expertise and guidance is needed most) from once each week to every other week.

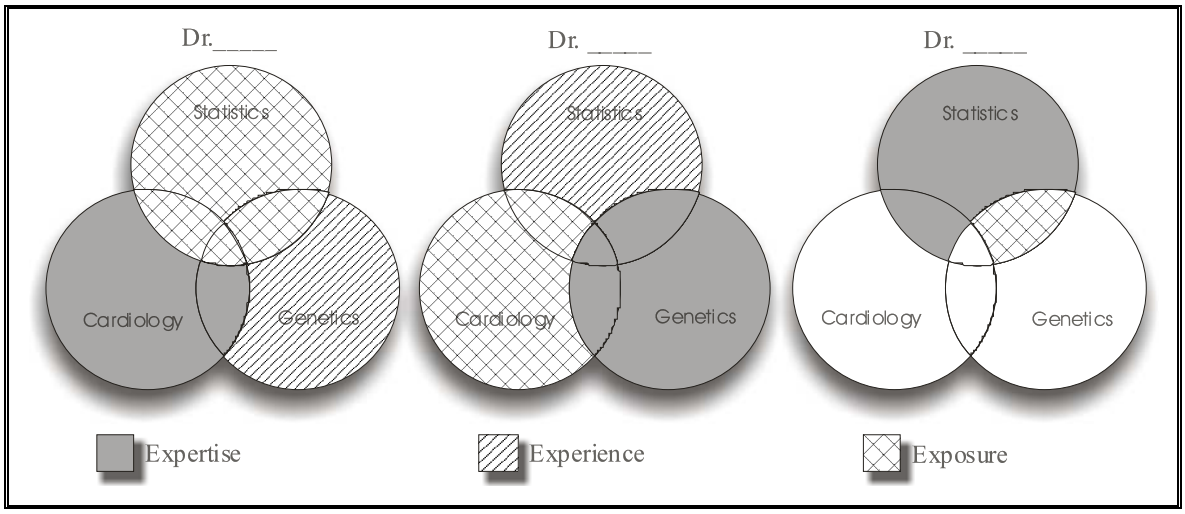

Figure 2: Areas of Expertise, Experience, and Exposure of Mentors in Areas of Research Relevant to this Application.

Presentation and Writing Skills:

An important component of developing into a successful independent researcher is the ability to communicate ideas and scientific results, both orally and written. Developing skills in each of these areas will be an important component of my career development plan. In order to sharpen my written communication skills I will take the _____ writing course taught by Professor of the Practice of Rhetoric and Director of the _____ University writing program, Dr. _____. Throughout the duration of this award I will continue to present my research in local, regional, and national settings. This will include both local, seminars and conferences, and national meeting forums. In addition, I will continue to develop manuscripts for submission to peer reviewed journals. The number of papers resulting from research in this proposal will vary according to where I am in my career development. One to two papers per year will be considered reasonable productivity at the beginning of this program, while three or more papers per year may be considered reasonable towards the end of the program.

Protection of Time:

Throughout the duration of the K25 award, I will devote at least _____ to the research proposal and career development plan outlined in this application. Because of the increased demand on my time during the early phases of the career development plan, I will devote _____ for the first 2 years. My salary during this time will be covered both by the K25 award (_____) and institutional funds (_____, see Dr. _____'s letter in section 5.b.)

Advisory Committee:

As part of the mentoring process associated with this career development award, I have identified a committee of successful statisticians, cardiologists, and genetic epidemiologists. In addition to my mentors, Drs. _____ (chair), _____, and _____, this committee will also include Dr. _____, Associate Vice Chancellor of Clinical Research and Director of the _____. Dr. _____ is a leading figure in cardiovascular clinical trials with a wealth of research and mentoring experience (Dr. _____'s letter of commitment is found in section 5.b.). I will submit a yearly progress report to the committee and meet with them on a quarterly basis. During these meetings the committee will evaluate my progress both in terms of the research plan highlighted in section 6 and in overall career development. The committee will make recommendations regarding changes in my career development and research plan as needed.

4. Statements By Sponsor(s), Consultant(s), And Collaborator(s)

Sponsors/Mentors: [Information redacted]

5. Environment and Institutional Commitment to Candidate

A. Description of Institutional Environment

The educational, research, and intellectual environment at _____ University is unique and ideally suited for the proposed career development and research plans. _____ University is a vibrant, interactive academic community where ideas are actively discussed. There is a spirit of support and collaboration throughout the different components of the University that provides the opportunity for multi-disciplinary research. In fact, throughout the Institution there is an emphasis on fostering interdisciplinary research at the interface of historically distinct fields. This is evidenced by _____ University's recent 200 million dollar investment establishing the _____ . This institute is comprised of several research centers which together form an integrated approach to advancing genomic knowledge and techniques while, at the same time, addressing ethical and policy issues that will necessarily arise with such advances. There are several reasons why the environment created by the _____ will advance the career development plan and future academic career of the candidate. First, two centers within the _____—the _____ and the _____—offer a joint program in Bioinformatics and Genome Technology. Courses from this program factor significantly in the didactic career development experience of Dr._____ . Second, another center within the_____, the _____ is a leading organization in mapping the genetic basis of human disease and plays a leading role in the execution of the _____ and the pending _____ study. Dr. _____, a co-mentor on the current grant, is a member of the _____ and their senior genetic epidemiologist with a focus on cardiovascular genetics. Finally, cardiovascular genomics is one of the 4 "areas of excellence" identified by the _____ for particular emphasis and additional financial support. Thus the career path mapped out in this application is in complete harmony with the larger institutional vision, ensuring that _____ will continue to be an environment rich in stimulating academic and collaborative opportunities conducive to the continued career development of Dr. _____.

The physical resources of the organizations directly involved in the proposal—(1) the _____, (2) the _____, and (3) the_____—are described in section I.7 (resources). In addition to these physical resources, their intellectual environment is conducive to the development of an academic career as is outlined _____.

[Information redacted.]

In summary, the educational, research, and intellectual environment at _____ University provides a highly compelling setting within which to develop an independent interdisciplinary research career. Moreover, the institutions strong interdisciplinary orientation and cardiovascular genomics focus make _____ University the ideal environment for the specific career goals of Dr. _____.

B. Institutional Commitment to Candidate's Research Career Development

The institution is committed to protecting my time so that I may devote _____ for the first two years and _____ thereafter to the research described in this proposal and will reduce my teaching and administrative responsibilities as needed to meet that goal. Additionally, the institution is committed to ensuring that the resources are available for the proposed research to succeed. The following letters from Drs. _____, _____, and _____ speak to commitment at the institute, center, and departmental levels. The letters from Drs. _____ and _____ speak to program commitment to specific educational components of the proposed career development plan. Please see the letters, which follow, from: (Redacted)

6. Research Plan

A. Statement of Hypothesis and Specific Aims

Modeling human genetic variation is critical to understanding the genetic basis of complex disease. The Human Genome Project has discovered millions of DNA sequence variants (single nucleotide polymorphisms or SNPs), and millions more may exist. As the coding for proteins takes place along chromosomes, SNP organization along each chromosome, the haplotype structure, will be most useful, almost essential, for discovering genetic variants associated with disease.

Haplotype-based association studies can be powerful procedures for detecting genetic influences on complex diseases. However, association tests of haplotype effects with unphased genotype data can be extremely sensitive to estimates of haplotype frequencies even with family-based study designs and complete genotype information.

We hypothesize that robust haplotype-based methods will be important, even essential, tools for uncovering genetic components of complex disease, and that coarsened-data semiparametric model theory will provide an appropriate mathematical structure for deriving improved haplotype-based procedures.

The specific aims of this proposal will address these hypotheses and enhance the arsenal of statistical methods researchers use to dissect genetic factors in complex diseases. We propose to apply results from coarsened-data semiparametric efficient model theory to derive optimal tests and estimates of haplotype effects that are robust to haplotype frequencies using unphased genotype data. The data structures we will consider are motivated by those found in the _____ study and the pending _____ Study. In addition, we propose to apply these newly developed techniques to the _____ samples in fine mapping and candidate gene studies.

SPECIFIC AIM 1*: Develop statistical methodology for estimating and testing haplotype effects on disease risk that are robust to haplotype frequency misspecification using unphased but complete genotype data.* Using results from coarsened-data semiparametric model theory[8-14], we will develop optimal (i.e., most powerful) tests and estimators of the effect of haplotypes on disease that are robust to incorrect estimates of haplotype frequencies. We will first focus on the simplest nuclear family structure: a pedigree consisting of one affected proband and their parents. We will then move progressively to consider more and more complicated pedigree structures culminating in multiple generations of affected/unaffected individuals with complete genotype data and fully observed founders. Numerical examples and simulation studies will be used to assess the feasibility and performance of the developed methodology.

SPECIFIC AIM 2*: Develop statistical methodology for estimating and testing haplotype effects in cardiovascular disease that are robust to haplotype frequency misspecification using unphased and sometimes missing genotype data.* We will extend the methods developed in Specific Aim 1 to consider

different types of missing genotype data. This will include genotype data that are missing at a particular locus (perhaps due to the failure of an assay), as well as genotype data that are missing because an individual in the pedigree was not observed. The resulting range of pedigree structures considered in Specific Aims 1 and 2 reflects the spectrum of family types expected in the _____ and _____ Studies. Once again, numerical examples and simulation studies will be used to assess the feasibility and performance of the developed methodology.

SPECIFIC AIM 3*: Develop statistical methodology for estimating and testing haplotype/haplotype and haplotype/environmental interaction effects using unphased genotype data.* We will extend the methods developed in Specific Aims 1 and 2 to include haplotype/haplotype and/or haplotype/environment interaction effects. We will develop models that stress biologic plausibility and interpretability over mathematical convenience. Examples of such models exist for genotype level effects but have yet to be applied to haplotypes with unphased and possibly missing genotype data. Numerical examples and simulation studies will be used to assess the feasibility and performance of the developed methodology.

SPECIFIC AIM 4*: Develop software tools for implementing novel methodology developed in Specific Aims 1-3 and make them available to the general research community.* We will develop user-friendly computer software to facilitate dissemination of our statistical methodology for use by research scientists attempting to detect genetic variants involved in complex diseases. We will make our software and corresponding documentation available on the web.

SPECIFIC AIM 5*: Apply methodologies developed in Specific Aims 1-3 to SNP haplotypes in the _____ sample.* We will use the methods developed in Specific Aims 1-3 to estimate and test haplotype effects in the _____ sample. We hypothesize that these new methods will substantially enhance our ability to detect genetic variants that predispose individuals to cardiovascular disease.

B. Background, Significance, and Rationale

Overview of Coronary Artery Disease: Coronary artery disease (CAD) is the leading cause of death in the developed world. In the United States, more than 12.9 million individuals seek treatment for CAD each year resulting in medical costs in excess of $129 billion annually[15]. More than 1 million Americans experience CAD events each year with 515,000 resulting in death[15]. Identifying genetic variants associated with increased risk of CAD has the potential of translating into an improved prognosis for millions of individuals both by allowing those identified as genetically predisposed to make risk reducing lifestyle changes as well as by providing potential targets for novel therapies. Unfortunately, CAD is a very complex disease with a multitude of genetic and environmental influences and interactions. This complexity has made identifying genes that lead to increased risk and pathogenesis of CAD, with its plethora of phenotypic clinical presentations, very difficult, often leading to inconclusive or conflicting results.

The_____ and _____ studies: The Genetics _____ study is one of the largest family-based studies attempting to disentangle these many effects in an effort to localize CAD genes[1]. Dr._____, my primary mentor, and Dr. _____, one of my co-mentors, are both involved in this study. Dr._____ is the principal investigator while Dr _____ is the study's genetic epidemiologist and also the PI of an NIH grant to analyze the _____ database. The _____ study focuses on families with individuals who have early onset CAD (EOCAD). In _____, EOCAD is defined as having acute coronary syndrome (unstable angina or myocardial infarction), a revascularization procedure, or a positive functional imaging study at or before age 50 in men and age 55 in women. There are two reasons why focusing on an early onset sample may improve the chances of localizing CAD genes: (1) sampled individuals may be more similar in terms of their exposures and the underlying etiology of the disease, and (2) the genetic effect may be stronger in families with EOCAD.

The _____ study identified 933 nuclear families in the US and Europe with at least two siblings with early onset cardiovascular disease. Most of the families sampled consisted of an affected sibling pair though more extensive pedigrees were collected. Recently an initial genome-wide linkage scan has been conducted on 491 affected sibling pairs from 438 families and has identified 5 genomic regions with increased linkage signals and one region in particular with a strong and consistent signal across all a priori identified subgroups. However, due to the relatively sparse spread of microsatellite markers used (395 throughout the genome) the width of these peaks was quite large making a study of local candidate genes or haplotypes premature. The _____ investigators are currently following up with fine mapping in regions of interest in order to narrow their linkage peaks and arrive at a reasonable list of local candidate genes for further study. In addition, they are proposing to expand the ascertainment of _____ pedigrees to include the offspring, all siblings (including unaffecteds), parents, and spouses of affecteds in the original _____ sample. When completed, this will constitute one of the largest family-based samples of EOCAD to date. The wealth of pedigree/genetic information is complemented by the richness of the clinical data. These data include intermediate risk phenotypes such as lipoprotein levels, HbA_1C, C-reactive protein, and anthropomorphic measures, as well as environmental risk factors such as physical activity and diet.

This study provides the motivation and vehicle for my research proposal. As the _____ investigators refine identified linkage regions, there will be a need for powerful and unbiased methods for estimating and testing the effect of specific genetic variants on cardiovascular and intermediate risk phenotypes (see Specific Aim 5). This research proposal focuses on addressing this need by developing, and then applying to the _____ dataset, optimal haplotype-based association techniques that are robust to misspecification of haplotype frequencies.

The Promise of Haplotypes: Since the 90's, there has been a great deal of interest in single nucleotide polymorphism (SNP) based association studies. Such studies are reported to be more powerful than linkage-based methods for identifying genetic variants, especially when the variant has only a moderate effect on the disease phenotype[16]. Much of this excitement, however, is moderated by the fact that methods for detecting disease/SNP association rely on the existence of linkage disequilibrium between the SNP being tested and the actual disease locus. Thus, single locus tests may have limited power to identify genetic variants that influence complex disease risk.

Since haplotype-based methods incorporate linkage disequilibrium information from many markers, these methods should be more powerful than traditional linkage disequilibrium methods that focus on a single SNP. Haplotype-based methods also have the promise of being able to identify unique segments of DNA containing sequences predisposing individuals to disease. In addition, when multiple alleles at a single disease loci influence disease susceptibility, single marker tests can be severely under-powered relative to haplotype-based association methods[17]. Haplotypes are also useful when multiple variants on the same chromosome interact to cause disease, as haplotype methods allow for the joint effect of these multiple variants while single locus SNP methods do not.

The major problem of conducting haplotype studies is that the marker genotype data (often SNPs) available is unphased, resulting in haplotype ambiguity. For example, if four SNPs each with allele frequencies of 50% are in linkage equilibrium, 69% of all marker data will not allow haplotype reconstruction[18]. Thus one cannot simply combine the SNP loci into a large locus and apply single locus methods.

Statistical Methodology for Haplotype Analyses: The promise of haplotypes has spawned the development of a number of statistical approaches to deal with the ambiguous reconstruction of haplotypes from unphased genotype data. Most of these methods treat the haplotypes as missing data and apply the expectation-maximization (EM) algorithm[19] to infer haplotype frequencies by assuming these frequencies are in Hardy-Weinberg equilibrium (HWE)[20-22]. This assumption results in estimates and tests that are sensitive to the genetic structure of the sampled population, potentially leading to biased estimates and incorrect inferences. One of the key reasons for collecting family-based data is the fact that, in the full data case, methods are available that are invariant to any substructure that may exist in the population. Therefore methods that are sensitive to these effects in the context of family-based data are quite undesirable. Clayton[23] has proposed a

score test which can be used with haplotypes and appears to be less sensitive to the HWE assumption (which it still assumes) than competing likelihood-based approaches[24]. Clayton's method and the supporting software he provided offered a significant new tool for dissecting complex disease. Unfortunately, although intuitive, Clayton's approach is somewhat ad hoc and while it appears to do better in simulations than competing approaches, bias is still apparent and its general properties are unknown. In addition, Clayton's score test has been shown, when using affected sib-pairs, to yield inflated type one error rates when testing for disease/marker association in the presence of linkage[25]. Finally, though the score testing approach assesses the overall association between haplotypes and disease, it does not provide for the estimation and inference of specific haplotypes or haplotype features. Thus there is a need for robust methods for haplotype-based association analyses that are based on a firm theoretical foundation so that optimal procedures for estimating and testing genetic variants can be derived. We propose to address this need in the current research plan.

Coarsened Data Semiparametric Efficient Models: In order to develop methods that are robust to misspecified haplotype frequencies, we will draw upon the framework provided by coarsened data semiparametric efficient model theory. Semiparametric models are those in which one is willing to specify some features of the data parametrically, but are unwilling to assume anything about other features[8;9]. In a semiparametric approach to a family-based genotype association study, the conditional distribution of offspring genotype given parental genotype is specified, while no assumptions are made about the distribution of the parental genotype. When genotype data are complete, inference can be based on the conditional distribution of offspring genotype given parental genotype. Tests and parameter estimates based on this conditional distribution will be robust to the distribution of parent genotypes which is not specified[26-28]. However, when parental genotype is missing this approach breaks down and one must again consider the distribution of parental genotype. As mentioned above, most approaches to this problem fully parameterize the observed data likelihood and either use the EM algorithm[29] or directly compute the observed data score function[23]. An alternative approach, which would not introduce bias due to misspecification of the distribution of the parental genotype, would be to derive a score function that has zero expectation regardless of the distribution of the parental genotype. This approach was implemented by Rabinowitz[30] in the context of missing parental genotype data. Rabinowitz developed a score function that, under the null hypothesis, has zero expectation regardless of the distribution of the parental genotypes. He also proved that, in a neighborhood of the true nuisance parameter (distribution of parental genotype) value, this score function had the smallest variance among all score functions with this robustness property. Recently, Whittemore[31] has extended Rabinowitz's result so that the score function can be used in parameter estimation in addition to testing. Rabinowitz's approach is quite innovative and has the potential to significantly improve family-based association methods. However, neither Rabinowitz nor Whittemore explicitily used formal semiparametric estimation theory in presenting their estimators or in proving their optimality. As a result, Whittemore's optimality proof requires assumptions that are both unrealistic and unnecessary. In addition, neither Rabinowitz nor Whittemore considered haplotypes. In this research plan, we will use formal coarsened-data semiparametric efficient model theory to derive optimal tests and estimates of haplotype and haplotype interaction effects that are robust to haplotype frequencies using unphased genotype data.

C. Preliminary Studies and Any Results

Efficient Semiparametric Modeling Approach to Missing Parental Data in Case-Parent Designs. Recently, Dr._____ and I have used the semiparametric model framework to develop an estimating function for estimating and testing the effect of genotype on disease in a case-parent design with missing parental genotype data. This was in an attempt to understand and extend the Rabinowitz result and was prior to our learning about the still-in-press Whittemore paper. However, the framework and general argument here forms the basis for our approach to addressing Specific Aims 1-3. It is also a fundamentally different and, we feel, more powerful argument than those given by either Rabinowitz or Whittemore. The power of the argument derives from the machinery of semiparametric model theory which allows proofs that are more general while at the same time giving insight into how related problems may be approached. Such an argument forms the basis

for our proposed approach to Specific Aims 1-3. Therefore, giving an overview of this argument in the missing parental data case will be valuable in understanding what follows. We present an outline of our argument and the resulting estimating function below.

Assume a case-parent sampling design where individuals with disease or trait of interest are sampled and where possible their parents are also recruited. At a locus of interest let the proband genotype be denoted G_o and let G_p denote the parents', possibly missing, genotype. Let G_p^o be the observed portion of G_p (realizations are denoted by g_o, g_p, g_p^o). If we assume that the parental genotype information is missing at random (assumed in both the Rabinowitz and Whittemore approaches) we can write the *i*th family's contribution to the observed data likelihood as

$$\mathcal{L}_i^o = \sum_{g_p \in \mathcal{P}\left(g_p^o\right)} \Pr\left(G_{o,i} = g_o \mid G_{p,i} = g_p; \gamma\right) \Pr\left(G_{p,i} = g_p; \eta\right),$$

where γ (*q*-dimensional) are the parameters of interest, η (*r*-dimensional) are nuisance parameters, and $\mathcal{P}\left(g_p^o\right)$ is the set of all g_p consistent with g_p^o. Note that both probabilities involved in $\mathcal{L}_i^o$ are implicitly conditioned on the proband being diseased. We can compute score functions by taking the log of $\mathcal{L}_i^o$ (denoted by l_i^o) and differentiating with respect to γ and η, i.e.,

$$\frac{\partial l_i^o}{\partial \gamma} = E\left[\frac{\partial}{\partial \gamma} \Pr\left(G_{o,i} = g_o \mid G_{p,i} = g_p; \gamma\right) \mid g_o, g_p^o\right] = E\left[S_{\gamma,i} \mid g_o, g_p^o\right]$$

and

$$\frac{\partial l_i^o}{\partial \eta} = E\left[\frac{\partial}{\partial \eta} \Pr\left(G_{p,i} = g_p; \eta\right) \mid g_o, g_p^o\right] = E\left[S_{\eta,i} \mid g_o, g_p^o\right],$$

where $S_{\gamma,i}$ and $S_{\eta,i}$ are full data score functions. We will suppress the family index *i* in what follows.

The semiparametic model framework of Bickel et al.[8] treats these score vectors as geometric objects within a Hilbert space. Though the mathematical details will not be presented here, the basic idea is to create an estimating function for γ that is orthogonal to $E\left[S_\eta \mid g_o, g_p^o\right]$ and as a result is insensitive to the distribution of G_p. This is attained by projecting $E\left[S_\gamma \mid g_o, g_p^o\right]$ onto the linear space spanned by $E\left[S_\eta \mid g_o, g_p^o\right]$ (also called the nuisance tangent space, Λ_η) and subtracting the resulting projection from $E\left[S_\gamma \mid g_o, g_p^o\right]$.

In order to give the estimating function resulting from this projection argument, we need to define some additional notation. Let $\tau\left(g_o, g_p^o\right)$ be the position of g_o, g_p^o in a lexigraphical list of all possible combinations of offspring and observed parent genotype data given the observed pattern of missing data. Similarly, define $\tau\left(g_p\right)$ for all possible combinations of parent genotype pairs. Assume that $\tau\left(g_o, g_p^o\right)$ takes on values from 1 to *k* and $\tau\left(g_p\right)$ takes on values from 1 to *s.* Note that *k* is a function of the observed pattern of missing data. Define the following matrices: *X,* a $k \times s$ matrix with *j,k*th entry given by $\Pr\left(\tau\left(G_o, G_p^o\right) = j \mid \tau\left(g_p\right) = k\right)$; *W,* a $s \times s$ diagonal matrix with *j*th diagonal entry given by $\Pr\left(\tau\left(G_o, G_p^o\right) = j\right)$; and *Y, a* $k \times q$ matrix with *j*th row given by $E\left[S_\gamma \mid \tau\left(g_o, g_p^o\right) = j\right]$.

Result: We can show (see appendix III for derivation summary) that the orthogonal estimating function *U* resulting from projecting off the component of $E\left[S_\gamma \mid g_o, g_p^o\right]$ in the direction of the nuisance tangent space is

$$U = Y - W^{-1}\left(XW^{-1}X'\right)^{-} X'Y, \qquad (1)$$

(see Figure 3). Note that this defines the estimating function for all g_o, g_p^o. The variance of *U* can be computed via a robust, sandwich-like, estimator[32].

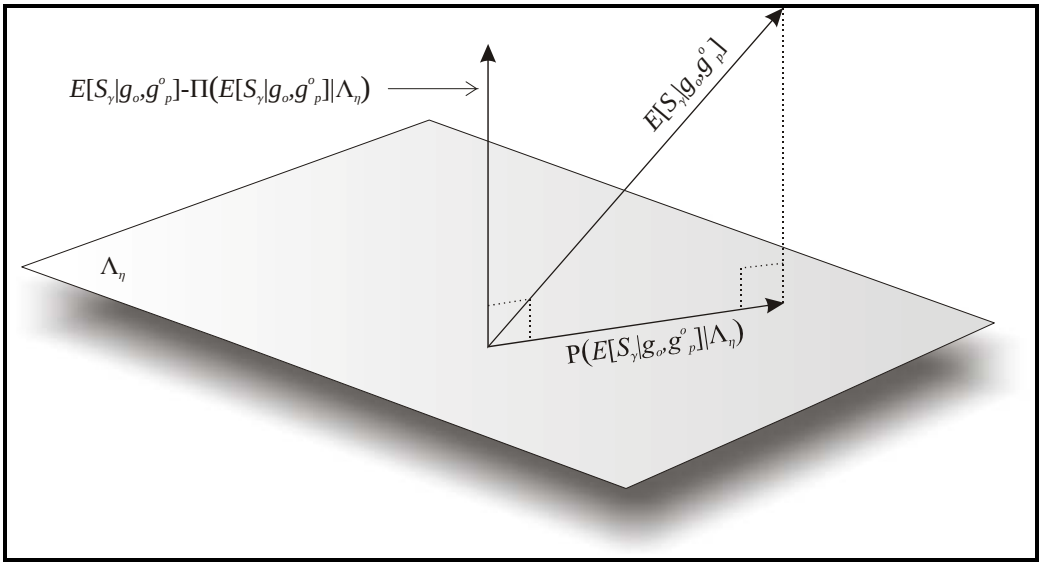

Figure 3: Geometric representation of the orthogonalization of *Y*.

Robustness: We can show (see appendix III for proof) that the estimating function (1) is robust to misspecification of the distribution of parental haplotypes. Therefore, regardless of the model used for $\Pr\left(G_p = g_p;\eta\right)$ (as long as its support is the same as the true model $\Pr\left(G_p = g_p;\eta_0\right)$), *U* will have mean zero for any correctly specified transmission model $\Pr\left(G_o = g_o \mid G_p = g_p;\gamma\right)$.

Optimality: Using the machinery provided by the coarsened-data semiparametric efficient framework, proving the optimality of the resulting estimator is not difficult. Though we will not present this proof here, we can prove that this estimator, in a neighborhood of the true η_0, has minimum asymptotic variance in the class of all estimators that are robust to parental genotype distribution. The asymptotic variance result is in terms of the standard definition of multi-parameter asymptotic efficiency, i.e., variance of *U* subtracted from the variance any competing robust estimator results in a matrix which is positive semi-definite. Thus we can see that the assumptions made by Whittemore (data being complete or component parameters being independent) are both unrealistic and unnecessary. Note that even though the estimated value of η does not affect the consistency of the estimates and the unbiasedness of resulting tests, it does affect their efficiency. *U* will only be optimal in a neighborhood of the true value of η_0. Thus there is still motivation for preferring reasonable and biologically motivated models for the distribution of parental genotypes.

The argument presented above has the additional benefit of being constructive. The estimator is derived directly from the appropriate likelihood and its robustness and optimality come directly from the geometry provided by the coarsened-data semiparametric efficient model theory framework. This allows one to apply a similar argument to related problems. This is the approach we utilize below.

D. Research Design and Methods

SPECIFIC AIM 1*: Develop statistical methodology for estimating and testing for haplotype effects on cardiovascular disease that are robust to haplotype frequency misspecification using unphased but complete genotype data.*

We will use the framework provided by coarsened-data semiparametric model theory to develop optimal tests and estimators of the effect of haplotypes on disease that are robust to incorrect estimates of haplotype frequencies. We will first focus on the simplest nuclear family structure: a pedigree consisting of one affected proband and their parents and then move progressively to consider more and more complicated pedigree structures culminating in multiple generations of affected/unaffected individuals with complete genotype data and fully observed founders.

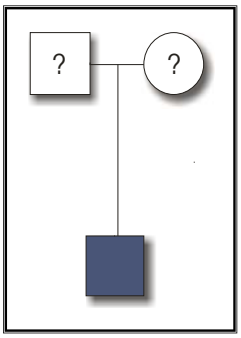

Figure 4: Case-Parent Design (question marks denote unknown affection status of genotyped parents, filled in square denotes affected male offspring).

Case-Parent Data: The triad sampling scheme (Figure 4) forms the simplest family-based design and, therefore, becomes our entry point for developing robust methods for haplotype association analyses. This was also the data structure considered in our previous work. If complete offspring haplotype data, H_o, and parental haplotype data, H_p, were available, and assuming no recombination within the gene cluster being characterized, one could base inference on a likelihood that conditions on parental haplotype. However, instead of haplotypes, frequently only unphased multilocus genotype data are available. Thus the data can be considered to be coarsened. Note that when genotypes are fully observed, the coarsening process is deterministic, so the coarsening process can be safely ignored. This can be seen by noting that, given the haplotype of an individual, their genotype is known with certainty. The coarsened data likelihood for the i^{th} family can be written as

$$\mathcal{L}_i = \sum_{(h_o,h_p)\in\mathcal{H}(g_o,g_p)} \Pr\left(H_{o,i}=h_o \mid H_{p,i}=h_p;\gamma\right)\Pr\left(H_{p,i}=h_p;\eta\right) \qquad (2)$$

where γ are the parameters of interest, η are nuisance parameters, and $\mathcal{H}\left(g_o,g_p\right)$ is the set of all haplotypes consistent with g_o and g_p. Once again, note that both probabilities involved in $\mathcal{L}_i$ are implicitly conditioned on the proband being diseased. This likelihood is very similar to that presented in the missing parental genotype case with the distinction that there is uncertainty in both the parental and offspring data. We will apply the same methodology highlighted above and derive an estimating function by taking the observed data score function for γ and subtracting off its projection onto the nuisance tangent space of η.

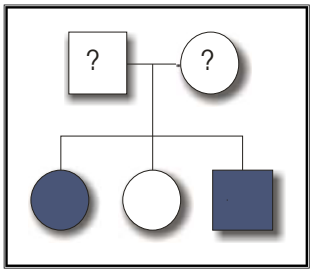

Figure 5: Inclusion of additional siblings (addition of an affected and an unaffected sister; question marks denote unknown affection status of genotyped parents).

Additional Siblings: Once we have characterized an estimator for the simpler case-parent scenario we will consider additional siblings in the design. Additional unaffected siblings will not be difficult to incorporate as transmissions to them by the parents will be independent of transmissions to either affected or unaffected siblings. Unaffected genotype data will affect the likelihood as an additional constraint on the haplotypes of the parents and affected offspring. However, when considering additional affected offsprings, transmissions from the parents to the affected offsprings may or may not be independent depending on whether or not there is linkage between the disease and the marker loci[25]. In the presence of linkage, these transmissions are correlated. Given complete data, a robust variance estimator will compensate for this effect. However, when parental genotype is not available, the robust variance estimator is no longer sufficient as the correlation also tends to bias the reconstruction of parental data[25]. It is unclear whether the reconstruction of haplotypes will be

significantly biased or whether the robustness to misspecification of parental haplotype frequencies property of our estimator will compensate for this effect. We will investigate these issues in the context of a richer likelihood that parameterizes the correlation due to linkage in terms of an identity-by-descent (IBD) parameter vector, π [25]. Note that π may or may not be considered a nuisance parameter depending on context. If one is conducting an association study as a follow-up to a linkage analysis that identified linkage in the region of the candidate markers, then π is a nuisance parameter. In this situation, π needs to be estimated only to give a valid test of association in the presence of linkage. This suggests that the coarsened-data semiparametric efficient framework may be useful in this context as well, deriving tests and estimators for the effect of genetic variants on disease in the presence of linkage that are robust not only to misspecification of haplotype frequencies but also to misspecification of IBD parameters. We will investigate these issues in this research proposal.

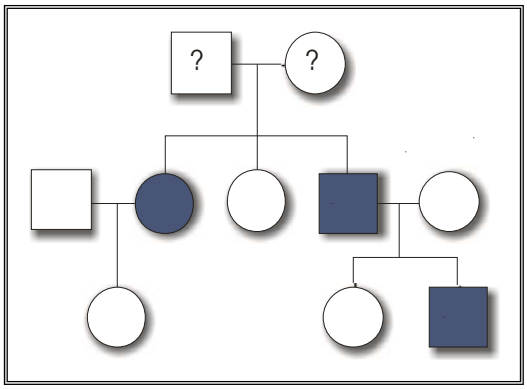

Figure 6: Sampling additional family members to create larger pedigree (question marks denote unknown affection status of genotyped parents).

Larger Pedigrees: We will consider more and more general pedigrees, using the same approach highlighted above. The likelihood considered will be expanded to have additional IBD probabilities corresponding to IBDs for different relational classes. Again, the semiparametric framework will be used to derive tests and estimators for the effect of genetic variants on disease in the presence of linkage that are robust to misspecification of the IBD parameters.

Quantitative traits: Above we considered discrete disease traits. However, there is also great interest in examining haplotype effects on intermediate continuous phenotypes such as lipid profiles, blood pressure, etc. Therefore we will also consider methods for estimating haplotype effects on quantitative traits that are robust to the misspecification of haplotype frequencies. We will begin by considering an adaptation of the variance-components model presented by Abecasis et al.[33] to haplotypes. We present the nuclear family case here but generalizations to more complex pedigrees are straightforward. Let y_{ij} be the quantitative trait of interest measured on the j^{th} offspring of the i^{th} family which consists of n_i offspring. Let H_i^* be the set of all offspring haplotypes consistent with the haplotypes of the parents (note: assuming no recombination, the cardinality of H_i^* will always be 4). Let Z_{ij} be a coded vector for $h_{o,ij}$, Z_{ik}^* correspond to the k^{th} member of H_i^* $(k=1,...,4)$, $b_i = \sum_{k=1}^{4} Z_{ik}^*$, and $w_{ij} = Z_{ij} - b_i$. Following Fulker et al.[34] and Abecasis et al.[35] we consider a model for the mean of y_{ij} to consist of orthogonal between and within family components of $h_{o,ij}$, i.e.,

$$E\left(y_{ij} \mid h_{o,i}, h_{p,i}\right) = \hat{y}_{ij} = \mu + \beta_b b_i + \beta_w w_{ij}.$$

For each family the $n_i \times n_i$ covariance matrix Ω_i has elements

$$\Omega_{ijk} = \begin{cases} \sigma_a^2 + \sigma_s^2 + \sigma_e^2 & \text{if } j = k \\ \pi_{ijk}\sigma_a^2 + \sigma_s^2 & \text{if } j \neq k \end{cases}$$

where π_{ijk} denotes the proportion of haplotypes shared IBD between siblings *j* and *k* in family *i*, σ_a^2 is the additive genetic variance of the major gene, σ_s^2 is the residual sibling resemblance, and σ_e^2 is the residual environmental variance component. If haplotypes were known, the regression coefficient β_w measures the

direct effect of haplotypes on y and is unaffected by population structure which is accounted for by β_b. Haplotypes, however are often unknown, and we are left with the "coarser" genotype data. Assuming a variance component model for y, the coarsened data likelihood component for the i^{th} family

$$\mathcal{L}_i = \sum_{\left(h_{o,i},h_{p,i}\right)\in\mathcal{H}\left(g_{o,i},g_{p,i}\right)} \left(2\pi\right)^{-n_i/2}\left|\Omega_i\right|^{-1/2} e^{-1/2\left(y_i-\hat{y}_i\right)'\Omega_i^{-1}\left(y_i-\hat{y}_i\right)} \Pr\left(h_{o,i},h_{p,i};\gamma,\eta\right) \qquad (3)$$

where y_i and $\hat{y}_i$ are n_i dimensional vectors with elements given by y_{ij} and $\hat{y}_{ij}$ respectively, and $h_{o,i}$ and $g_{o,i}$ consist of haplotypes or genotypes of all offspring in family *i*. As above, this procedure will tend to reintroduce sensitivities of parameter estimates and tests to haplotype frequency distributions. Therefore we will apply the coarsened-data semiparametric efficient machinery to this case as well, deriving tests and estimates that are invariant to the distribution of haplotype frequencies. Note that the transmission parameters, γ, may also be considered nuisance parameters in this case.

As discussed above in section 6.c., even though the methods proposed here will lead to consistent estimates and unbiased tests regardless of the estimated haplotype distribution, the optimality results implied by the projection argument framework apply only when the estimated haplotype distribution is in a neighborhood of the truth. It is a "local efficiency" result. Therefore, attention still needs to be paid to how haplotypes are modeled within this framework and those models that are clearly motivated by the underlying biology should be preferred.

SPECIFIC AIM 2*: Develop statistical methodology for estimating and testing for haplotype effects on cardiovascular disease that are robust to haplotype frequency misspecification using unphased but sometimes missing genotype data.*

We will extend the methods developed in Specific Aim 1 to consider different types of missing genotype data. These will include genotype data that are missing at a particular locus (perhaps due to the failure of an assay), as well as genotype data that are missing because an individual in the pedigree was not observed. The resulting range of pedigree structures considered will reflect the spectrum of family types expected in the _____ Study.

Missing Genotype at a Particular Locus: On any given individual there often will be genotype data missing at some loci but available at others, often due to the failure of an assay or a genotype reading that is hard to call. As long as the genotype can be assumed to be missing at random (MAR), this type of missing genotype data is particularly easy to deal with within the framework identified in specific aim 1. The likelihood given in (2) (or by (3) in the context of quantitative traits) is still valid: the only difference being that g_o and/or g_p is missing individual loci marker data, resulting in a larger set of haplotypes $\mathcal{H}\left(g_o,g_p\right)$ to sum over.

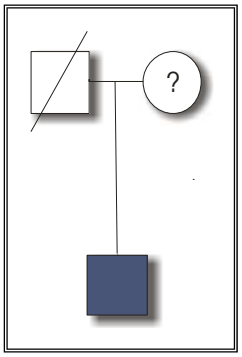

Figure 7: Triad data with missing parental genotype data (Father deceased; mothers affection status unknown).

Missing Pedigree Member: A more difficult type of missing genotype data to deal with is created when an individual in a pedigree is unavailable for study. Particularly problematic is missing parental (founder) data as transmission to any offspring will need to be inferred. Consider a case-parent design where one of the parents is deceased and unavailable for study. If the parent is missing at random, the likelihood can be written as

$$\mathcal{L}_i = \sum_{g_p \in \mathcal{P}\left(g_p^o\right)} \sum_{\left(h_o,h_p\right) \in \mathcal{H}\left(g_o,g_p\right)} \Pr\left(H_{o,i} = h_o \mid H_{p,i} = h_p; \gamma\right) \Pr\left(H_{p,i} = h_p; \eta\right) \qquad (4)$$

where $\mathcal{P}\left(g_p^o\right)$ is the set of parental genotype data g_p consistent with the observe parental data g_p^o.

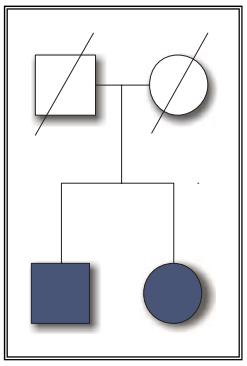

Figure 8: Affected sibpair data with missing parental genotype data (Both parents deceased).

When multiple affected offspring are available and at least some parental data are missing (see Figure 8), modeling haplotype/disease association in the presence of linkage requires additional IBD parameters to characterize the likelihood. We will consider the pedigree structures studied in Specific Aim 1 with differing levels of missing founder genotype data. We will apply the coarsened-data semiparametric framework and derive an estimating function by taking the observed data score function for γ and subtracting off its projection onto the nuisance tangent space of η.

Quantitative traits: When considering quantitative traits the same approach used in (4) can be applied to the variance-component likelihood given in (3), i.,e.,

$$\mathcal{L}_i = \sum_{g_{p,i} \in \mathcal{P}\left(g_{p,i}^o\right)} \sum_{\left(h_{o,i},h_{p,i}\right) \in \mathcal{H}\left(g_{o,i},g_{p,i}\right)} \left(2\pi\right)^{-n_i/2} \left|\Omega_i\right|^{-1/2} e^{-1/2\left(y_i-\hat{y}_i\right)' \Omega_i^{-1}\left(y_i-\hat{y}_i\right)} \Pr\left(h_{o,i}, h_{p,i}; \gamma, \eta\right).$$

Note that this likelihood already considers the correlation due to linkage. We will apply the coarsened-data semiparametric framework and derive a robust estimating function for this quantitative trait scenario as well. Note again that the transmission parameter γ can be considered part of the nuisance parameter.

Informative missingness: When missingness is non-MAR, the missingness mechanism will need to be considered as a necessary factor in the likelihood. We considered this problem in the context of a case-parent genotype association study[4;5]. We used a likelihood that conditioned on missingness pattern and found a class of identifiable missingness models. We also showed that, contrary to the standard situation, the complete-case estimating function is unbiased within the class of reasonable informative missingness mechanisms in which missingness is not a function of the genotype of any pedigree member except the missing individual. The fact that there exists (1) an unbiased estimating function and (2) a class of identifiable missingness models, suggests that the the semiparametric efficient framework can be used in this informative missingness context as well, improving the robust, but inefficient, complete-case estimator.

The pedigree data structures studied in specific aims 1 and 2 reflect the spectrum of family types expected in the _____Study, allowing us to use these novel methods in addressing Specific Aim 5.

SPECIFIC AIM 3*: Develop statistical methodology for estimating and testing for haplotype/haplotype and haplotype/environmental interaction effects using unphased genotype data.*

We will extend the methods developed in Specific Aims 1 and 2 to include haplotype/haplotype and/or haplotype/environment interaction effects. Note that we are referring to interaction between haplotypes at different loci, perhaps even on different chromosomes.

The concept of interaction has been somewhat controversial in epidemiology generally[36-39] and in genetic epidemiology in particular[40-42]. Much of this debate has centered on whether a statistically quantified interaction can adequately characterize biologic interaction (epistasis), and, if so, what scale of measurement is most appropriate[43]. In most discrete response models haplotype/haplotype or haplotype/environment interactions are most naturally parameterized in terms of departures from multiplicity of the risk ratios for each haplotype pair. Some have argued that an additive model for interaction is more compatible with the concept of biologic interaction, and propose parameters for quantifying this effect statistically[43]. Fortunately, most of these parameters can be expressed as functions of parameters from the multiplicative model, so additive interaction effects can be estimated via a, mathematically more convenient, multiplicative model. In the case of a discrete disease trait, this extension involves parameterizing $\Pr\left(H_{o1}=h_{o1},H_{o2}=h_{o2}\mid H_{p1}=h_{p1},H_{p2}=h_{p2};\gamma\right)$ in terms of the main effects of h_{o1},h_{o2} and their interaction, usually on the multiplicative scale (here h_{o1} denotes the offspring haplotype pair at loci 1 and h_{o2},h_{p1},h_{p2} are defined similarly). Robust methods for haplotype/haplotype interactions can then be obtaining by substituting

$$\Pr\left(H_{o1}=h_{o1},H_{o2}=h_{o2}\mid H_{p1}=h_{p1},H_{p2}=h_{p2};\gamma\right)\Pr\left(H_{p1}=h_{p1},H_{p2}=h_{p2};\eta\right)$$

for

$$\Pr\left(H_{o}=h_{o}\mid H_{p}=h_{p};\gamma\right)\Pr\left(H_{p}=h_{p};\eta\right)$$

into (2) or (4) and proceeding as in specific aim 1 or 2. Estimating haplotype/environment interaction will be addressed in a similar manner, working with $\Pr\left(H_{o}=h_{o}\mid H_{p}=h_{p},X;\gamma\right)$, where X is an environmental covariate, and substituting it for $\Pr\left(H_{o}=h_{o}\mid H_{p}=h_{p};\gamma\right)$ in (2) or (4). Methods for haplotype/haplotype and haplotype interaction with a quantitative trait can be addressed similarly. Note that for quantitative traits, much of the above discussion about the scale of measurement does not apply since the variance-components model is already parameterized on the additive scale.

Numerical Examples and Simulation Studies: We will examine numerical examples and conduct simulation studies to assess the feasibility and performance of the methods developed in Specific Aims 1-3. Special attention will be devoted to evaluating the robustness and power of the developed methods under a variety of penetrance, substructure, sample size, and pedigree structure scenarios.

SPECIFIC AIM 4*: Develop software tools for implementing novel methodology developed in Specific Aims 1-3 and make them available to the general research community.*

We will develop user-friendly computer software to facilitate the dissemination of our statistical methodology for use by research scientists attempting to detect genetic variants involved in complex diseases. This software will fully implement the methods developed in Specific Aims 1-3. We will develop the program to allow the standard linkage analysis pedigree file format (pedfile) to be read, making it easy for almost anyone to use the software with their data. Since the software will be developed concurrently with Specific Aims 1-3, much of the error checking, debugging, and quality control will be done at this time via the proposed simulation studies. We will also check the program by applying it to real data examples using other methods to confirm the results are reasonable. Though my mentors and I have considerable programming experience, we have budgeted some professional programmer support throughout the duration of this K25 award with greater support in the later part of the grant. This support will be especially valuable in making the software easy to use with pull-down menus and other prompts to lead the user through the analysis features via a graphical user interface. We will publicize the software through presentations at national meetings and in journal articles

detailing its capabilities. We will make this software and corresponding documentation freely available on the web, and will provide links to it from both the _____ and the _____ websites.

SPECIFIC AIM 5*: Apply methodologies developed in Specific Aims 1-3 to SNP genotypes in the _____ sample.*

We will use the methods developed in Specific Aims 1-3 to estimate and test haplotype effects in the _____ sample. We hypothesize that these new methods will substantially enhance our ability to detect genetic variants that predispose individuals to cardiovascular disease.

In collaboration with the _____ analysis team (lead by my mentor Dr. _____), I will identify candidate genes for association analysis on the basis of the strength of linkage evidence in the gene's region, published reports of candidate genes in the region as well as candidate genes from expression studies performed at _____ and beyond. Multiple SNP polymorphisms within each gene complex will be identified to comprehensively search for the true effect. The density of SNP development will depend on the size of the candidate gene, but we will isolate a minimum of 10 useful SNPs per gene. SNPs will be prioritized based on potential for biological effect (coding SNPs, 5′/3′ untranslated, regulatory regions), physical position, and allele frequency. Evolving mouse genome sequence and improving annotation of the human genome will facilitate identification of conserved sequence elements, which are presumably more likely to harbor functionally significant DNA. We anticipate the application of both well-characterized polymorphisms and novel markers derived from the databases.

Once we have identified and genotyped the collection of SNPs for a given candidate gene or set of candidate genes, we will apply the methods developed in Specific Aims 1-3. We will investigate haplotype effects on CAD as well as on intermediate phenotypes such as lipid levels or blood pressure. We will also investigate haplotype/haplotype interaction effects on CAD and intermediate phenotypes. Finally, we will determine whether differential haplotype effects on CAD or intermediate phenotypes exist depending on levels of an environmental exposure such as exercise or diet.

Potential Problems and Solutions

The main challenges facing this research program are computational. As the number of marker genotypes used to characterize a genomic region increases, the number of possible SNP haplotypes increases exponentially. Missing genotype data increases the number of haplotypes even further since missingness increases the number of genotypes that must be considered. The approach outlined in section 6.c.(previous studies) involves matrices whose dimensions depend on the total number of offspring parent genotype combinations consistent with the observed data. By analogy, a similar argument applied to the haplotype problem (specific aim 1) will likely involve matrices whose dimensions depend on the total number of offspring parent haplotype combinations consistent with the observed offspring parent genotype data. Thus the storage requirements of a naïve implementation of this approach will likely be enormous. Fortunately, modern computing is to the point where such computations can easily be accommodated for most reasonably sized problems. In addition, there is a great deal of structure in these matrices suggesting that more efficient algorithms or expressions may be found. To explore this possibility we plan to examine relatively simple examples (a few SNP loci) analytically in detail. We will use mathematical symbolic manipulation software (Mathematica or Maple) to examine the structure of specific components of the estimating function *U*. By comparing this structure over several examples it should be possible to identify simplifications of the formulation, leading to more efficient computational algorithms and thereby bypassing much of the computational complexity.

Summary

Haplotype-based statistical methods are quickly becoming important, even essential, tools for uncovering genetic components of complex disease. However, statistical tests of haplotype effects with unphased genotype data can be sensitive to estimates of haplotype frequencies even with family-based study designs and complete genotype information.

The broad objectives of this proposal focus on enhancing the collection of statistical methods researchers use to dissect genetic factors in complex diseases. Specifically, we propose to apply results from coarsened-data semiparametric efficient model theory to derive optimal tests and estimates of haplotype effects that are robust to haplotype frequencies using unphased genotype data. Missing genotype data and haplotype interaction effects will also be considered. User-friendly software tools that implement the novel methodology will be developed and made freely available to the general research community. Though the data structures considered are motivated by those found in the _____ study and the pending _____ Study, the methodology and software will be relevant to most family-based haplotype association studies. Finally, we will illustrate the utility of these newly developed techniques by applying them to the _____samples in fine mapping and candidate gene studies.

E. Human Subjects [Redacted]

1) Women and Minority Inclusion [Redacted]

2) Inclusion of Children [Redacted]

F. Vertebrate Animals: N/A

G. Literature Cited: [Redacted]

H. Consortium/Contractual Arrangements: N/A

I. Consultants: N/A

7. Checklist

8. Appendix