Bioinformatics Core Design Proposal

This manuscript (permalink) was automatically generated from alemenze/Core_Proposal@d2726d2 on January 22, 2021.

Authors

- **Alexander Lemenze**
  0000-0003-3053-0849 · alemenze
  Department of Pathology, Immunology, and Laboratory Medicine, Rutgers University- New Jersey Medical School; Center for Immunity and Inflammation, Rutgers University- New Jersey Medical School

Bioinformatics Core facility need

What is Bioinformatics

“A hybrid science that links biological data with techniques for information storage, distribution, and analysis to support multiple areas of scientific research, including biomedicine. Bioinformatics is fed by high-throughput data-generating experiments, including genomic sequence determinations and measurements of gene expression patterns. Database projects curate and annotate the data and then distribute it via the World Wide Web.” – Encyclopedia Britannica

Need for a Core Facility

Data generation in biomedical and life science research is exponentially increasing in volume and scale. This requires a strong support network to facilitate the acquisition, management, analysis, visualization, and sharing of data. A Core facility can offer a shared resource to provide bioinformatics expertise, reduce overall hardware and personnel costs, and minimize duplication events.

Overview of Core Facility Design

Core facilities have been implemented with a variety of designs depending upon their physical needs and the particular establishment. A multitude of models have been used for designing bioinformatics core facilites. Often these end up swinging towards the poles of “biology” or “informatics” that must come together for an effective implementation. A few implementations are described below, followed by the proposed implementation.

1. Diffuse model

A diffuse bioinformatics model is what often occurs without a semi-centralized or centralized effort. This avenue entails individual laboratories hiring and maintaining bioinformatics staff/equipment. As each laboratory is responsible for the staff, they are intimately familiar with the associated biology. This intimate association with the biology enables the diffuse bioinformaticians to work in an isolated environment and operate with full expertise of the project. Inherent in this model is increased costs for bioinformatics across the university, as individual laboratories each have increased staffing, and each will purchase and maintain their own informatics tools- such as local servers, HPC units, and software licenses. Additionally, as these bioinformaticians are focused on the biological questions at hand, they more often are entrenched in a canonical academic career path within that domain-specific field, not within a bioinformatics service effort or assisting on projects outside of their specialty.

2. Pure Centralized model

A purely centralized model would be designed to build a pillar of bioinformatics at a central site. Though this can potentially be very strong, it also requires the largest buy in- both financial and investigator usage. As this type of model is often integral in developing infrastructure to facilitate high throughput data, this model can skew towards the informatics aspects. Inherent in this model is a lower overall cost of bioinformatics across the university, as the centralized core can leverage economy of scale for purchasing power for informatics tools. On the flip side, this model is often the least biologically focused, and can create a divide between the biologists and informaticians instead of uniting as bioinformaticians.

3. Semi-Centralized model **(Proposed model)** The Core facility proposed herein shall be a semi-centralized effort. A semi-centralized design incorporates the strengths of both previous models by integrating into an overarching centralized infrastructure and providing branches of staff into semi-specialized workflows. One major downside to this model is requiring a centralized infrastructure team, which fortunately in this instance is already in place with the Office of Advanced Research Computing. By nesting the bioinformatics core within the existing infrastructure, the semi-centralized model gains the benefits of lower operating costs and existing informatics knowledge. The bioinformatics team can then focus on tuning the informatics to fit biological questions. With this, the bioinformatics team will have the availability to branch into data-acquisition specialities (such as genomics/proteomics etc described below), to then build a strongly collaborative environment with domain-specific investigators.

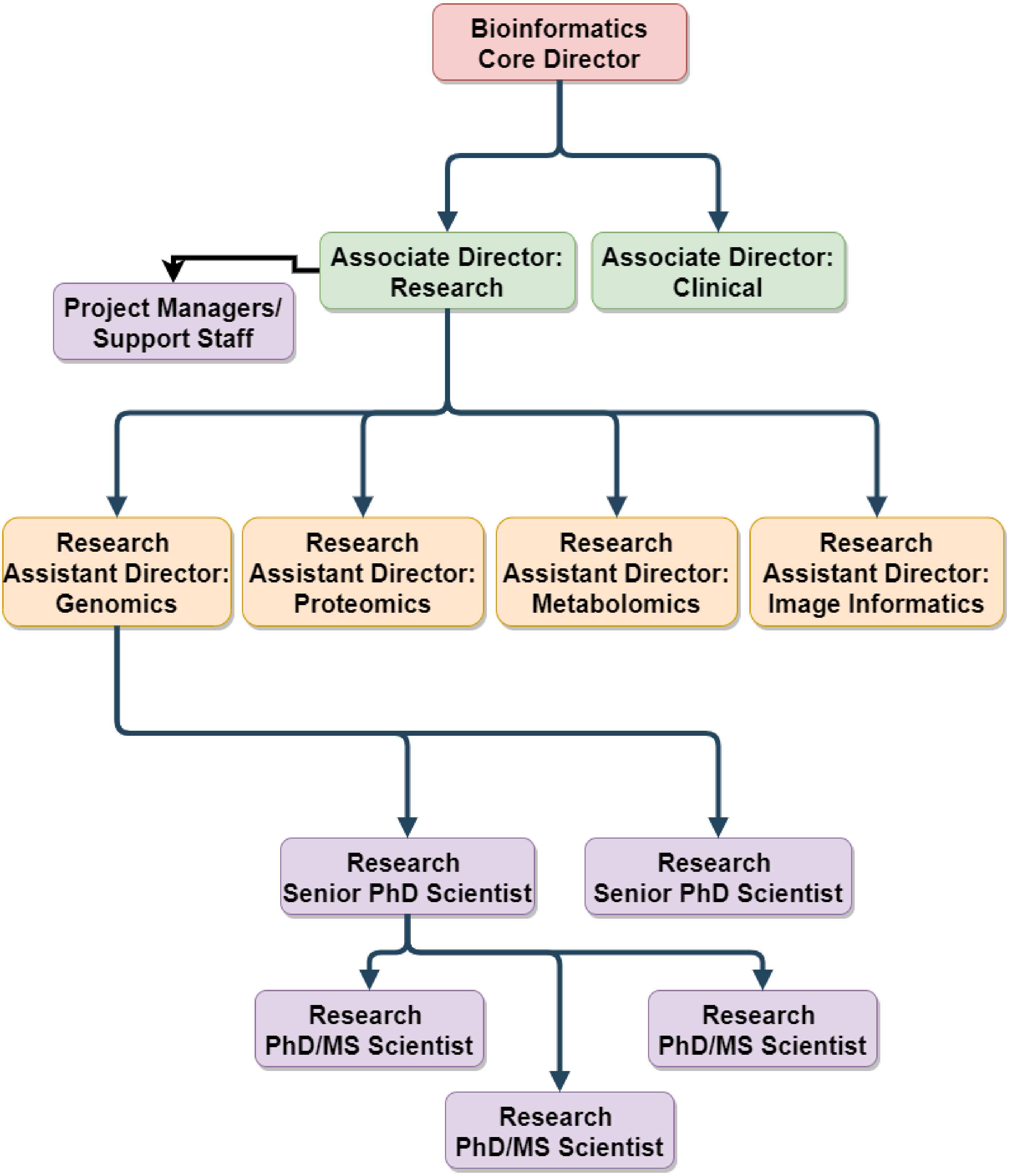
Bioinformatics Core Mission

TLDR: To continually evaluate best-in-class bioinformatics solutions, implement bleeding-edge bioinformatics techniques, and offer training of basic bioinformatics methods to advance reproducible analysis of biological data.

Core facilities serve as a source of expertise in a particular field/topic to be shared amongst investigators. In the case of bioinformatics, the core facility is to provide expertise in computational analysis of biological datasets encompassing the fields of genomics, proteomics, metabolomics, and imaging based datasets. This includes operational know-how for computational tools, balancing hardware requirements for computational tools, benchmarking comparative computational tools, understanding the data generation techniques, and executing analyses. By leveraging the Core's expertise with these topics, investigators will be enabled to generate biologically driven discoveries.

Core Organizational Structure

Core organization shall follow basic principals of ICS. Staffing shall be filled from the top down as demand is generated for specific roles. All roles will maintain the responsibilities of lower branches until subsequent steps are implemented. For example: If no Research Assistant Director for Metabolomics is assigned, the Associate Director of Research will bear the responsibilities for Metabolomics. At any stage if >7 staff are reporting up a tier to a single staff member an additional tier must be implemented. As within emergency services, this design is inherently flexible for scaling up or down as needed.


Bioinformatics Core Director
Associate Director: Research
Associate Director: Clinical
Project Managers/ Support Staff
Research Assistant Director: Genomics
Research Assistant Director: Proteomics
Research Assistant Director: Metabolomics
Research Assistant Director: Image Informatics
Research Senior PhD Scientist
Research Senior PhD Scientist
Research PhD/MS Scientist
Research PhD/MS Scientist
Research PhD/MS Scientist

Example for a full administration build out of organization

Core Revenue Streams

This Core will focus on 4 primary revenue streams to both cover operational costs and provide for growth of the Core. These four revenue streams are designed to operate in synergy to meet the mission of the core.

All project consultation will be provided free of any charge.

Fee per operation Services

Fee per operation services will be primarily prescribed functions and automated for maximum efficiency. These will involve upfront time investment, but long term will provide staff availability to focus on collaborative services focused on biological interpretation.

Primary Data Analysis

The first facet of fee per operation would be primary data analysis. Many raw data outputs are natively human-unreadable. The primary data analysis will at minimum convert machine raw data to common data formats and/or human readable data structures. This will become a standard add-on to all data generating facilities services, and will be priced as an absolute minimum- covering purely the computational costs associated with the processing. These will be designed for standardizing output of data for other pre-defined pipelines, ease of staff use for collaborative usage, and input to SaaS tools.

Example: Transcriptome analysis - bcl to FASTQ conversion and demultiplexing to individual samples for NGS approaches - Transcriptome alignment from FASTQ to bam files - Differential gene expression analysis to human readable CSV (excel) files.

Pre-defined Pipelines

The second facet to fee per operation would be some automated secondary data analysis. This is comparable to many contract research organization (CRO) offerings for analysis. All outputs will be explicitely pre-defined.

Example: Transcriptome analysis + visualizations - bcl to FASTQ conversion and demultiplexing to individual samples for NGS approaches - Transcriptome alignment from FASTQ to bam files - Differential gene expression analysis to human readable CSV (excel) files. - Principal component analysis and global hierarchal clustering - Per-comparison volcano plots - Heatmaps for top differentially expressed genes - Basic Pathway analysis

Training Services

Training services will be offered to increase the bioinformatics education level of university members. In alignment with traditional open-source values, **all** training materials will be made open-access. Charges will only be incurred for staffed courses.

Prescribed courses

Prescribed courses will be developed for standard bioinformatics education. These will involve an initial time investment, but once developed will require minimal maintenance to stay up-to-date. Additionally, these courses can be made in-person, virtual, or interactive depending upon the topic involved. These will be offered at a regular interval, and potentially could grow in to a revenue stream of degree courses.

Spot courses

Spot courses will be offered at a higher rate than prescribed courses. These will be custom designed for investigators requests on bioinformatics techniques. For example, if a laboratory wishes to learn more about alignment algorithms but no prescribed course covers these topics, one will be developed and provided to the investigator’s group.

Software as a Service (SaaS)

Software as a Service is designed to enable investigators to perform biological analyses using informatically developed tools. There are an increasing number of companies exclusively designed around SaaS for bioinformatics, such as Partek, Basepair, Rosalind. This will be an in-house developed effort, designed both for customization and individualization of tools, as well as reduced costs compared to corporate efforts. The key is the core provides the bioinformatics expertise to manage the tools, the parent organization (IE OARC) can provide infrastructure, and the investigator provides their biological expertise.

Access subscription

This will generate revenue in a subscription style. Individual investigator's labs, departments, or major centers can purchase subscriptions for their members.

Examples: - RNA-seq Shiny app. Investigators can take differentially expressed gene matrices (generated automatically in primary processing pipeline), and interogate their data at further depth with tools designed by bioinformaticians to enable their publication quality plot of data.

Collaborative Services

Collaborative services will be the primary time utilization of core personnel. With collaborative services core staff will directly interface with investigators, pairing the core staffs computation expertise with the investigators domain-specific knowledge. This collaboration should be proactively discussed to set expectations for the project. Core personnel will be often expected to learn portions of the domain-specific knowledge, but should **not** be expected to become domain-specific experts. That should remain the responsibility of the investigators, and is why this is a collaboration, not the diffuse bioinformatics model that requires domain-specific bioinformaticians.

Grant percent Effort

As a collaborative service, grant percent effort is when Core staff are covered as primary personnel on sponsored research. The percentage of effort should be commesurate with the expenditure of time and higher-level analyses. For ease, this can be targeted at a scale of 20% increments- with each 20% corresponding to 1 day/week of the staff members time dedicated to that specific project. At minimum, it is highly recommended for bi-weekly meetings between the Core staff and investigator to ensure alignment of workflows and timely progress.

Department/Unit support

Department/Unit support is designed to provide collaborative services to non-sponsored research. This often will encompass investigators side projects, development of tools/workflows for investigators usage, and most importantly preliminary data for grant applications. This should begin the integration of Core staff in to a project, ensuring appropriate data practices are being followed to strengthen the application as well as encourage applications.

References