Proposed Master's Program in Computational Biology and Statistical Genetics

Departments of Biostatistics and Epidemiology
February 2013

BACKGROUND AND DEVELOPMENT OF THE DEGREE

Current Biostatistics Master's Programs:
The Departments of Biostatistics and Epidemiology both offer Master's degree programs that emphasize competence in biostatistical and epidemiological methods and both have been highly successful in training students. However, new technologies developed through the Human Genome Project are altering both the types of data that quantitative biological scientists now work with and the types of questions people hope to ask of the data. From searching for rare genetic variants, to using genomic markers in adaptive clinical trials, to searching for compound biomarkers and studying how microbial populations correlate with disease, understanding genomics has become a critical element in training our students for modern biomedical research. Even our prospective students are increasingly indicating their interest in the "Big Data" questions in biology, suggesting that they see a growing need for competence in these areas.

To address these issues, we are proposing to create a new MS in Computational Biology and Quantitative Genetics. This 80 credit program, to be offered jointly through the Departments of Epidemiology and Biostatistics, will build on our existing core competencies in these disciplines by integrating genomic methods and examples into our existing courses and by creating a small number of new courses and leveraging existing courses in other departments. The goal will be to provide students with a working knowledge of basic molecular biology, an understanding of DNA sequencing technology and data analysis, to give students working knowledge of computational and systems biology methods, and to provide broader training in the use of genomic data to address biomedical research questions.

We feel there is a need for a new Master's program that better prepares graduates to handle "Big Data" in addressing the biomedical research questions that are becoming increasingly commonplace in hospitals and universities, research organizations, and the pharmaceutical and biotechnology industries. As personalized genomic medicine increasingly becomes a reality, students with training in analyzing and interpreting the underlying data are going to become an invaluable resource driving discovery and innovation.

Summary of New Program:
- Allow admittance of students without a post-baccalaureate degree to a new Master's program in Computational Biology and Quantitative Genetics (80 credits), provided they have appropriate quantitative training as undergraduates or before arrival to the program.

- We are requesting to offer the degree of Master's of Science in Computational Biology and Quantitative Genetics to graduates of this new program, to better clarify the distinctions with our current MS/SM programs. We wish to obtain the appropriate approvals to use this new name. In the interim (pending approval for this new degree name), we would offer the Master's of Science in Biostatistics, with the area of interest being Computational Biology and Quantitative Genetics. In the proposal below we use the Computational Biology and Quantitative Genetics name.
- The new program will require course work (minimum of 60 credits), followed by a required supervised Collaborative Research Thesis (minimum of 20 credits, possibly performed off-site), for a total of 80 credits. This new program would typically be of 18-24 months duration. The Collaborative Research Thesis must be accomplished at an authorized location where trainees will have access to mentoring by experienced quantitative scientists with expertise in the analysis of genomic data.
- The program would be designed to give students essential skills for the job market while at the same time providing a route to a PhD in Biostatistics or an SD in Epidemiology with an emphasis in computational biology.
- No changes to our current Master's programs in either Biostatistics or Epidemiology.

RATIONALE/PURPOSE

- The first human genome was sequenced little more than 10 years ago; an undertaking that spanned nearly 15 years and cost billions of dollars. Technologies available today allow genomes to be sequenced in days at a cost of thousands of dollars—and the cost continues to fall. This has resulted in a flood of genomic data and a host of new applications for its use.
- There is a critical shortage of well-trained Master's level scientists skilled in the analysis of such genomic data, and a particularly acute shortage of those with rigorous quantitative training. Demand for such people exists in both industry and university-based health research settings, in the Boston area, the U.S., and across the world.
- The Departments of Biostatistics and Epidemiology already offer many excellent courses that would provide rigorous training in quantitative methods required for this Computational Biology and Quantitative Genetics Master's program. The existing courses are often undersubscribed, and thus could accommodate additional enrollees. Department faculty are also well integrated in various clinical and epidemiologic research projects at HSPH, Harvard University, and the affiliated hospitals that will provide excellent collaborative research training. We also have industry contacts and partnerships that we can use to develop potential Collaborative Research Thesis positions and help us with job placement. This new program will also dovetail well with Harvard's recently submitted Clinical and Translational Science Awards (CTSA) proposal for which HSPH played a key role.
- These existing courses, however, need to be updated to include relevant examples of genomic data and genomic data analysis that emphasize the relevance of the course material to the analysis of such data and the use of genomic data in a range of applications.

- Implementing this new program will provide an important impetus for our Departments to review and update our intermediate Biostatistics and Epidemiology course curricula. We plan to coordinate our course offerings to help with student course scheduling and cross-enrollment, to adapt our existing courses to include genomic applications, to add an introductory course in molecular biology, to expand our programming course offering, and to add additional courses emphasizing essential skills and applications in computational It is clear that these improvements will be helpful not only to Master's students in Biostatistics and Epidemiology, but also to many other HSPH students who take intermediate courses in our departments and who are interested in applications of genomics in public health.
- Although the goal of this proposal is to create a terminal Master's program for students interested in careers involving quantitative analysis and interpretation of genomic and other data, we recognize that this program will also help strengthen our offering in computational biology and bioinformatics and may serve as a mechanism for recruiting students interested in pursuing a doctoral degree in Biostatistics or Epidemiology with an emphasis in quantitative genomics. Direct knowledge of such students through this program would allow us to identify exceptional students who are appropriate for a doctoral program.
- The current Master's programs in Biostatistics and Epidemiology attract students with different career histories and different career paths than does the proposed program. While we hope to continue to attract students with strong quantitative backgrounds, we also recognize that the proposed program is likely to attract students who have stronger biological training and who are interested in developing better quantitative skills. We believe that the students interested in this program will likely have career goals that are complementary to the existing Master's programs and will not adversely affect enrollment in these.

Types of Students that would enroll:
All candidates for admission to the MS in Computational Biology and Quantitative Genetics program should have successfully completed calculus through partial differentiation and multivariable integration, one semester of linear algebra or matrix methods, and either a two-semester sequence in probability and statistics or a two-semester sequence in applied statistics. Students should have at least one semester of training in biology, with some familiarity with molecular biology and genetics. Practical knowledge of computer scripting and programming as well as experience with a statistical computing package such as R would be highly desirable. Applicants should also show excellence in written and spoken English. Evidence that these requirements have been fulfilled should form part of the application. Applicants would also be encouraged to have completed other courses in quantitative areas and in areas of application in the biological sciences. Additional research or work experience would be considered beneficial, but not required. Students who enroll in this program will come with undergraduate degrees in the mathematical sciences or from allied fields (biology, psychology, economics, etc.). Most students would have the goal of working in a position such Bioinformatics Analyst or Bioinformatics Engineer in teaching hospitals, universities, research organizations, or the pharmaceutical and biotechnology industries.

Expected Size:
Approximately 20-30 students per year, though starting smaller (10-12 students).

NEW DESCRIPTION FOR HSPH CATALOG

Master of Science in Applied Biostatistics (60-credit program)
The master's degree program in Computational Biology and Quantitative Genetics is aimed at students seeking both theoretical and practical training in the quantitative analysis and interpretation of large-scale, public health, genomic data. Students will receive training in quantitative methods, including linear and logistic regression, survival analysis, longitudinal data analysis, statistical computing, clinical trials, statistical consultation and collaboration, and epidemiology. Students will also be provided with firm grounding in modern molecular biology and genetics, computer programming, and the use and application of tools for analysis of genomic data, and in methods for integrative analysis and meta-analysis of genes and gene function. Typically of 18-24 months duration, the focus of this program is on training graduates for positions in which they can apply their new-found expertise to the analysis of biomedical and genomic data in teaching hospitals, universities, research organizations, and the pharmaceutical and biotechnology industries.

Applicants to the MS in Computational Biology and Quantitative Genetics program should have successfully completed calculus through partial differentiation and multivariable integration, one semester of linear algebra or matrix methods, and either a two-semester sequence in probability and statistics or a two-semester sequence in applied statistics. Students should have at least one semester of training in biology, with some familiarity with molecular biology and genetics. Practical knowledge of computer scripting and programming as well as experience with a statistical computing package such as R would be highly desirable. Applicants should also show excellence in written and spoken English. Evidence that these requirements have been fulfilled should form part of the application. Applicants would also be encouraged to have completed other courses in quantitative areas and in areas of application in the biological sciences. Additional research or work experience would be considered beneficial, but not required. Students who enroll in this program will come with undergraduate degrees in the mathematical sciences or from allied fields (biology, psychology, economics, etc.). Most students would have the goal of working in a position such Bioinformatics Analysts or Bioinformatics Engineer in teaching hospitals, universities, research organizations, or the pharmaceutical and biotechnology industries.

(Minor modifications would be made in the rest of text as necessary.)

DEGREE REQUIREMENTS

Course Requirements:
A total of 80 credits are required for the MS in Computational Biology and Quantitative Genetics. The School of Public Health requires that these include EPI ~~200 or~~ 201 and 202.

A minimum of 60 credits of course work is required. This includes a 47.5 credit ordinally graded core curriculum consisting of:

BIO 210	Analysis of Rates and Proportions (5 credits)
BIO 211	Regression and ANOVA for Experimental Research (5 credits)
BIO 222	Basics of Statistical Inference (5 credits)
BIO 223	Applied Survival Analysis (5 credits)
BIO 226	Applied Longitudinal Analysis (5 credits)
BIO 290	Genomics and Genetics for Health Research (2.5 credits)
BIO 506	Introduction to Computational Biology (5 credits)
BIO 508	Genomic Data Manipulation (5 credits)
EPI 201	Introduction to Epidemiology Methods: 1 (2.5 credits)
EPI 202	Elements of Epidemiologic Research: Methods 2 (2.5 credit)
EPI 249	Molecular Biology for Epidemiologists (2.5) credits
EPI 288	Data Mining and Prediction (2.5 credits)

Students with prior equivalent background to any of these courses or strong reasons to take a different course can request permission from the Director of Master's Programs for a substitution of one or more of these required courses.

A minimum of 12.5 additional credits will come from the following list of elective courses:

BIO 212	Survey Research Methods in Community Health (2.5 credits)
BIO 214	Principles of Clinical Trials (2.5 credits)
BIO 227	Fundamental Concepts in Gene Mapping (2.5 credits)
BIO 257	Advanced Statistical Genetics (5 credits)
BIO 283	Spatial Statistics for Health Research and Social Inquiry (5 credits)
BIO 287	Public Health Surveillance (2.5 credits)
BIO 503	Introduction to Programming and Statistical Modeling in R (1.25 credits)
BIO 504	Introduction to Geographical Information Systems using ArcGIS (1.25 credits)
BIO 505	Database Design and Use for Health Research (1.25 credits)
BIO 513	Advanced Computational Biology and Bioinformatics (2.5 credits)
BIO 514	Introduction to Data Structures and Algorithms (2.5 credits)
BIO 521	Introduction to Social and Biological Networks (5 credits)
EPI 203	Study Design in Epidemiologic Research (2.5 credits)
EPI 204	Analysis of Case-Control and Cohort Studies (2.5 credits)
EPI 215	High Dimensional Data Analysis for Epidemiologists (2.5 credits)
EPI 221	Pharmacoepidemiology (2.5 credits)
EPI 222	Genetic Epidemiology of Diabetes (5 credits)
EPI 240	Biomarkers in Epidemiology Research (1.25 credits)
EPI 271	Propensity Score Analysis (1.25 credits)
EPI 289	Causal Inference (2.5 credits)

EPI 293	Analysis of Genetic Association Studies Using Unrelated Subjects (2.5 credits)
EPI 507	Genetic Epidemiology (2.5 credits)
EPI 511	Advanced Population and Medical Genetics (2.5 credits)
ID 271	Advanced Regression for Environmental Epidemiology (2.5 credits)
RDS 280	Decision Analysis for Health and Medical Practices (2.5 credits)
RDS 282	Cost-Effectiveness and Cost-Benefit Analysis for Health Program Evaluation (2.5 credits)
RDS 285	Decision Analysis Methods in Public Health and Medicine (2.5 credits)

Additional courses from Harvard Medical School's Program in Biomedical Informatics

BMI 701	Introduction to Biomedical Informatics I (4 credits)
BMI 702:	Introduction to Biomedical Informatics II (4 credits)

Other courses may also be acceptable to satisfy these 15 additional credits. Students are advised to consult with the Director of Master's Studies to check prior to enrolling in the courses in question.

Research Ethics:
Students must satisfy a research ethics requirement through attendance at a lecture series or satisfactory completion of a web-based training program. Students who feel they have already completed an equivalent training program must submit adequate documentation to, and receive approval from, the Director of Master's Studies during the first semester in residence.

New Courses:
Wherever possible, we have tried to leverage the existing course offered at HSPH to develop this Master's program. It may require some shuffling of course, and reviving courses that are, at present, not offered (such as BIO 290). We may also require the addition of a second section of BIO 512, to be offered on the Longwood Campus. The current, highly-successful course if offered at FAS and many students interested in the course have expressed concerns regarding travel to and from Cambridge for the class. Creating a second section would allow us to continue to meet the needs of students in Cambridge while better serving students in this proposed program.

Further development of these courses will proceed in the coming months.

Further Curriculum Development:
In developing this proposal, we recognized that our goal was to provide students with a set of essential skills, outlined in the final pages of this proposal. Looking carefully at our existing course offerings, we concluded that students were receiving training in many of what we perceived as essential, but that the courses often failed to include concrete examples that included the analysis of genomic data. For many students interested in training in computational biology and quantitative genetics, this failure left an "application gap" in which the might not fully understand how the methods they were

being taught could be applied to genomic data. We also recognized that given the rapid proliferation of genomic data into nearly every area of biomedical research, that developing examples leveraging genomic data could be of great service to our existing Master's and Doctoral students.

To address both these needs, we are proposing to invest in curriculum development focused on developing genomic data examples to be included in the core and elective courses outlined in this proposal. We believe that this is a necessary ingredient for the success of the proposed program, but it will also have the additional benefit of making our overall course offerings more useful for our students and relevant to the data types they are increasingly likely to encounter.

CULMINATING EXPERIENCE: COLLABORATIVE RESEARCH THESIS

Students must satisfy a 20 credit ordinally graded Collaborative Research Thesis (???) or Thesis, taken after the required course work has been completed. This will typically involve data analysis for a research project under the direction of one or more mentors. These projects could be supervised primarily by a faculty member in Biostatistics or Epidemiology, or co-supervised by a doctoral-level investigator (at Harvard or elsewhere) and a faculty member in Biostatistics or Epidemiology.

In this Collaborative Research Thesis, students will perform activities related to the design, conduct, and analysis of research studies involving genomic data, with a focus on data analysis and interpretation as well as scientific presentation. Students will carry out an extensive data analysis, including data summaries and graphical displays, regression methods, biological data interpretation, and comparison of alternative methods. They will then write a Master's paper of approximately 20-25 double-spaced pages excluding tables, figures, and references that describes the medical or public health problem of interest, summarizes the appropriate data analyses, and provides a scientific interpretation of the data, in a standard scientific writing style. The student will also orally present this work in a seminar of approximately 30 minutes in length. The Master's paper and oral presentation will primarily be the work of the student, with only advisory input from the mentor(s). The Master's paper and oral presentation will be evaluated by a review committee consisting of three members. The members will include the student's Thesis mentor(s), the Director of Master's Programs, and other Biostatistics faculty members as needed. The Master's paper must be submitted to the review committee at least two weeks prior to the oral presentation. A written evaluation will be provided to the student.

PROPOSED CURRICULUM:

Summer

SAS Programming Bootcamp/EdX-type on line version

Fall Semester, first year:

BIO 210 Analysis of Rates and Proportions (5 credits)
BIO 211 Regression and ANOVA for Experimental Research (5 credits)

EPI 201 Introduction to Epidemiology Methods: 1 (2.5 credits, Fall 1)
EPI 202 Elements of Epidemiologic Research: Methods 2 (2.5 credits, Fall 2)
EPI 249 Molecular Biology for Epidemiologists (2.5 credits, Fall 1)
BIO 290 Genomics and Genetics for Health Research (2.5 credits, Fall 2)

Winter Session first year:
EPI 288 Data Mining and Prediction (2.5 credits)

Spring Semester first year:
BIO 223 Applied Survival Analysis (5 credits)
BIO 508 Genomic Data Manipulation (5 credits)
BIO 512 Introduction to Computational Biology and Bioinformatics (5 credits)
Electives (5 credits)

Summer/Fall Semester first year:
XXX Collaborative Research Thesis (15-20 credits)

Fall Semester, second year:
EPI 507 Genetic Epidemiology (2.5 credits)
EPI 293 Analysis of Genetic Association Studies Using Unrelated Subjects (2.5 credits)

PROCESS FOR DOCUMENTING COMPLETION OF THE DEGREE:

The Department of Biostatistics will develop a Degree Program Form for this new program. The purpose of this form is to permit the Director of Master's Studies to verify that all degree requirements are being met. Details of the degree requirements will be given in the Graduate Student Handbook of the Department. Students are responsible for the requirements at the time they enter the program. Any change from Departmental requirements must be approved by the Director of Master's Studies, including course waivers or substitutions.

This Degree Program Form will include core requirements, the epidemiology requirement, the research ethics requirement, and electives. It will be filled out by the student in January, reviewed and signed by the student's academic advisor, and then final approval must be obtained from the Director of Master's Studies. The Department will also follow student course selections and grades as students move through the program.

In addition, we will develop a Master's Paper Completion Form. This form will be filled out after the Master's paper has been submitted and the oral presentation made. It will confirm that the Collaborative Research Thesis has been successfully completed, and will also include a written evaluation of the Master's paper and oral presentation.

DEGREE TO BE AWARDED:

Master's of Science in Computational Biology and Quantitative Genetics. We wish to obtain the appropriate approvals to use this new name. In the interim (pending approval for this new degree name), we would offer the Master's of Science in Biostatistics or Masters of Science in Epidemiology, with the area of interest being Computational Biology and Quantitative Genetics.

DESCRIPTION OF LIKELY CAREER PATHS FOR GRADUATES:

The exponential drop in the cost of genome sequencing has produced a strong and growing demand for quantitative scientists trained in the analysis and interpretation of genomic data. The demand for such biological "Big Data" scientists is strong and growing. We anticipate that graduates will take positions as Bioinformatics Analysts or Bioinformatics Engineers in teaching hospitals, universities, research organizations, or the pharmaceutical and biotechnology industries. They will be working on cutting edge research projects at the interface of biostatistics, epidemiology, computer science, and biology, leveraging individual and population-level genomic data in applications ranging from basic research to personalized medicine. After a few years, our graduates would be expected to expand to project leadership and management and in the supervision of junior computational biology and bioinformatics staff. Some graduates may decide to pursue a doctoral degree at some point in time after completing the MS.

FACULTY MEMBERS ASSOCIATED WITH THE PROGRAM:

William Barry (BIO)
Immaculata deVivo (EPI)
David Christiani (EPI)
Aditi Hazra (EPI)
Winston Hide (BIO)
Curtis Huttenhower (BIO)
Christoph Lange (BIO)
Liming Liang (BIO/EPI)
Xihong Lin (BIO)
Marc Lipsitch (EPI)
Xiaole Shirley Liu (BIO)
Peter Kraft (EPI)
Franziska Michor (BIO)
Murray Mittleman (EPI)
Loreli Mucci (EPI)
Megan Murray (EPI)
Shuji Ogino (EPI)
Jukka-Pekka Onnela (BIO)
Giovanni Parmigiani (BIO)
Alkes Price (BIO/EPI)
John Quackenbush (BIO)

Eric Tchetgen Tchetgen (BIO.EPI)
Lorenzo Trippa (BIO)
Shelley Tworoger (EPI)
Guo-Cheng Yuan (BIO)

Other Biostatistics and Epidemiology faculty members will be involved with academic advising, course instruction, and mentoring students in this program.

ESSENTIAL SKILLS:

The proposed curriculum for the Master's in Computational Biology and Quantitative Genetics is designed to provide students with what we perceive to be essential skills to contribute to research projects involving large, complex genomic datasets that are becoming increasingly common in all areas of biomedical, biological, and public health research. We see these skills as representing core competencies in five areas: biological background to understand and interpret the data, bioinformatics background providing familiarity with essential tools and data resources, computational skills required for analyzing and managing large data, statistical skills required to appropriately analyze large quantitative datasets, and epidemiological skills necessary for experimental design, conduct, and analysis. These essential skills are summarized below:

<u>Biological Background</u>

1. Working knowledge of molecular genetics, including Mendelian inheritance, complex trait inheritance, and the essentials of DNA and its role
2. Working knowledge of the Central Dogma of Molecular Biology (DNA->RNA->protein), including an understanding of transcription, splicing, and translation
3. Working knowledge of feature organization within the genome (including genes and regulatory regions) as well as the challenges of gene finding
4. Familiarity with processes that regulate gene expression and protein translation, including transcription factors, miRNAs, etc
5. Familiarity with epigenetic regulation, including DNA methylation and histone modification
6. Familiarity with gene functional descriptions (such as the Gene Ontology) and basic signal transduction and metabolic pathways
7. Familiarity with modern technologies, including genotyping and gene expression arrays, genome-seq, exome-seq, RNA-seq, ChIP-seq, etc, and their applications
8. Understanding of metagenomics and its importance

<u>Bioinformatics Background</u>

1. Familiarity and ability to use the major genomics data resources, including those at NCBI, EBI, and UC Santa Cruz
2. Understanding of sequence alignment algorithms
3. Basic knowledge of gene finding methods and challenges
4. Familiarity with gene functional annotation, including Gene Ontology and Pathway databases
5. Ability to write simple scripts to download and transform data into useful formats
6. Working knowledge of basic data analysis and data mining techniques such as hierarchical clustering, k-means clustering, PCA, SVD

7. Working knowledge of basic statistical tests including t-tests, ANOVA, linear modeling, Fisher's Exact Test, Kolmogorov-Smirnov statistics, chi-squared tests, and their applications
8. Familiarity with Bayesian statistical approaches, MCMC, Gibbs Sampling, and HMMs
9. Familiarity with machine learning approaches such as Bayesian Networks and Artificial Neural Networks, Classification and Regression Trees, and genetic algorithms
10. Familiarity with bootstrapping, jackknifing, and sensitivity/recall/ROC curve analysis, and other empirical method
11. Familiarity with modern network theories, including scale-free network models and their measures.

Computational Skills

1. Working knowledge of UNIX
2. Working knowledge in a scripting language such as perl or python
3. Working knowledge with an advanced programming language such as c, c++, or java
4. Working knowledge of R/Bioconductor
5. Familiarity with database programming and modern web technologies

Biostatistics Skills

1. Fundamentals of experimental design
2. Rates and proportions
3. Parametric and non-parametric statistical methods
4. Basic inference
5. Regression and ANOVA
6. Applied survival analysis
7. Applied longitudinal analysis
8. Bayesian statistical analysis

Epidemiology Skills

1. Ability to critique the existing evidence for a particular research topic, review and summarize information from many studies
2. Ability to develop a research question and formulate study objectives, define a set of related specific aims, write a research protocol for a given study question
3. Ability to identify relevant ethical issues in a given study
4. Ability to develop sampling procedures and be able to undertake calculations for sample size and power requirements

5. Ability to identify methods of data collection appropriate to the study design and population
6. Ability to design efficient data collection and data management procedures
7. Ability to choose and use the techniques appropriate for estimation and hypothesis testing in selected situations
8. Ability to perform data cleaning and data management operations to prepare for data analysis
9. Familiarity with data cleaning and management techniques used to prepare unconventional data sources (such as large pharmacoepidemiology data, vital records, EMR, and cohort data) for causal inference exercises
10. Familiarity with a comprehensive set of statistical methods suitable for a wide range of epidemiological situations; ability to summarize and present data in graphs and tables
11. Familiarity with methods to assess and possibly correct for measurement error
12. Familiarity with methods for managing missing data problems
13. Ability to interpret the results of statistical procedures and draw appropriate conclusions