**Budget for Unified Statistical Methods for Sequence-based Association Studies**

Momiao Xiong (30% ) $ 39,600

Eric Boerwinkle (6%) $ 10,782

Ming Cao (100%) $23,000

Programmer (70% ) $42,000

Consultant:

Goncalo Abecasis, Felix E. Moore Collegiate Professor of Biostatistics

University of Michigan at Ann Arbor $3,800

Tuition $5,000

Grants Manager (1%) $ 324

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  |  |  |  |

Liming Liang

Assistant Professor of Statistical Genetics

Department of Epidemiology

Department of Biostatistics

Harvard University

E-mail: lliang@hsph.harvard.edu

Postdoctoral Fellow (50%)

Contact E-mail: [OIEP@hsph.harvard.edu](mailto:OIEP@hsph.harvard.edu)

Total: $65,000

**Equipment**

Qty Price

GPU-Powered Cluster: 1 $26,250

Data Server (20Tb storage): 1 $11,250

Total $37,500

**Supplies:** $ 7,143

Computer Media: $ 1,000

Printer Supplies: $ 1,000

Software:

Matlab and toolboxes $ 1,000

Visual Studio 2011 $ 1,400

Adobe Photoshop $ 1,000

Intel compiler for windows $ 743

Misc: $ 1,000

Total: $ 7,143

**Travel:**

2,000 for Investigators $ 5,000

**Other Expense:**

Publication Fees $ 3,040

### BUDGET JUSTIFICATION

**PERSONNEL:**

Momiao Xiong (3.6 cal mo/year), Ph. D., Professor, Division of Biostatistics, Human Genetics Center, at the School of Public Health at The University of Texas Health Science Center at Houston, will devote 30% of effort to the proposed research as Principle Investigator. He has developed numerous statistical methods for genetic studies of complex diseases involved in both qualitative and quantitative traits, DNA sequence analysis, detection of gene-gene interaction and gene-environment interaction, gene expression data analysis, pathway analysis, characterization and construction of genetic networks and metabolic networks, and computational systems biology. Dr. Xiong will have overall responsibility for directing all aspects of the research and for communicating critical issues with the entire investigative team . He will interact on a regular basis with Dr. Liming Liang, in the Department of Biostatistics at the Harvard University, Dr. Eric Boerwinkle, in the Human Genetics Center, School of Public Health at The University of Texas Health Science Center at Houston. Dr. Xiong will have primary responsibility for directing the development of unified statistical methods for sequence-based association studies which unify family and population design, will collaborate with Dr. Liang in devising analytic strategies for population and family-based eQTL analysis and with Dr. Boerwinkle in developing biologically motivated pathway analysis. Dr. Xiong will also supervise the software development to implement the proposed algorithms. Dr Xiong and Dr. Boerwinkle will jointly supervise real data analysis to evaluate the developed methods.

Eric Boerwinkle (0.72 cal mo/year), Ph. D., Kozmetsky Family Chair in Human Genetics, Professor, Director of Division of Epidemiology and Disease Control and Director of Human Genetics Center, at the School of Public Health at The University of Texas Health Science Center at Houston will devote 6% of effort to the proposed research as an investigator. Dr. Boerwinkle, an internationally recognized geneticist, has conducted numerous NIH funded projects in genetic studies of complex diseases as principle investigator. Dr. Boerwinkle is currently developing technologies and statistical methods for analysis of large-scale data from next-generation sequencing. His knowledge, experience and resources in performing genetic studies of common diseases will provide valuable guidance in the development of statistical methods. Dr. Boerwinkle will closely work with Dr. Xiong in developing biologically motivated pathway analysis and provide next-generation sequencing data to evaluate the proposed methods. He also will jointly with Dr. Xiong supervise real data analysis.

Goncalo Abecasis, Felix E. Moore Collegiate Professor of Biostatistics at the University of Michigan at Ann Arbor. Dr. Abecasis is an internationally recognized statistical geneticist and computational biologist. Dr. Abecasis' current research focuses on the development of statistical tools for the identification and study of genetic variants important in human disease with next-generation sequencing data. $3,800 annually is being request for the collaboration of Dr Abecasis. His role on the project will serve as a consultant. He will provide DNA variation and RNA-Seq data for 1000 individuals in the pedigrees sampled from the genetically isolated population of Sardinia and provide advice to develop statistical methods for population and family-based association studies as well as guidance in the analysis of Sardinia data.

.

Ming Cao (12 cal mo/year). A Ph D. student will devote 100% of effort to the proposed research for developing statistical methods, conduct simulations and real data analysis. He will be primary responsible for developing functional mixed model with both scalar and functional response for both family and population-based QTL and eQTL analysis for RNA-seq and methylation-seq data. He will also develop software for implementing the proposed methods and conduct large-scale real data analysis. He will assist with writing of research reports and articles.

Programmer (TBA**),** will devote 8.4 cal months of effort to the proposed research as a programmer. He will be responsible for developing software for implementing the proposed methods for testing association of full spectrum of DNA variation sampled from pedigrees and population in the presence of population structure, functional mixed linear models with scalar and functional responses for the family and population-based QTL and eQTL analysis, and identifying pathways associated with the complex diseases. He will also be responsible for developing simulation software and real data analysis to illustrate and validate the proposed statistical methods.

**Equipment**

Purpose of this application is to develop statistical methods and algorithms for genetic analysis of next-generation sequencing data. Specifically, we will perform genome-wide association analysis, family and population-based QTL and eQTL analysis and pathway analysis with next-generation sequencing data for thousands of individuals. We also need to perform large-scale simulations of the same size as real data for thousands of times. Therefore, this application raises great challenges to both computation and data storage hardware. For example, whole-genome sequencing data file for one person would consume 100GB, and 1000 individuals would require 100 Terra Bytes. Take the marker information, which is typically SNP, into memory needs more than 10G physical RAM. It is clear that this project will carry new demands for computational infrastructure. The current traditional single CPU serial processing would need years to complete one run of sequence-based association analysis. It is infeasible to use the traditional computers for sequence-based association studies and QTL (eQTL) analysis. In the recent years, all of the major computer hardware vendors have been moving to multi-core and GPU-powered architectures. This new architecture of high-performance computing provides powerful tools for large-scale computation. Some statistical analyses have been reported to speed up by 5000 times on GPU powered cluster. And for the data storage, we need several data server with high-capacity hard disk. Messages transfer from different nodes use lots of computing time. To improve the communication speed from nodes to nodes in the cluster system, we need to connect these nodes with a fast network switch. With the development of Graphics Computing, now we can use GPU to handle some specific problem at low cost and low power consumption. Analyzing of a large amount of SNPs with same method can be well paralleled by GPU computing. Therefore, we request $37,500 for purchasing 14 T storage ($9,700) and a GPU-powered cluster with 16 GPUs ($27,800) in year 1, which can shorten computation time from 1 year to two hours.

**Software:**

Also it is very important to have some software to analyze data and build our own applications. In order to rapidly implement developed new statistical methods and algorithms for family and population-based qualitative and quantitative association studies, pathway analysis, we use analytics platform such as Matlab, SAS, Adobe Photoshop and Visual Studio 2011for fast modeling and implement the final algorithm with C++. These software would definitely accelerate our development.

**Subcontract with Harvard University**

Liming Liang (1.44 cal mo/year)

A Postdoctoral fellow (6.0 cal mo/year)

Liming Liang (1.44 cal mo/year), Ph. D, Assistant Professor, Department of Epidemiology and Department of Biostatistics at the Harvard University, will devote 12% of effort to the proposed research. Dr. Liang is an outstanding young scientist. He has a Ph.D in biostatistics from the University of Michigan (with Dr. Gonçlo Abecasis). His specific training and expertise in key research areas include statistical genetics, population genetics and practical experience in genome-wide association study for complex diseases, gene expression quantitative trait loci (eQTL) mapping and DNA sequencing data since his involving in the early stage of the 1000 Genomes Project. His contribution resulted in 5 first/co-first author and 2 second-author papers in Nature, Nature Genetics, Nature Review Genetics, Bioinformatics, Genetic Epidemiology and Human Heredity, and 22 other publications in top biomedical journals including Nature, Nature Genetics, PLoS Genetics and others. Dr. Liang will contribute his expertise in the methodology development using DNA sequencing data in genetic association studies with pedigree and cases-control samples in the presence or absence of population structure (Aim 1), quantitative trait loci mapping based on RNA-sequencing data with family-based and population-based samples (Aim 2) and pathway analysis using DNA sequencing data from family-based and population-based samples (Aim 3). Dr. Liang will also apply the above developed methods to his ongoing breast cancer target sequencing project involving >4000 case-control samples from multiple populations (data expected early 2012, see support letter from Dr. Peter Kraft at Harvard School of Public Health), the RNA-sequencing data from a subset of the family-based sample that he previous led the statistical analysis (Dixon et al. 2007) via collaboration with Dr. William Cookson at Imperial College London, and the SardiNIA DNA and RNA sequencing data via collaboration with Dr. Goncalo Abecasis at the University of Michigan.

The 50% post-doc will be based at Harvard School of Public Health and co-advised by Dr. Liang and Dr. Momiao Xiong.