

# Bioinformatics Core Strategy Document 2009-2011

Harvard School of Public Health  
Oliver Hofmann ([ohofmann@hsph.harvard.edu](mailto:ohofmann@hsph.harvard.edu))  
Winston Hide ([whide@hsph.harvard.edu](mailto:whide@hsph.harvard.edu))  
Department of Biostatistics  
SPH2, 4th Floor, Room 437A  
655 Huntington Ave Boston, MA 02115

## Preamble

*“An opportunity exists for the Harvard School of Public Health to establish the lead in public health bioinformatics. The School is positioned to harness its considerable potential in intellectual capacity, areas of study and its access to exceptional resources for genome scale data and analysis. The School can achieve its potential in leadership through establishment of an interdisciplinary, research driven training program with a common identity grounded in bioinformatics.”*

Public Health Bioinformatics at the Harvard School of Public Health, strategic review document.  
Winston Hide, 2007

The Harvard School of Public Health is uniquely positioned to provide leadership in the field of bioinformatics and public health. As a result of a strategic review, the School has undergone a planning process and has subsequently committed ongoing resources to the development of bioinformatics capacity as part of its research program in public health.

In October 2008 the School appointed Winston Hide to the faculty of the Department of Biostatistics to lead Bioinformatics development through scientific oversight of the HSPH Bioinformatics Core (HBC). He has recruited a Core Manager (Oliver Hofmann) and two dedicated developers (Alyssa Porter and Ray McGovern). As the Core has roles across the School and within the Harvard-wide [Catalyst](#) project, it is essential that we seek input on the development and roles of the HBC.

This document provides an overview of the aims and objectives for the HSPH Bioinformatics Core from 2009 to 2011.

# Mission

To develop and advance the application of bioinformatics and computational biology in public health

## Summary

In this document we summarize our aims and objectives for the next three years, provide an overview of the organization of the HBC and its relationship with other Cores and centers at Harvard and in the Longwood community, and describe the current staff and hiring plans. An overview of supported experimental studies is given along with current plans for IT infrastructure development, teaching and outreach.

The Harvard School of Public Health is positioned to take a leading role in interdisciplinary research involving the computational analysis of complex relationships between genes and their environment as well as basic biological and quantitative sciences. The analysis of high-throughput genomic and proteomic data in the context of Public Health studies is a rapidly expanding area, providing a more rapid route to health interventions. High dimensional genomic data is increasingly becoming a mainstream tool, requiring expertise and infrastructure for its exploitation. The Harvard School of Public Health Bioinformatics Core (HBC) aims to establish HSPH and its researchers as competitive leaders in the application of new technologies and methods to the study of Public Health.

The HBC is integral to the Program for Quantitative Genomics. It provides a single point of contact for computational biology, providing venue to apply genomic approaches together with established and developing Public Health methodologies, basic biology, epidemiology, environmental health, biostatistics and bioinformatics to develop novel approaches to improve human health. As a result, HSPH exploits data-driven research, improving grant funding opportunities which are becoming increasingly computational. By providing access to bioinformatics expertise for researchers at HSPH, the HBC opens new areas of research, enhances the quality and consistency of high-throughput data analysis and improves the School's ability to support research in this mission-critical area.

The HBC will improve consistency of data analysis, reduce redundancy across studies, provide an environment for individual research groups to hire and retain expert staff to be "hosted" at the core and provide connections to other ongoing developments in the field of computational biology such as the CTSA Catalyst program. Through triage processes developed in collaboration with Catalyst, we will ensure that studies requiring assistance will be supported through HBC staff, or importantly through consultation with Biostatisticians, Epidemiologists, Genome Scientists or any other field specialist relevant to our remit.

## Aims

The Core will aid researchers at the School with the management, integration and contextual analysis of biological high-throughput data, provide training on tools, databases and best practices, foster collaboration and a community of bioinformatics activities, and help build a unified infrastructure supporting a diverse set of experimental systems and high-throughput biological data.

### **Aim 1: Provide support on management and analysis of high-throughput biological data**

- establish “best practices” for data analysis and software development
- develop customized software to assist researchers with data analysis
- establish a dialogue between biomedical researchers, biostatisticians and computational biologists at HSPH to improve analytical methods
- Provide a venue for establishment of collaborations between public health researchers such as biostatisticians, bioinformaticists, epidemiologists and environmental health scientists
- support the development of research grants

### **Aim 2: Disseminate information and provide training**

- advertise the availability and establish processes to interact with the HBC among the HSPH community and to the Catalyst system
- coordinate the development of new methods and infrastructure with other service cores in the Longwood area
- organize short courses on current topics in bioinformatics
- provide training sessions and ad hoc help for supported software and databases
- disseminate established best practices for the management and analysis of high-throughput data among members of the School
- support the development of the community of bioinformatics researchers and users in the Longwood area

### **Aim 3: Build a unified IT and software infrastructure**

- assist the HSPH IT department in the development and maintenance of a computational and storage infrastructure that readily supports the demands for high-dimensional bioinformatics data analysis
- provide access to large public health data sets (such as the Framingham Heart Study)
- develop biological databases to provide essential contextual information to specific research projects
- install and maintain a web-based portal to provide basic storage and analytical requirements (self service)
- identify, license and support commercial systems biology databases

## Objectives

Since hiring a Bioinformatics Core Manager in October 2008 (Oliver Hofmann, Department of Biostatistics) the HBC has set up a collaborative infrastructure to provide *ad hoc* support to ongoing projects, generated an overview of current resources and experimental data used at HSPH, and initiated communication with other service cores and projects (Center for Computational Cancer Biology, DFCI; Environmental Statistics and Bioinformatics Core; Genetics Consulting; Catalyst / CTSA; Bioinformatics, Channing Laboratories and others) to explore current needs and applied solutions in Public Health Bioinformatics.

Consistent with a user survey that was conducted in 2006, direct interaction with over ten research groups at HSPH has revealed strong demand for:

- the management of large scale biological data sets
- statistical analysis and quality assessment of these data sets
- a resource providing biological context information
- information and training on best practices and existing data and software resources at the School and within the bioinformatics community

We will address these needs in three stages and focus efforts on supporting studies involving pathogens, model organisms and human data:

### **Short term (next six months):**

- provide direct support to existing projects to for pre- and post-grant development — measurably improving the ability for the School to secure grants
- recruit additional staff members to support the identified needs
- develop a strategy to deploy IT infrastructure supporting demands in high-throughput data analysis
- provide an updated HBC web site with current information, description of processes, contact forms and links to existing resources
- advertise basic services and contact information for the Core within the School
- establish a triage process with other Catalyst-related cores and projects and integrate this approach with other cores and centers at DFCI and HSPH
- development of standard operating procedures to handle grant support requests

### **Mid term (next twelve months):**

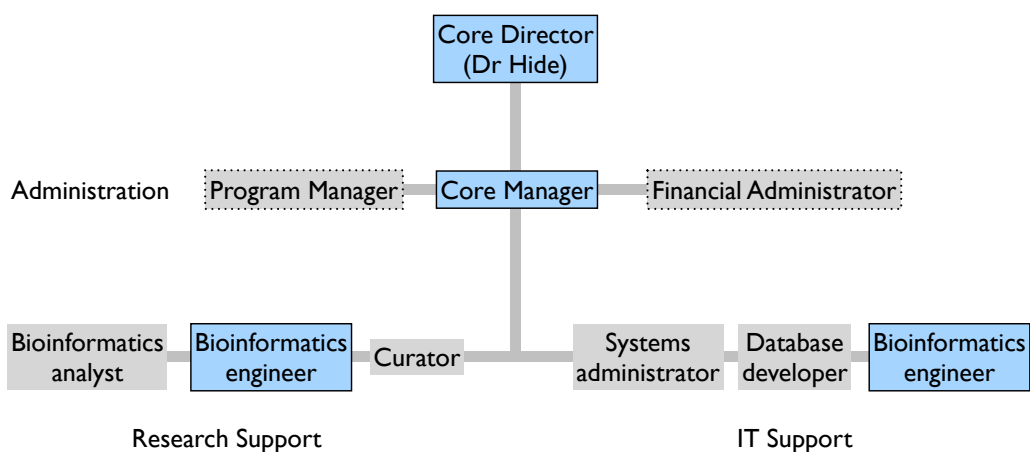
- deploy and utilize new IT infrastructure
- develop and distribute standard operating procedures for the most commonly requested experimental data
- establish a basic self-service environment and make it available to the HSPH community
- hold training sessions to utilize the self-service environment
- advertise Core services within the Longwood community

**Long term (by July 2011):**

- deploy infrastructure and databases to provide additional contextual (internal) information to the established self-service environment
- provide a bioinformatics web portal on a web server embedded in the new IT infrastructure environment
- manage, process and analyze data from all platforms and studies utilized at the School
- organize regular short courses on popular bioinformatics topics and methods

## Organizational structure of the HBC

The HBC is overseen by a faculty member (Dr. Winston Hide) who provides scientific guidance and direction. Day-to-day activities are managed by the Bioinformatics Core manager (Dr. Oliver Hofmann) who will oversee a team of six staff members (software and database engineers, IT support staff, curator and research analyst). Two software engineers have been recruited to date. The manager liaises with researchers at the School and maintains relationships to other bioinformatics cores at Harvard and within the Longwood community.



*Organizational structure of the HBC. Positions already filled are highlighted in blue and outlined. We expect the boundaries between "research support" and "IT Support" positions to shift depending on the projects supported.*

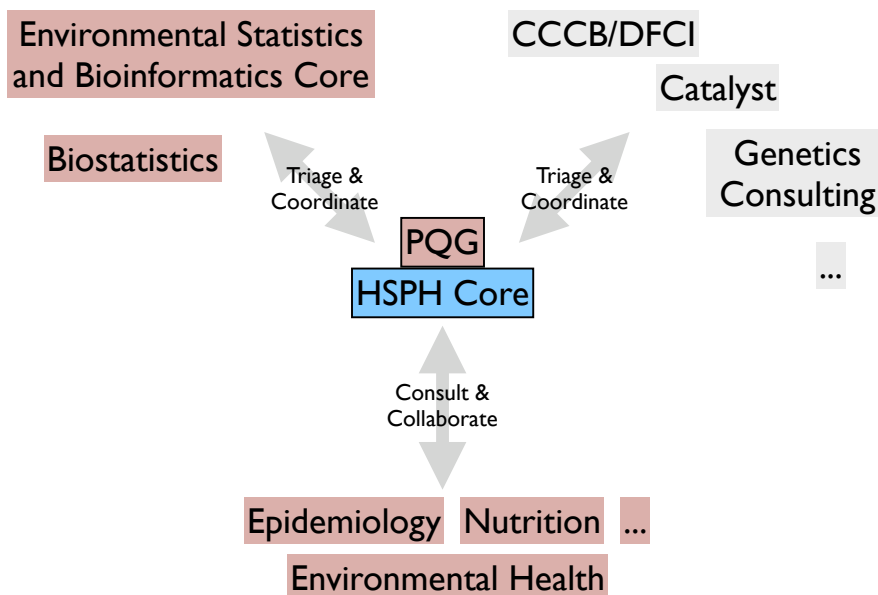
## External Relationships

Services provided by the HBC overlap significantly with the interest and expertise of other Cores and centers such as the CCCB (DFCI), the Catalyst / National Clinical and Translational Science Award Consortium, Genetics Consulting and others. To coordinate infrastructure deployment, methods development and service provision we will grow and where appropriate, formalize the interaction with these groups, by regular face-to-face meetings as well as a mailing list to ensure tight collaboration.

In particular, the HBC and DFCI Core — currently being established by Dr. John Quackenbush, HSPH Professor of Computational Biology — will have their own natural constituencies, based on research focus and geography. While the DFCI Core will serve some non-cancer clients through Catalyst, the bulk of its business will be cancer-focused research which cannot be funded by Catalyst. It is anticipated that over time, each Core will develop its own specialties and identity, based on the needs

of its core constituencies, and will refer clients to each other as appropriate. They will collaborate rather than compete. Policies and consulting rates established by the HSPH Bioinformatics Core will be consistent with those set in the DFCI Bioinformatics Core.

We will identify similar overlapping needs and resources with other Cores and centers, with a triage process ensuring that requests are handled by the center best suited to support a given project.



HBC members will work closely with members of the Department of Biostatistics and the Biostatistics Core, collaborating on service requests requiring (bio)statistical expertise and driving establishment of collaborations.

## Oversight

Dr. Winston Hide will serve as the Director of the HBC and will report to the Dean through the Director of the Catalyst Program, James Ware. Dr. Hide will be responsible for providing the scientific direction of the Core, setting the priorities of collaborative research projects and those undertaken through the support activities of the Core. He will represent the Core within the Catalyst infrastructure.

An internal oversight committee will be established as the Core becomes operational. The members of the oversight committee will be appointed by the Dean. The committee will meet once a year and will:

- evaluate the progress of HBC service activities in related to stated objectives as well as to overall performance expectations for Catalyst Cores
- provide advice regarding the overall scientific direction for HCB and advise on ongoing research support activities

- provide guidance regarding the HCB relationships and collaborations with other HSHP groups, programs, departments and strategic initiatives to ensure the success of the HCB, and to avoid duplication of effort.
- work with HSPH administration to ensure the HCB is meeting expectations of the Dean's Office

The oversight committee will report to the Dean of HSPH.

The establishment of an external advisory board (EAB) providing guidance on scientific matters is planned in order to ensure a close adherence to current best practices and community standards. The EAB will also evaluate the progress of the HCB and provide the HCB with relevant experience and perspectives, with regard to its complementary scientific and research support missions. Members of the EAB will be selected by Dr. Hide with the advice and consent of the HCB Oversight Committee

Up to ten members of the advisory board will be selected from experts in the field of bioinformatics and public health. They will represent leaders in the areas of training, scientific leadership, technologies, genome translation and bioinformatics core management. Key members of the Boston bioinformatics community and others further afield are currently being approached for their willingness to serve on this board.

## Measures of activity and impact

Measures of the success of the HCB will be developed in collaboration with the advisory board and oversight committee. As the Core is not itself a pure research group, measurements will qualify how well the group does in enabling and strengthening ongoing research from other laboratories at HSPH. Sample measurements of impact and activity include, but are not limited to:

- grant funding applied for and grant funding secured
- number of pre-grant projects supported
- number of other consulting support efforts
- hours billed to projects
- number of publications with Core members as co-authors or acknowledged as contributors
- successful establishment of collaboration with other Harvard institutions and Cores
- development of re-usable tools
- database resources made available both locally and publicly
- adherence to community-accepted standards in software development and data handling
- training sessions held (in collaboration with Countway, PQG and other entities)
- creation of a website for information dissemination
- number of users and unique hits for the website and software platform
- hiring of bioinformatics staff (internal and for other HSPH groups)
- degree of recognition external to HSPH, and external to Harvard

## Implementation and Staffing

Our staffing model: Funding and staffing for the Core HBC been approved by the Dean's office (see the organizational chart). We will assist laboratories and programs in hiring, supporting and training their bioinformatics staff as affiliate members of HBC. Affiliated members will be provided with a work environment and be able to draw on the expertise of the Core, but are employed by their respective research programs.

Position	FTE	Shared with	Hired	FTE / Year		
				Year 1	Year 2	Year 3
Core Manager	0,5	Group of Dr Hide, NIEHS	2008	25%	25%	25%
Bioinformatics engineer	0,5	PQG	2008	25%	25%	50%
Bioinformatics engineer	0,5	PQG	2008	25%	25%	50%
Bioinformatics engineer	1			75%	100%	50%
Bioinformatics analyst	1			75%	100%	50%
Data Curator	1			75%	100%	50%
System Administrator	1			75%	100%	50%
<b>Total</b>	<b>5,5</b>					

### The current status of staffing:

Since October 2008 the Bioinformatics Core has hired a Core Manager (Oliver Hofmann) who splits an FTE between research in the lab of Dr Hide and managing the HBC. Funds for managing the Core are provided by the Dean's Office and NIEHS in equal parts. The HBC has also hired two bioinformatics developers (Alyssa Porter and Ray McGovern), with 50% of their time dedicated to and supported by the PQG.

### Our future plans for staffing:

The remaining open positions will be advertised and filled within the next 3-6 months based on the job descriptions listed below. The HBC has financial support from the Dean's Office for an additional bioinformatics engineer, an analyst, data curator and systems administrator until 2011, with the expectation of external funds and consulting fees to cover half of their salaries by 2011.

### Job Descriptions

Core Manager (0.5 FTE, shared with lab of Dr Hide and NIEHS, hired)

The platform manager will have an MD or PhD in the biological or computational sciences and extensive expertise in Bioinformatics and Computational Biology and extensive experience in providing bioinformatics support to laboratory and/or clinical/translational research scientist.

This individual will have responsibilities for managing the day-to-day operations of the core as well as for providing high-level consulting support to the core's clients.

**Bioinformatics Engineer (2.0 FTE, shared with PQG, hired)**

The Bioinformatics Engineers will have a BS or MS in computer science or in the biological or physical science or a related discipline and experience in software development in support of bioinformatics applications. These individuals will be responsible for providing consulting support for the HBC's clients, with a focus on those needing greater assistance in developing new software tools or applications.

**Bioinformatics Analyst (1.0 FTE, to be hired)**

The Bioinformatics Analyst will have a BS or MS in the biological or physical sciences or a related discipline and experience in the use and application of bioinformatics software and tools to address data analysis needs. This individual will also be expected to possess some software development skills in support of developing data analysis solutions. The Bioinformatics Analyst will have primary responsibility for working with HSPH researchers and provide analytical support for available data sets. Based on an initial needs analysis the analyst should have expertise in the handling of array data with a particular focus on supporting genome-wide association studies.

**Systems Administrator (1.0 FTE, to be hired)**

The Systems Administrator will have primary responsibility for managing the computer hardware systems necessary for supporting the computational needs of the Core, including maintaining our computing and database clusters and other computational servers, overseeing our network architecture and firewalls, and coordinating systems support efforts with the School's IT department.

**Data Curator (1.0 FTE, to be hired)**

The data curator will have an extensive background in at least one biomedical field and ideally be experienced with existing community standards for the curation of biological data (typical areas of expertise include Gene Ontology, the OBO ontologies as well as standardized formats such as MIAME, ISA-TAB etc.). The curator will support the developers in implementing software and data exchange formats that are standards-compliant, ensure data consistency of Core-managed external data from research labs at the School and work with international collaborators to develop and extend existing ontologies and standards.

**Hiring Timeline**

<b>Task</b>	<b>Deadline</b>
Decision on requirements for systems administrator, developer, analyst and curator	March 2009
Job descriptions provided to HR and approved	April 2009
All open positions advertised (Nature Jobs, ISCB, and other venues)	April 2009
Interviews and reviews	May 2009
Systems administrator and developer hired	June 2009
Curator and analyst hired	July 2009

## Supported representative studies

We will develop, document and share standard operating procedures in the form of method scripts offering standard analysis for relevant experimental data for scientists comfortable with command-line scripting, enabling them to initiate their analysis with a high confidence set of cleaned up data. Where possible, standard operating procedures will also be offered as part of a web server within the HSPH intranet, allowing trained researchers to quickly apply these workflows to incoming data. Upon request we will work with researchers to identify data management and storage requirements for larger data sets in order to develop optimal solutions that are scalable and can be applied seamlessly to future studies.

Infrastructure developed for the HBC is being made available to all HSPH researchers where possible (with exceptions in the case of privacy restrictions or proprietary models). Software development is coordinated with the PQG, NIEHS and other facilities, and any software developed by the HBC will be documented and made freely available to the community as open-source software where possible.

By the end of the initial funding period the Core will support experimental data from the studies listed below. Support includes data management and storage, development of quality assessment procedures, installation and maintenance of standard analytical software packages, collaborative efforts with researchers within the HSPH community to improve analytical methods and the visualization of data in their biological context where applicable:

- **Genome-wide association studies (GWAS) – SNP:** High-throughput GWAS play an important role in a large variety of public health studies to associate genotypes with disease phenotypes.
- **GWAS – Copy Number Variation (CNV):** We expect an increasing demand to utilize information on structural genetic variation, in particular variation of gene copy numbers likely to affect gene expression levels, as part of ongoing and future GWAS.
- **ChIP-chip/ChIP-seq:** The study of heritable changes such as DNA methylation and regulatory effects (including transcription factor binding) is rapidly becoming a standard analysis, frequently used to complement gene expression studies.
- **Large-scale gene expression analysis (microarray):** Despite an increasing demand in support for next-gen whole-transcriptome sequencing microarrays continue to be a standard tool, in particular to support studies in GWAS, proteomics and other areas.
- **Next-gen sequencing data:** High-throughput sequencing data, whether capturing genome or transcriptome data, continues to grow at a rapid pace.
- **Proteomics:** A number of projects, particularly in the clinical/translational bioinformatics area, utilize proteomic studies to identify large numbers of peptides and proteins of interest.

Integration with contextual biological information is key to support of the above domains, allowing researchers to draw conclusions and test hypothesis based not on the original list of genes or proteins, but their additional features and interactions. Linking SNPs to genes and transcripts, identifying functional and regulatory annotations of significance for gene lists or exploring metabolic and signal transduction pathways associated with a case/control proteomic study are just a small subset of use cases to enhance data generated in high-throughput experiments with biological knowledge. We will provide training and assistance in the application of standard tools licensed at the Countway Library (such as IPA, MetaCore and Explain), but also host our own database system to provide cross-platform data and knowledge integration from different platforms.

Prioritization of data will be based on the needs analysis and current demands of groups at the School. Support for experimental data and procedures associated with grant or project deadlines take priority, followed by support for study types benefiting multiple groups and new researchers at the School.

## IT infrastructure

Computational demands required to process biological high-throughput data continue to rise at an exponential pace likely to outpace current increases in CPU power, in turn driving the increasing complexity and parallelization of bioinformatics algorithms for distributed data processing. Likewise, the size of biological data sets is increasing beyond the parallel decrease in storage cost, requiring a balanced approach between low-cost, long term data archives and more expensive high-performance storage environments used to store actively used data. The increasing data set size also requires processing units with sufficient built-in memory to be able to handle sets of data that can not be trivially broken down into smaller subsets.

The Bioinformatics Core will work with the HSPH IT department to develop solutions that address these challenges. In particular, we plan to install and maintain a high performance multi-core server with sufficient memory to handle large data sets on a single machine. This server will complement the existing high performance cluster that is already operational and in high demand. To alleviate storage cost we will implement a tiered storage model that distinguishes between shared data sets, data in active use and long-term storage requirements and is priced competitively.

The new IT infrastructure will be supported by dedicated IT staff within the HBC who will work in close collaboration with the HSPH's IT group to ensure tight integration of the new servers and storage facilities into the existing infrastructure. Operating system and supported hardware will match already deployed solutions at the School. User accounts, backup strategies, data tracking / version control support and other administrative tasks will be kept consistent between both IT environments. Existing access control mechanisms will be leveraged to enable sharing and collaboration of HSPH-internal data sets among researchers at the School.

## Website

The HBC website serves as an initial point of contact and provides information on:

- current events in computational biology in the Longwood area
- available bioinformatics services and data sets at the Core
- additional services offered by other cores and centers
- tutorials and documentation on any self-service tools provided
- an archive of software and algorithms maintained by the Core
- a help desk and contact form for HSPH researchers
- an overview of consulting fees and models
- links to project management sites, knowledge bases and wikis for individual projects

Hosted bioinformatics services will be deployed on the same web server where feasible. These will be limited to members of Harvard University and require a Harvard ID and password.

## Teaching & Outreach

The Bioinformatics core will provide support to external training sessions and seminars and help advertise events in computational biology and bioinformatics in the Longwood area. Specialized training courses for software licensed or supported by the Core will be coordinated with existing training sessions provided by the Countway Library. The HBC also supports the Bioinformatics Forum, a monthly seminar series providing an overview on current questions and methods in the field of Computational Biology with a focus on their application to support studies in public health.

We offer ad hoc training for supported tools and databases to researchers at the School as part of our regular consultation system. In addition, we will organize workshops and short-courses on standard bioinformatic tools and software such as R, Bioconductor or Galaxy in collaboration with the Genetics Consulting group and Catalyst.

# Appendix

## Finances

The HBC and DFCI Bioinformatics Core (CCCB) will coordinate efforts, thereby achieving economies of scale and ensuring uniform policies with regard to federal funding. Rates charged by the HSPH Bioinformatics Core (and by extension the DFCI Core) will be consistent with those charged by bioinformatics cores at other leading universities.

- Most universities subsidize their Bioinformatics Core rates, presumably because part of the infrastructure associated with providing these services is built into their F&A rates.
- We believe charging higher hourly rates to federal grants at Harvard than generally prevailing rates at other universities would invite federal scrutiny.

For compliance reasons, pre-award work cannot be subsidized by funded projects — rates charged to grants cannot be set artificially high by building in costs related to pre-award work.

- Most HSPH researchers do not have funding available for pre-award services (initial data analyses, help with preparing competing grant proposals, etc.), and HSPH will need to support that work in the long term.
- While it is expected that most clients of the HBC will be based in the Longwood Medical Area or Cambridge, in the future there may also be external clients from private corporations. Rates charged to external clients may be set higher, and thus subsidize the rates charged to Harvard-based researchers.
- Long-term support from HSPH will involve both consultant-hours and IT infrastructure support. In some cases, research requests may require the Bioinformatics Core to build capacity, for example by purchasing new biological databases, solving new IT problems relating to confidentiality of data (as was the case with the Framingham Heart Study data), or purchasing additional storage for massive databases.
- In FY11, the final year of the current funding agreement for the Bioinformatics Core, the Dean's Office will need to decide how much pre-award consulting work it will support in FY12 and beyond. The Bioinformatics Core will set policies regarding the level of pre-award help it can provide to HSPH researchers accordingly.

Rates charged to the HSPH Core's "clients" will be consistent, yet still take into account peculiarities associated with each type of funding source

Currently, funding to support HBC activities is provided by the Dean's Office, the NIEHS Center Grant, and Catalyst. If possible, each source would be billed an hourly rate for services, plus, in some cases,

a percentage of effort for higher-level, sustained involvement. As the Bioinformatics Core and Program in Quantitative Genomics are both in early stages of development, there is a certain amount of fluidity between their activities and staff. However, in the long run, the PQG would most likely be billed as a “client” for bioinformatics services.

- Budget, Revenue Assumptions, Expense Assumptions: The current 3-year budget for the Core assumes a gradual shift of funding from resources provided by the Dean's Office towards financial independence, but does not impose any restrictions on how this can be achieved. To maintain a high amount of flexibility we will adopt a chargeback model based on fixed fees for different tasks (general consulting, training, infrastructure development), with initial project analysis' and preliminary grant support as well as general administrative expenses being covered by the Dean's Office.
- Policies: We will develop standard operating procedures to prioritize projects and provide estimated costs for different services. This includes an overview of common (sample) tasks in each category as well as providing boundaries to what services the Core is not equipped to handle (e.g., generic IT infrastructure development). Policies will include expected input (specifications provided by researchers), standard outputs / deliverables and evaluation criteria (feedback forms, timely delivery of milestones, etc). Policy development will be a joint effort with CCCB management and is expected to be supported by a Catalyst Research Navigator to ensure consistency with standards developed at other groups and universities.

## Image information

- Front page: DNA sequence, “genome” (<http://www.flickr.com/photos/hydepodcorner/1411610758/>)
- Front page: Ethernet, “Core B&W” (<http://www.flickr.com/photos/mightyboybrian/126630218/>)
- Front page: Circos Genome Visualization (<http://mkweb.bcgsc.ca/circos/>)