Summer August 22, 2016

Supporting Biomedical Research in the Era of Omics and Precision Medicine

Rolando Garcia-Milian
Denise Hersey, *Yale University*
Nathan Rupp

ANNUAL REPORT 2015-2016

SUPPORTING BIOMEDICAL RESEARCH IN THE ERA OF OMICS AND PRECISION MEDICINE

Yale *Harvey Cushing / John Hay Whitney Medical Library*

END-USER BIOINFORMATICS SUPPORT PROGRAM

ANNUAL REPORT 2015-2016

YALE HARVEY CUSHING/ JOHN HAY WHITNEY MEDICAL LIBRARY

Rolando Garcia Milian, Biomedical Sciences Research Support

Denise Hersey, Clinical Support Librarian

Nathan Rupp, Head of Collection Development & Management

Lei Wang, Instructional Design Librarian

Contents

Summary

This annual report (2015-2016) provides a continuing view on the position of the Cushing/Whitney Medical Library End-user Bioinformatics Program. Besides the report on the three main areas of training, resources and tools, and consultations, it contains the results of the recent assessment "Information and Needs Assessment for Biomedical Research in the Omics Era" During this period, 741 Yale affiliates attended (out of 1240 registered) the end-user bioinformatics training and presentations organized by the Medical Library. This year, the number of Ingenuity Pathway Analysis and MetaCore accounts continued to grow. Consequently, the number and length of research support consultations (130 researchers benefited) on these tools also increased. In addition, the Medical Library added the professional version of the Human Gene Mutation Database, and Biocyc database, and it has begun license Partek Flow as part of a pilot project to support the secondary stage of the omics high throughput data analysis cycle. According to the aforementioned assessment, the data and information needs differ between graduate students, postdocs, and faculty. These and other results are discussed in this report.

Introduction

High throughput technologies such as next generation sequencing, microarray, and mass spectrometry, are rapidly generating high amounts of diverse omics data (see Figure 1). Although this offers a great opportunity for discovering the molecular basis of diseases, it represents significant challenges in terms of information and knowledge extraction from these data.

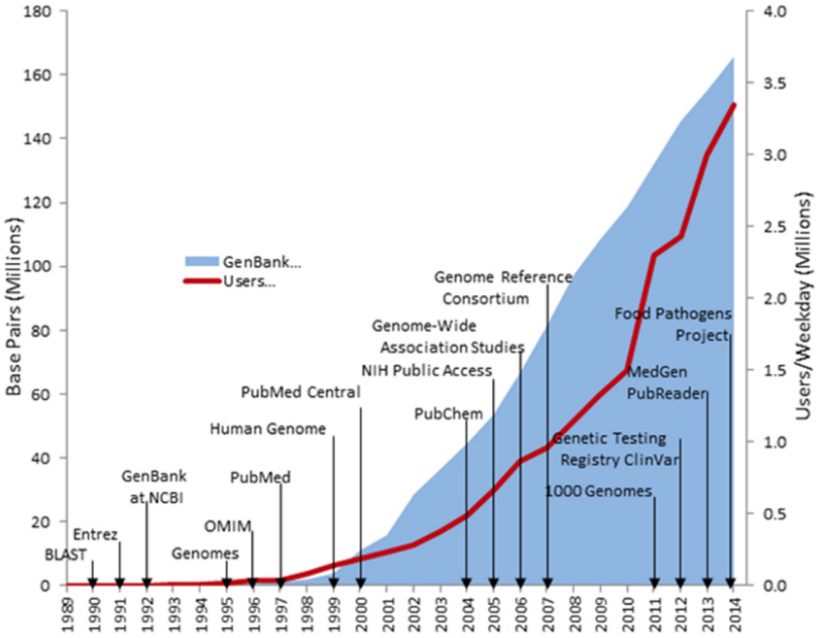


Figure 1. Increase in the amount and diversity of data and databases. NLM Congressional Justification FY 2016 https://www.nlm.nih.gov/about/2016CJ.html

Besides the rapid generation of diverse omics data, some of the main challenges faced by biomedical researchers are:

- Exponential grow of biomedical literature (e.g. PubMed has more than 26.3 million records). In 2009, a study showed that one third of PubMed searches resulted in 100 or more records (Dogan, Murray, Neveol, & Lu, 2009).
- Integration of these diverse omics (genomics, proteomic, metabolomics, etc.) data (Wang et al., 2015).
- Analysis and visualization of networked data. The majority of the omics studies include a pathway or network analysis of the data (Villaveces, Koti, & Habermann, 2015).
- Low reproducibility of experimental research (Begley & Ioannidis, 2015).

- Translating omics information into human health benefits. Despite predictions and progress in research, the introduction of omics into clinical practice has been slow (Burke & Korngiebel, 2015).

In order to provide relevant services, avoid duplication, and foster collaboration with other units on campus, it is important to understand the Research cycle of high-throughput omics data at Yale University. The cycle consists of the following stages (see Figure 2):

- The **primary data analysis** is performed by the sequencer instrument which transfer the signal generated to the ACTG code and where the calculation of the quality indicators for the data takes place. It generates a huge FASTQ file. This step takes place mainly at the Core units (e.g. Yale Center for Genome Analysis, Yale Stem Cell Core, Proteomics Core, etc.)
- The **secondary data analysis** is the arrangement of the obtained fragment sequences alignment of reads and their assembly. This is done by bioinformaticians/ biostatisticians in a High Performance Computer (HPC) environment.
- The **tertiary analysis** is the final stage of data analysis and provides clues on the biological significance of the data (relevant networks, pathways, biomarkers, etc.). These are small files (e.g. Excel, text files, etc.). The Yale Medical Library is currently providing support for this stage of the data life cycle in terms of resources, training, and consultations.

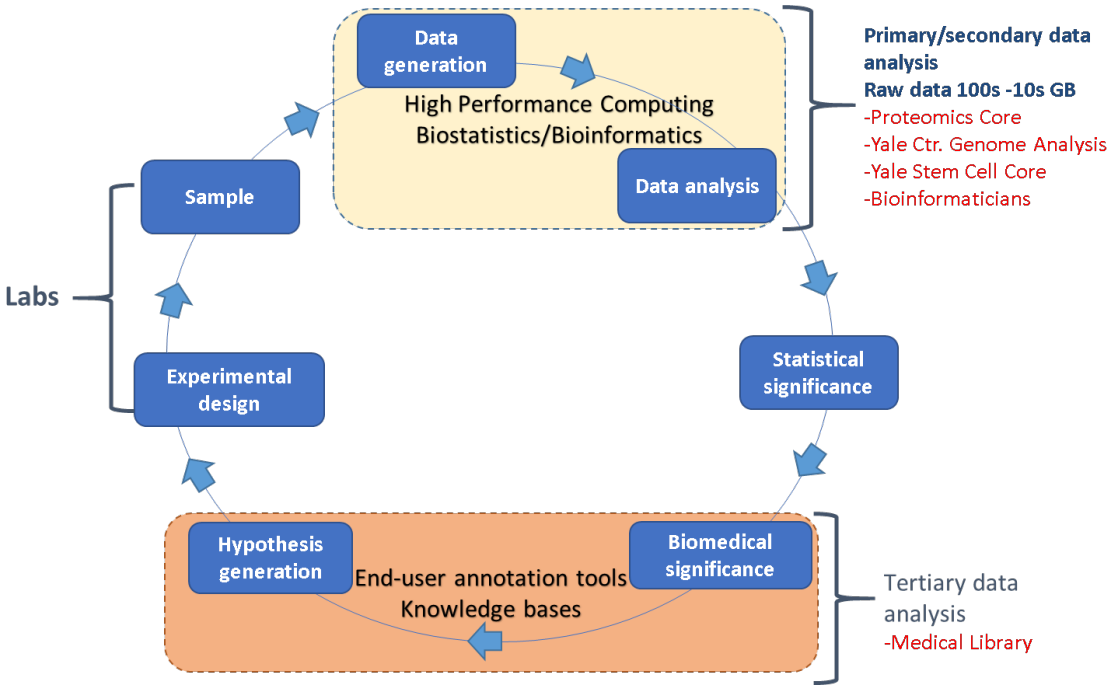


Figure 2. Research cycle of high-throughput omics data showing the position of the Medical Library in support of tertiary data analysis stage.

When designing the trainings and providing the support, it is important to recognize that while bioinformatics experts tend to push for more mathematical and statistical content, biomedical researchers, generally not formally trained in computing, prefer an end-user tool approach to bioinformatics (Tan, Lim, Khan, & Ranganathan, 2009).

Precision medicine use of high throughput omics (e.g. genomic, transcriptomic, proteomic, metabolomics, etc.) for the characterization of patients in order to inform medical decisions. Ultimately, it allows physicians to tailor treatment based upon the molecular profile of a patient's disease and to develop diagnostic tests to predict who is at risk for certain diseases. In addition, work in this field should result on the development of companion diagnostics tests for the safe and effective use of a therapeutic product (Vicini et al., 2016). Precision medicine entails compiling big data on individuals, and then combining that data and analyzing it to determine how it can be transformed from the bench to the bedside. Scientists and clinicians working in this endeavor will require new tools and support to make this initiative successful and beneficial to patients.

Consequently, in President Obama's 2015 State of the Union Address, he introduced a new initiative to support research for precision medicine, sometimes labeled "personalized medicine." (https://www.whitehouse.gov/precision-medicine). To this end, the President has included funding in his 2016 budget; $215 million for the National Institutes of Health (NIH), along with the Food and Drug Administration (FDA), and the Office of the National Coordinator for Health Information Technology (ONC) to support work in precision medicine. The National Cancer Institute (NCI) is receiving $70 million of this funding to help identify genomic drivers in cancer and translate that information into cancer treatments (Collins & Varmus, 2015).

Trainings and Presentations

Results from a recent survey[1] (see Information Needs of Biomedical Researchers section) show that not having adequate training is one of main challenges that Yale biomedical researchers face when analyzing their data (see Figure S5). Based on this, the Cushing/Whitney Medical Library will continue to provide end-user bioinformatics training including those developed in-house, and in collaboration with other campus units, outside partners (e.g. National Center for Biotechnology Information, Mouse Genome Informatics, etc.), and other experts.

The concepts of *training* and *end-user* are used here as previously defined by Schneider (2010). Training refers to a short session aimed to deliver skills that allow the attendee to use bioinformatics resources and tools. End-user refers to a user who accesses these resources and tools through a Graphical User Interface (GUI). This concept is different from the bioinformatician or computational scientist who actually develop the resources and tools (Schneider et al., 2010)

Dr. Cooper presenting at the NCBI Regional Workshop. April 5, 2016 Yale School of Medicine, SHM, Room C-103

Besides meeting the information and data literacy needs of biomedical researchers, in-house end-user bioinformatics trainings sessions were developed as marketing tools to highlight services and resources provided at the Medical Library in support of basic biomedical research; increase the visibility of the Biomedical Sciences Research Support Librarian and showcase his expertise; and present him as a potential partner in the research process. During this year, 192 individuals registered for the medical library in-house bioinformatics training sessions with 113 attending (Table 1). The success of this approach can be measured by the significant increase not only in the number of consultations but also in their length and complexity compared to FY2014-15 (see Figure 5 in the Consultations section).

Overall, 741 individuals out of 1240 registrations attended the bioinformatics sessions organized by the Yale Medical Library in FY 2015-16. This represents a slight increase in

[1] "Information and Needs Assessment for Biomedical Research in the Omics Era" (HSC# 1511016778)

both the number of registrants and attendees (198 and 153 respectively) compared to FY 2014-15 (see Figure 3).

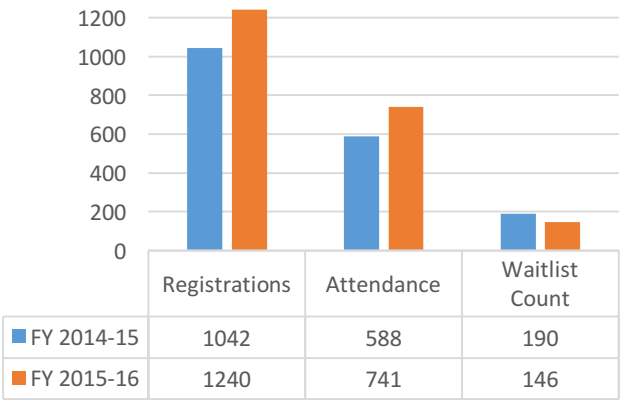
Besides the in-house bioinformatics sessions, a number of presentations and training sessions on commercial bioinformatics software were offered during the year. These sessions expose biomedical researchers to novel bioinformatics tools that may be of help in their workflows. They also help librarians decide which, if any, to license. In addition, we coordinated several trainings on the tools licensed by the Medical Library (e.g. TRANSFAC, IPA, and MetaCore) (Table 2)

"The classes that you have given us are all highly valuable. I am using some of them to solve the problems in my experiments and clinics well."

Visiting Research Scientist in Obstetrics, Gynecology, and Reproductive Sciences. Nov. 13, 2015

This year the Medical Library hosted a two-day National Center for Biotechnology Information's Regional Workshop at the Yale School of Medicine. (Table 3)

A



B

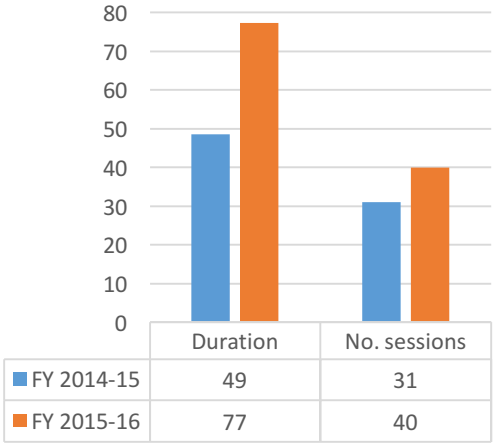


Figure 3. Training and presentations: comparison between FY2014-15 and FY2015-16. A) Registrations, attendance and waitlist. B) Duration and number of sessions offered.

When selecting the training topics, it is important to consider that the large amount of data generated by high-throughput technology and the exponential growth of biomedical literature require new strategies for extracting hidden information. Biological networks showing interconnected elements are suitable for discovering hidden biological information behind omics data as well as in the large amount of text of the biomedical literature. (Kim, Kim, & Lee, 2010). In a network of interconnected data the nodes represent knowledge (e.g. facts, observations, structures, behaviors, etc.) and the edges represent relationships between pieces of knowledge (Figure 4). Hence, it is not surprising that the majority of omics studies contain a network analysis. One of the most popular tools for network analysis is Cytoscape an open source software that allows networked data visualization, exploration, manipulation, and analysis (Villaveces et al., 2015).

Cytoscape training session March 3, 2016. Yale School of Medicine, Room C103

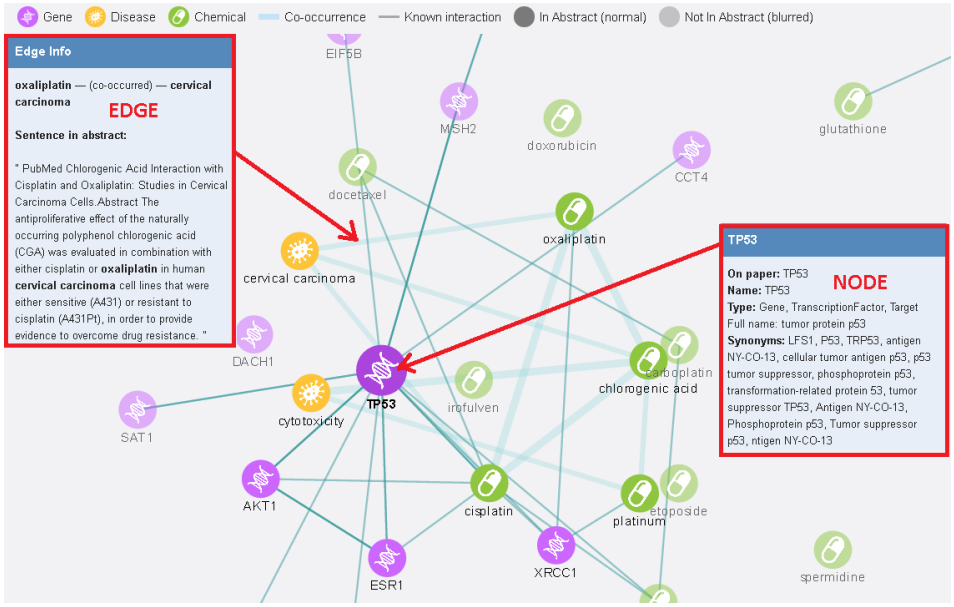


Figure 4 Network representation of a PubMed record using HiPub application. Nodes are circles and edges the lines connecting these. The node 'TP53' is highlighted as well as the edge (relationship) between the nodes 'cervical carcinoma' and 'oxaliplatin'.

During FY 2015-16, a webinar and two on-site training sessions on Cytoscape were offered for Yale biomedical researchers. This was part of a collaboration project with the Taubman Health Sciences Library, University of Michigan. The two on-site sessions were supported by the Jay Daly Technology Grant from the North Atlantic Health Science Libraries. Out of 77 registered, 42 attended.

User feedback was also carefully considered when deciding what training to offer. According to a recent survey administered by the Medical Library, there is a perceived need among biomedical researchers that basic programming and Unix command-line skills are important for the analysis of high-throughput omics data (e.g. Python, R, Unix, etc.). See the "Information Needs of Biomedical Researchers" section. Consequently, on June 4, 2015, Lei Wang, the Instructional Design Librarian, began to offer the "The Very Basics of the Unix Command-Line" class. This popular hands-on class was offered 5 times in the Medical Library computer classroom during FY 2015-16 and was taken by 61 out of 136 registered attendees. Registration was usually full within minutes of announcement and had long wait lists. Basic Unix skills are a prerequisite to take more advanced bioinformatics courses (e.g. Canadian Bioinformatics Workshops https://bioinformatics.ca/ , Cold Spring Harbor Courses & workshops http://meetings.cshl.edu/courseshome.aspx), which may also explain the popularity of this class.

"I have been enjoying the workshops, seminars and training sessions you organize with/for the Medical library. I think you cover most if not all topics and areas of biomedical information"

Associate Research Scientist
Yale University School of Medicine
Department of Internal Medicine
Section of Cardiovascular Medicine.
Sept. 25, 2015

Table 1. Summary of the in-house bioinformatics training sessions offered at the Medical Library - FY2015-16

Title	Seats	Registrations	Attendance	Waitlist
Novel Online Tools for Mining the Biomedical Literature	27	27	15	0
BioMart: A Research Data Management Tool for the Biomedical Sciences	24	17	6	0
Tools for Enrichment Analysis	26	26	23	1
Introduction to Genome Browsers	24	24	19	2
My Bibliography and SciENcv: grant reporting, compliance and biosketch through MyNCBI	24	7	3	0
Making Sense of Genomic Variation: SNP Annotation	24	24	10	0
Tools for gene enrichment analysis	20	20	11	10
Introduction to Genome Browsers	20	20	9	5
BioMart: A Research Data Management Tool for the Biomedical Sciences	20	13	8	0

Novel Online Tools for Mining the Biomedical Literature	20	6	5	0
My Bibliography and SciENcv: grant reporting, compliance and biosketch through MyNCBI	15	8	4	0

Table 2. Summary of the presentations and trainings on commercial bioinformatics software at the Medical Library - FY2015-16

Title	Presenter	Seats	Registrations	Attendance
Introduction to Ingenuity Pathway Analysis	Dr. Kate Wendelsdorf, QIAGEN Informatics	68	64	57
Advance training on Ingenuity Pathway Analysis: Integrated Analysis and Interpretation of Variant and Gene Expression Data from Breast Cancer Subtypes	Dr. Kate Wendelsdorf, QIAGEN Informatics	60	50	38
MetaCore WORKSHOP: Enabling Systems Biology Research Through Pathway Analysis	Matthew Wampole, Bioinformatics Solution Specialist Discovery Science (MetaCore)	120	56	47
Interpretation of variants from human next-generation sequencing studies using Ingenuity Variant Analysis	Matthew Wampole, Bioinformatics Solution Specialist Discovery Science (MetaCore)	50	35	24
Golden Helix Webinar: SNP GWAS and Whole Exome Analysis	Ashley Hintz, Field Application Scientist, Golden Helix, Inc	24	24	27
Introduction to Ingenuity Pathway Analysis	Field Scientist QIAGEN Informatics	51	37	28
MetaCore: Getting the most from your "omics" analysis (Introductory session)	Matthew Wampole, Bioinformatics Solution Specialist Discovery Science (MetaCore)	47	47	32
MetaCore: Getting the most from your "omics" analysis (Advanced)	Matthew Wampole, Bioinformatics Solution Specialist Discovery Science (MetaCore)	45	34	17
Using a Variant Analysis Tool to Study Rare Mendelian Disorders	Bryn D. Webb, MD	40	32	17
Introductory Workshop to MetaCore and Key Pathway Advisor – Pathway Analysis of “Omics” Data	Deborah Riley, PhD, Senior Solution Scientist – Thomson Reuters Life Sciences	45	44	27
Start-to-finish Analysis Software for NGS & Microarray Data	Dr. Eric Seiser, Field Application Scientist, Partek Incorporated	55	55	32
Ingenuity Pathway Analysis Hands On Training	Devendra Mistry, PhD, Field Application Scientist, Ingenuity Products, QIAGEN	55	49	32
CLC Genomics Workbench	Devendra Mistry, PhD, Field Application Scientist, Ingenuity Products, QIAGEN	50	40	27

Table 3. Summary of the invited presentations on bioinformatics offered at the Medical Library - FY2015-16

Title	Presenter	Seats	Registrations	Attendance
Webinar: Introduction to Cytoscape: network visualization software	Marci Brandenburg, Bioinformationist, Taubman Health Sciences Library, Univ of Michigan	24	24	22
Webinar: Introducing SmartBLAST a Rapid Protein Identification Tool	National Center for Biotechnology Information (NCBI)	24	13	9
Hands-on workshop in mouse genetics and Mouse Genome Informatics	Dr. Joanne Berghout, Outreach Coordinator, Mouse Genome Informatics, The Jackson Laboratory	50	25	15
Cytoscape: Going from Raw Data to a Publishable Image	Marci Brandenburg, Bioinformationist, Taubman Health Sciences Library, Univ of Michigan	45	41	16
Cytoscape apps with a Focus on MetScape	Marci Brandenburg, Bioinformationist, Taubman Health Sciences Library, Univ of Michigan	40	12	4
Nat. Ctr. Biotech. Info. workshop: A Practical Guide to NCBI BLAST	Dr. Peter Cooper, Staff Scientist, National Center for Biotechnology Information (NCBI)	46	44	26
Nat. Ctr. Biotech. Info. workshop: Accessing Genomes, Assemblies and Annotation Products	Dr. Peter Cooper, Staff Scientist, National Center for Biotechnology Information (NCBI)	55	53	35
Nat. Ctr. Biotech. Info. workshop: Accessing NCBI Human Variation and Medical Genetics Resources	Dr. Peter Cooper, Staff Scientist, National Center for Biotechnology Information (NCBI)	55	48	24
Nat. Ctr. Biotech. Info. workshop: Exploring Gene Expression Information at the NCBI	Dr. Peter Cooper, Staff Scientist, National Center for Biotechnology Information (NCBI)	65	63	28

Resources and Tools

The use of command-line or scripting languages for omics data analysis requires substantial learning curves for non-programmers. The actual process of developing, debugging and maintaining scripts, and even learning basic scripting languages to analyze omics data can be laborious and time-consuming. It presupposes an availability of time and an interest in programming that many biomedical researchers do not have (Kumar & Dudley, 2007). This is why the development of user-friendly tools has been seen as a major need in omics research (Gomez-Cabrero et al., 2014). Kumar and Dudley (2007) promote the development of graphical user interfaces that enable biologists to analyze and visualize biological data in a biologist-centric mode, rather than large, cryptic ASCII text files (Kumar & Dudley, 2007).

"Thanks for all of your efforts in helping us have access to the best tools for these downstream analyses"

Assistant Professor in the Child Study Center and of Psychiatry Yale Child Study Center. May 31, 2016

During FY2015-16 the Yale Medical Library has evaluated a series of tools for the analysis of omics data (Table 1). Biomedical researchers play an active role during the evaluation period by attending vendor presentations and ultimately deciding whether a tool is relevant and will be supported or not by the library.

Table 4. Resources evaluated and licensed by the medical library. Strikeout text indicates that the software was not licensed after evaluation.

Commercial bioinformatics software	Licensed	Seats	Use
BIOBASE TRANSFAC (QIAGEN)	YES	Unlimited	Transcription Factor
BIOBASE Proteome (QIAGEN)	YES	Unlimited	Knowledge base
BIOBASE Human Gene Mutation Database (QIAGEN)	YES	Unlimited	Variation analysis
Ingenuity Pathway Analysis (QIAGEN)	YES	2 Concurrent	Knowledge base, functional analysis
MetaCore (Thomson Reuters)	YES	Unlimited	Knowledge base, functional analysis
~~Golden Helix~~	~~NO~~	~~0~~	~~Variation~~
~~Ingenuity Variant Analysis~~	~~NO~~	~~0~~	~~Variation~~
Partek Flow (Partek Incorporated)	YES	2 Concurrent	Next Gen Seq Analysis
~~CLC Bio Workbench (QIAGEN)~~	~~NO~~	~~0~~	~~Next Gen Seq Analysis~~

The commercial bioinformatics software Golden Helix, Ingenuity Variant Analysis, and CLC Bio Workbench were evaluated by the medical library but not licensed. Among reasons for not licensing these software are: inadequate cost/benefit, not enough number of users interested, inadequate infrastructure for installation, licensing of an equivalent software.

As of June 3rd, 2016, 303 Yale affiliates have an Ingenuity Pathway Analysis (QIAGEN) accounts using the Yale Medical Library license while 232 have a MetaCore (Thomson Reuters) account. The departments with the highest number of accounts are Genetics, Internal Medicine, Pathology, and Immunobiology. (Table 5)

Table 5. Departments with higher number of MetaCore and Ingenuity Pathway Analysis accounts.

Department	MetaCore	Ingenuity Pathway Analysis
Genetics	33	30
Internal Medicine	31	32
Pathology	23	27
Immunobiology	18	23
Public Health	11	9
Psychiatry	10	8
Pharmacology	8	1
Child Study Center	8	5
Neurosurgery	7	5
Neurology	6	1
Cell Biology	5	8
Comparative Medicine	5	8
Pediatrics	5	7

The specific combination of genetic variations in an individual defines not only the external appearance but also susceptibility to diseases, cancer, genetic disorders, and drug response. This explains the great interest in discovering and cataloging these variations and using them for disease association and functional studies. Last fall, Yale biomedical researchers were invited to attend presentations/demos in order to evaluate two resources for the analysis of genomic variation: Ingenuity Variant Analysis (QIAGEN), and Golden Helix (Golden Helix, Inc.). Although both sessions were well attended, there was not enough interest (except for a handful of users) in having access to these tools.

However, On January 6, 2016, a request for access to the professional version of the Human Gene Mutation Database (QIAGEN) was made by a group of unit directors and faculty of the Yale School of Medicine. This database contains comprehensive data on

published human inherited disease mutations. Nathan Rupp, Head of Collection Development, negotiated and secured a campus-wide license to this important resource.

Published articles using the Medical Library's IPA or MetaCore license

In order to further understand the usage and the value of the resources the Medical Library is licensing, librarians asked those biomedical researchers with an Ingenuity Pathway Analysis and MetaCore account to submit research papers in which they have used these tools. In addition, we searched the literature to identify papers from those researchers who did not respond to our request. As of June 3rd, 2016 we identified 10 research papers that used the IPA or MetaCore license provided by the Medical Library, illustrating the importance of these resources in the research process.

Lennington JB, Coppola G, Kataoka-Sasaki Y, Fernandez TV, Palejev D, Li Y, Huttner A, Pletikos M, Sestan N, Leckman JF, Vaccarino FM (2016) Transcriptome Analysis of the Human Striatum in Tourette Syndrome. Biol Psychiatry 79(5):372-82 PMID: 25199956

Mariani J, Coppola G, Zhang P, Abyzov A, Provini L, Tomasini L, Amenduni M, Szekely A, Palejev D, Wilson M, Gerstein M, Grigorenko EL, Chawarska K, Pelphrey KA, Howe JR, Vaccarino FM (2015) FOXG1-Dependent Dysregulation of GABA/Glutamate Neuron Differentiation in Autism Spectrum Disorders. Cell 162(2):375-90 PMID: 26186191

C Cappi, H Brentani, L Lima, S J Sanders, G Zai, B J Diniz, V N S Reis, A G Hounie, M Conceição do Rosário, D Mariani, G L Requena, R Puga, F L Souza-Duran, R G Shavitt, D L Pauls, E C Miguel, and T V Fernandez (2016) Whole-exome sequencing in obsessive-compulsive disorder identifies rare mutations in immunological and neurodevelopmental pathways. Transl. Psychiatry 6(3): e764. doi: 10.1038/tp.2016.30

Zhou Q, Wu SY, Amato K, DiAdamo A, Li P (2016) Spectrum of Cytogenomic Abnormalities Revealed by Array Comparative Genomic Hybridization on Products of Conception Culture Failure and Normal Karyotype Samples. J Genet Genomics. 43(3):121-31. doi: 10.1016/j.jgg.2016.02.002.

Srivastava A, Shinn AS, Lam TT, Lee PJ, Mannam P (2016) SILAC based protein profiling data of MKK3 knockout mouse embryonic fibroblasts. Data Brief. 7: 418-22. doi: 10.1016/j.dib.2016.02.034.

Guzeloglu Kayisli O, Kayisli UA, Basar M, Semerci N, Schatz F, Lockwood CJ (2015) Progestins Upregulate FKBP51 Expression in Human Endometrial Stromal Cells to Induce Functional Progesterone and Glucocorticoid Withdrawal: Implications for Contraceptive- Associated Abnormal Uterine Bleeding. PLoS One. 10(10):e0137855. doi: 10.1371/journal.pone.0137855.

Eric Lau, Yongmei Feng, Giuseppina Claps, Michiko N. Fukuda, Ally Perlina, Dylan Donn, Lucia Jilaveanu, Harriet Kluger, Hudson H. Freeze, and Ze'ev A. Ronai (2015) The transcription factor ATF2 promotes melanoma metastasis by suppressing protein fucosylation Sci. Signal. 8(406): ra124. doi: 10.1126/scisignal.aac6479

Haskins JW, Zhang S, Means RE1, Kelleher JK, Cline GW, Canfrán-Duque A, Suárez Y, Stern DF (2015) Neuregulin-activated ERBB4 induces the SREBP-2 cholesterol biosynthetic pathway and increases low-density lipoprotein uptake. Sci Signal. 3; 8(401):ra111. doi: 10.1126/scisignal.aac5124.

Elisa Araldi, Marta Fernández-Fuertes, Alberto Canfrán-Duque, Wenwen Tang, Julio Madrigal-Matute, Abdul Basit, Aránzazu Chamorro-Jorganes, Gary Cline, Miguel Angel Lasunción, Dianqing Wu, Carlos

Fernández-Hernando and Yajaira Suárez (2016) "Lanosterol modulates innate immune responses in macrophages" (Sent to Cell Metabolism)

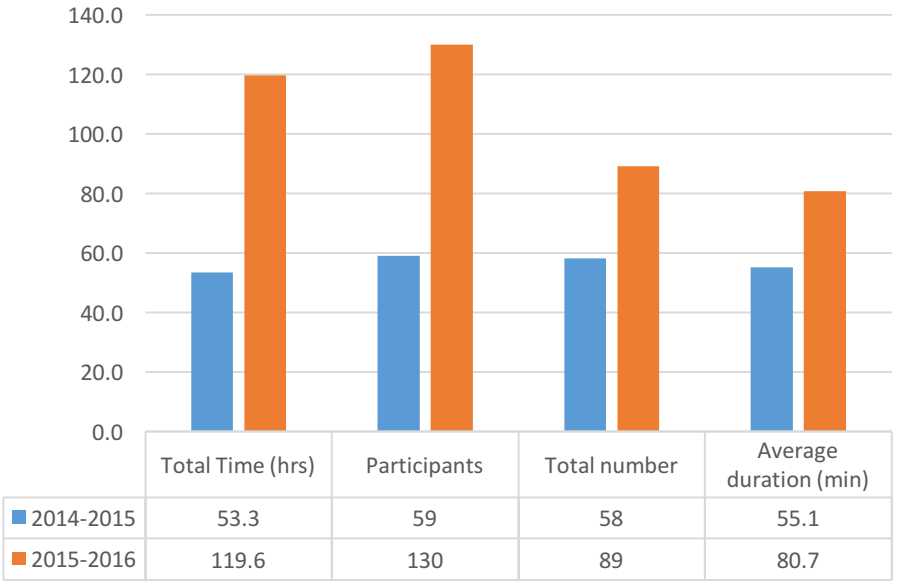
Guillermo C. Rivera-Gonzalez, Brett A. Shook, Katherine Bollag, Brandon Holtrup, Matthew S. Rodeheffer and Valerie Horsley. Adypocyte stem cell self-renewal is regulated by a Pdgfa/Akt signaling axis in the skin (Submitted)

Consultations

As mentioned above, one of the objectives of the training sessions was to showcase resources and expertise available at the medical library in support of end-user bioinformatics. During FY2015-16, both the consultation hours and number of participants increased 2.2 times while both the average time and total number of consultations increased 1.5 times compared to previous FY2014-15 (Figure 5). Besides this increase, it is important to notice that more than 42 % of these consultation used the tools licensed by the Medical Library (e.g. IPA and MetaCore). Finally, these sessions are beginning to translate into long term collaboration projects that result either in presentations or papers (see below). A summary of the consultations offered during this year can be seen in Table 6.

"Your help made my studies more effective and interesting as we kept digging in the genomic data. Needless to say that you will be a coauthor on this paper."

Associate Research Scientist in Genetics, Yale School of Medicine



	Total Time (hrs)	Participants	Total number	Average duration (min)
2014-2015	53.3	59	58	55.1
2015-2016	119.6	130	89	80.7

Figure 5. Face-to-face consultation sessions: comparison between FY2014-15 and FY2015-16

Resulting paper/presentation in collaboration with biomedical researchers FY2015-16:

Zuo, Lingjun; Garcia-Milian, R, Li, Chiang-Shan Ray; Luo, Xingguang Replicable risk nicotinic cholinergic receptor genes for nicotine dependence (Submitted to Genes 2016)

Canaan A, Arjona C, Seay M, Garcia-Milian R. How does FAT10 silencing extend lifespan in mice? Nathan Shock Center Summit: Personalized Geroscience Bell Harbor Convention Center Seattle, Washington June 3, 2016

Table 6. Summary of face-to-face consultations during FY2015-16

Duration (Minutes)	Participants	Position	Department	Topic
90	1	Assistant Professor	Epidemiology (Chronic Diseases)	Functional analysis of gene lists resulting from methylation studies
60	1	Research Scientist	Genetics	Search databases and knowledge bases for gene or protein aberrations associated to human neurodegenerative disorders
30	1	Research Scientist	Dermatology	Functional analysis of RNAseq data with IPA and MetaCore
75	1	Postdoc	Neurobiology	Analyzing a data set from GEO followed by functional analysis of differentially expressed genes
30	1	Postdoc	Pharmacology	How to cross-referencing datasets using IPA
120	2	Associate Professor	Medicine (Digestive Diseases)	Functional analysis of proteomics data - MetaCore and IPA
120	1	Associate Research Scientist	Cell Biology	Training on how to use IPA
120	1	Graduate Student	Cell Biology	Consultation on how to use different databases: TCGA COSMIC, MetaCore, IPA and GEO datasets.
95	1	Graduate Student	Pathology	Protein-protein Interaction networks, modeling and prediction on IPA and MetaCore
150	1	Associate Research Scientist	Cell Biology	Functional analysis of RNAseq data
60	1	Associate Professor	School of Nursing	Searching databases to find relationship between gene aberrations and phenotypes
150	1	Research Scientist	Immunobiology	Functional analysis of RNAseq data MetaCore
70	1	Graduate Student	Cell Biology	Functional analysis of RNAseq data IPA.
20	1	Postdoc	OBGYN	Mining a dataset for SNPs and phenotype association
1	60	Visiting Fellow	Genetics	Ffunctional analysis of RNAseq data IPA
120	1	Associate Research Scientist	Genetics	Functional analysis of RNA seq data using IPA and Metacore
90	1	Graduate Student	Pathology	Interaction between two list of genes MetaCore, BioMart
120	1	Research Scientist	Yale Cardiovascular Research Center	Functional analysis of RNAseq data MetaCore
30	1	Graduate Student	Mol Biophysics and Biochemistry	How to convert IDs into official gene symbols
60	1	Visiting Research Scientist	OBGYN	Functional analysis of RNAseq data- MetaCore.
45	1	Associate Research Scientist	Psychiatry	Finding SNPs associations with phenotype
90	1	Visiting Research Scientist	OBGYN	Functional analysis of RNAseq data MetaCore
35	1	Assistant Professor	Genetics	NIH PA compliance- MyBibliography
180	1	Visiting Research Scientist	OBGYN	Functional analysis of RNAseq data- MetaCore
20	1	Visiting Research Scientist	OBGYN	Functional analysis of RNAseq data
120	1	Research Scientist	Immunobiology	How to use Metacore for enrichment analysis.
55	1	Research Scientist	Pathology	Cleaning RNA seq data for subsequent functional analysis on MetaCore/IPA
90	1	Associate Research Scientist	Epidemiology	How to navigate, find information and use genome browsers: UCSC genome browser and Ensembl
120	1	Visiting Research Scientist	Epidemiology	Understanding the biological significance of gene variations associated with a specific phenotype- different tools used including MetaCore

15	1	Professor	Genetics	Moving different collections into MyBibliography- MyNCBi
65	1	Associate Professor	Yale Cancer Center	Searching for specific gene expression on different databases
25	1	Research Scientist	Epidemiology	Finding all the orthologues for a phylogenetic study
20	1	Administrative Assistant	Genetics	NIH PA compliance- MyBibliography, delegating accounts
25	1	Research Scientist	Pathology	Literature-mediated search for a review paper
20	1	Administrative Assistant	Genetics	NIH PA compliance- MyBibliography, delegating accounts
50	1	Postdoc	Therapeutic Radiology	How to retrieve variation data from the TCGA
65	3	Postdoc	Biostatistics	Demo on how to use IPA and MetaCore.
55	1	Research Assistant	Therapeutic Radiology	Finding protein interaction networks- IPA
30	1	Associate Research Scientist	Pathology	Functional analysis of proteomics data using MetaCore
70	1	Administrative Assistant	Pathology	Reference management with EndNote
1	1	Assistant Professor	Epidemiology	Question on whole genome methylation analysis
75	1	Research Assistant	Therapeutic Radiology	How to find variations of a gene related to squamous cell carcinoma of head and neck
25	2	Graduate Student	BBS	Finding information on transcription factor-diseases, interactions, functions
65	1	Research Associate	Immunobiology	IACUC search training
45	1	Research Scientist	Neurobiology	Training on how to do an IACUC search
35	1	Research Scientist	Genetics	Searching ClinicalTrials.gov
25	1	Administrative Assistant	Pharmacology	NIH Public access compliance MyBibliography
1	1	Postdoc	Pathology	Database search for protein and RNA expression of a gene in cell lines
55	1	Graduate Student	Cell Biology	How to compare two or more bacterial genomes based on homology, or phenotypes
50	1	Graduate Student	Pathology	Finding cell lines that express a specific protein
180	1	Research Scientist	Pathology	Functional analysis with MetaCore
45	1	Research Scientist	Pathology	Finding the sequence for the a gene to create PCR primers
60	1	Postdoc	Pathology	Reference management with EndNote
45	1	Postdoc	Internal Medicine	Functional analysis of data on MetaCore
60	1	Associate Professor	Pathology	literature-mediated search

Medical Library Role in Support of Precision Medicine

Precision Medicine (also called as personalized medicine) is the use of high throughput genomic, transcriptomic, and proteomic to characterize patients in order to guide diagnostic, prognosis, treatment, and prevention of diseases. It is seen as the combination as the combination of biomarker, molecular information, and clinical phenotype at the individual patient level (Vicini et al., 2016).
Yale University researchers and clinicians are already working on precision medicine-related projects. The Yale Cancer Center and Smilow Cancer Hospital are performing molecular profiling of cancer patients (Lynch, 2015), and Yale University has launched a multicenter clinical trial, sponsored by Stand Up to Cancer and Melanoma Research Alliance. The latter, will apply the advances in personalized medicine technology to treat metastatic melanoma (Kashef, 2015).
During the FY2015-16 the Clinical Support Librarian Denise Hersey and the Biomedical Sciences Research Support Librarian have begun to take steps in order to understand the role of the medical library in the new age of precision medicine. This will allow for better supporting the information and data needs of researchers, students, and clinicians, and the delivery of relevant resources and services.

Early this year, we submitted an application to the SPARK (Institute of Museums and Libraries) to assess the information and data needs of those working on precision medicine-related projects. In addition, clinicians and researchers working on precision medicine at Yale have been identified on further discussion, as well as some information and data resources/tools available for precision medicine. For example, in June 2016, Denise and Rolando coordinated and attended (along with Nathan Rupp, Head of Collection Development) a presentation on Thomson Reuters's new Precision Medicine Intelligence database: "Supporting Evidence-Based Genomic Interpretation with the Thomson Reuters Precision Medicine Intelligence".

Information Needs of Biomedical Researchers

Between January and March 2016 an online Quatrics survey "Information and Needs Assessment for Biomedical Research in the Omics Era" (HSC# 1511016778) was emailed to 860 Yale-affiliated individuals who had registered for at least one training session on bioinformatics-related topics offered by the Cushing/Whitney Medical Library. This survey tool was design as a collaboration project between Dr. Milica Vukmirovic (Pulmonary, Critical Care & Sleep Medicine, at the Yale School of Medicine), Denise Hersey, and Rolando Garcia-Milian (Curriculum & Research Support Dept.). One hundred and seventy-six individuals responded to this survey, for a 20.4% response rate. Overall, we acknowledge that the present survey may not represent the views of the entire Yale biomedical research community but it does highlight relevant questions and provides initial insights into their information and data needs issues.

Three well-defined groups of respondents can be identified: faculty, postdocs, and graduate students (50, 50 and 38 respondents respectively). None of the respondents identified themselves as Resident or Clinical Fellow.

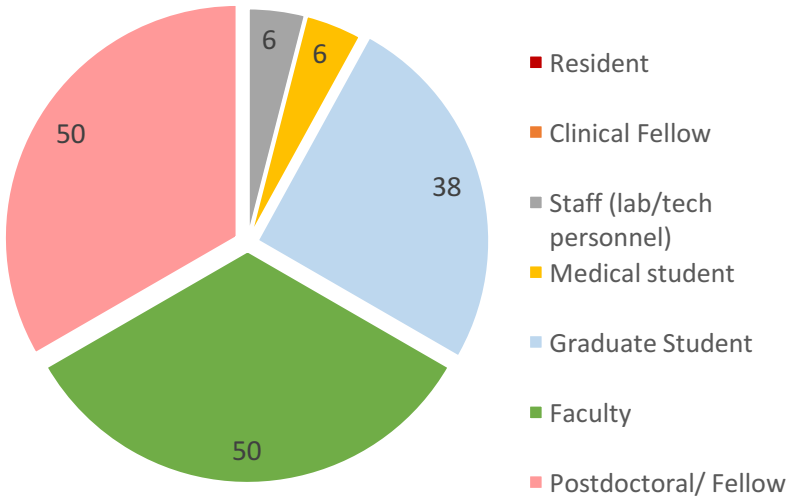


Figure S1. Which of the following best describes your role? Please select all that apply. Total responses: 146

The most represented respondents were affiliated with the Genetics and Immunology departments (10 individuals each) followed by the Department of Molecular Biophysics and Biochemistry (MB&B) and Pathology (9 and 8 individuals respectively). The rest of departments represented can be seen in Figure S2.

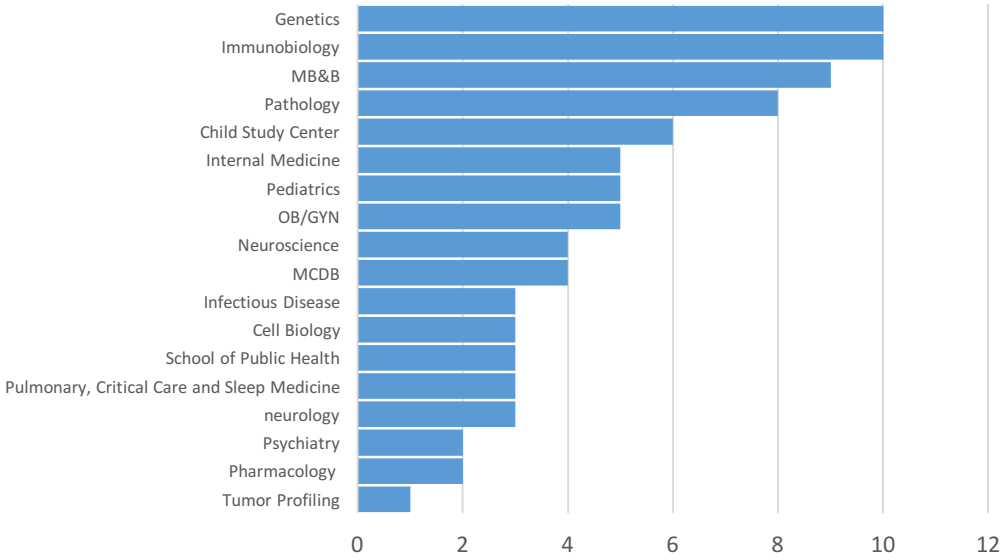


Figure S2. Please provide the name of your department. Total responses: 134

The majority of respondents (44 individuals) located in the Anlyan Center (TAC) and Sterling Hall of Medicine (SHM) at 333 Cedar St. (26), followed by the West Campus (12).

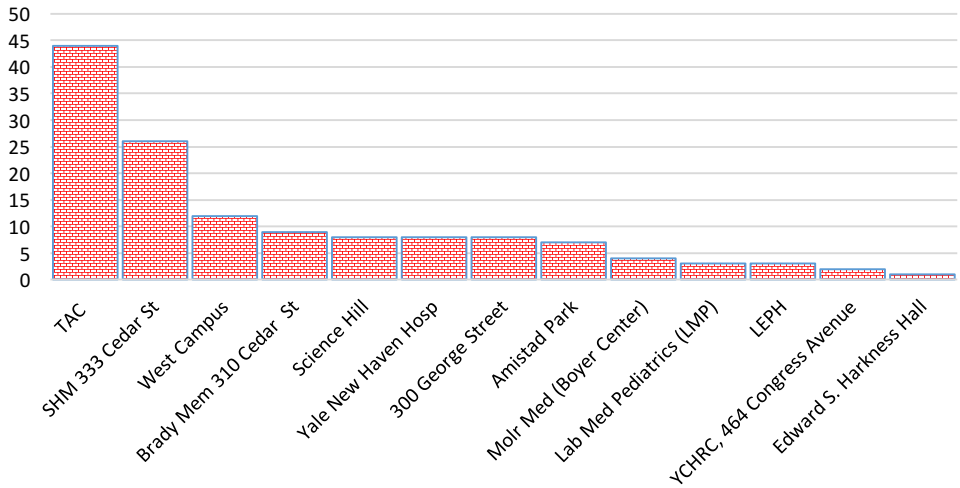


Figure S3. Please select your main location on campus. Total responses: 144

Human and mouse are the primary organisms of studies (104 and 73 individuals respectively- see Figure S4). The "Other" category includes planarians, frogs, *Caenorhabditis elegans*, non-human primates, and Arabidopsis.

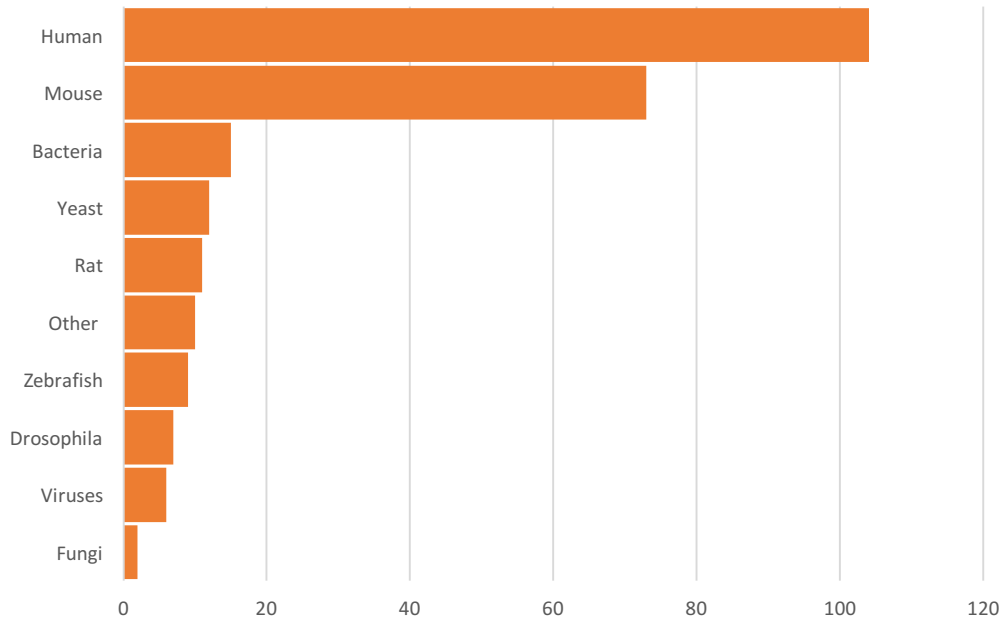


Figure S4. What is (are) the primary organism(s) that you study? Total responses: 144

High-throughput data analysis (RNAseq, DNA-seq, microarray, etc) was selected by respondents as their most important type of analysis. We are coordinating presentations on suitable tools for this stage of the high-throughput data life cycle such as: CLC Main Workbench (CLC Bio QIAGEN), and Partek Flow (Partek Incorporated, http://www.partek.com/partekflow). In addition, conversations are already in progress with the leadership of Yale Research Computing (High Performance Computing) to evaluate the implementation of the Galaxy open source software for the analysis of Next Generation Sequence data. (Please see the **Resources and Tools** section in this report for more details on these tools)

Table S1. Please indicate how important are the following types of data analysis for your research. Total responses: 134

Question	Not Important	Important	Very important	Total Responses
Analysis of high throughput data (e.g. microarray data, RNA/DNA-Seq)	16	21	97	134
Signaling, network, and pathway analysis	13	33	84	130
Functional analysis of high throughput data	20	36	74	130
Transcription factor and gene regulatory sequence analysis	25	38	68	131
Integrated searches of literature and high throughput data	15	50	64	129
DNA/protein sequence manipulation and analysis	17	50	61	128
SNP, genetic variation, Genome wide association data analysis	42	42	48	132
Other data analysis needs (please list below)	11	4	17	32

Other data analysis needs

- ✓ Statistical analysis, including the use of programs like SPSS but also a brief background on various statistical tests and regressions to inform the appropriate selection for data sets. Would be great to have a workshop specifically tailored to working with high throughput DNA data from Annovar.
- ✓ Custom software analysis of fMRI – Matlab software
- ✓ Analysis of flow cytometry data is very critical to our daily work.
- ✓ ANOVA MANOVA
- ✓ Sequence alignment visualization
- ✓ Methylation/epigenetic analysis
- ✓ Transcriptome assembly
- ✓ Clustering (hierarchical, PCA etc)
- ✓ Image analysis
- ✓ Comparative genomics
- ✓ Proteomics and metabolomics data
- ✓ Statistical analysis of mutations identified in whole exome sequencing; basic statistical analysis in genetics (ie what is a benjamini hochberg correction or a Bayseian approach vs. classical statistics)
- ✓ Mass spectrometry, small angle x-ray scattering
- ✓ Published ChIP-seq data
- ✓ Mass Spec data analysis
- ✓ Integrative Genomics Viewer

- ✓ Meta analyses
- ✓ Visualization of alignments and CRISPR alignment visualization.
- ✓ Somatic variant analysis and annotation
- ✓ Two photon calcium imaging data
- ✓ Epigenome analysis
- ✓ Epigenetics, genomics, omics
- ✓ Anything related to cancer biology"
- ✓ Metabolomics
- ✓ Clustering, adjusting data for multinomial regression
- ✓ These are all great topics to teach but first, a core course in shell programming and scripting is necessary
- ✓ FACS data
- ✓ Proteomics and metabolomics
- ✓ Proteomics
- ✓ DNA Methyaltion/epigenomics
- ✓ Phosphosite-specific tools (role of particular phosphosite in regulating protein)
- ✓ MS

Overall, the majority of the respondents (78%) identified lack of adequate training as the main challenge they face in their research. This is followed by not having the proper database or software (54%). However, not having the proper training is perceived as a bigger challenge for graduate students and postdocs (94% and 85% of respondents) than for faculty (62%). In contrast, the main challenge for faculty is not having the adequate software, database or tool (73% of respondents) followed by postdocs (63%), but this is not perceived as a main challenge by graduate students (29%) (see Figure S6)

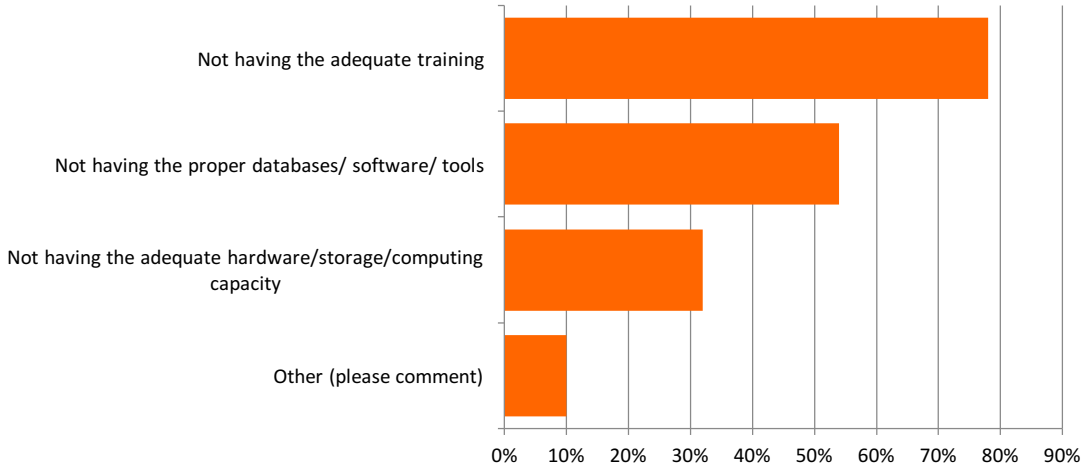


Figure S5. What are the main challenges with data analysis that you are facing in your research? Please check all that apply and comment if you would like. Total responses: 130

Other comments

- ✓ I would greatly benefit from training beyond the basic level for Unix command line coding, especially as it relates to processing sequencing data from YCGA, as well as information on various statistical analyses to inform choices about which tests to use.
- ✓ I am still learning these bioinformatics techniques - it takes me a while to teach myself.
- ✓ We have the appropriate software I need, like IPA, but because Yale only has 2 "seats" on these program licenses, I can rarely get on the software when I need to. I'm having to a lot of my data analysis at "off hours," and it's making my lab progress really inefficient and frustrating.
- ✓ I'm taking a course to give me adequate training and learn about resources for help through the genetics department this semester, so it should not be an issue in the future.
- ✓ Adequate ITS support. Issues of their system losing all backups (only found this out when desktop failed). Inability to quickly fix computer problems.
- ✓ Resource on how to use R
- ✓ Lack of software like Lasergene DNAStar
- ✓ Not having a single software that can manage all formats of bioinformatics files.
- ✓ Combination of lack of training and subsequently not knowing which software is the best one to use for a given problem.
- ✓ It would be great to have training sessions on (1) intro to public bioinformatics databases for gene, protein, and transcriptome data (like the Cold Spring Harbor Labs 2 day course) and (2) how to make heat map figures/R for biologisits--introductory level

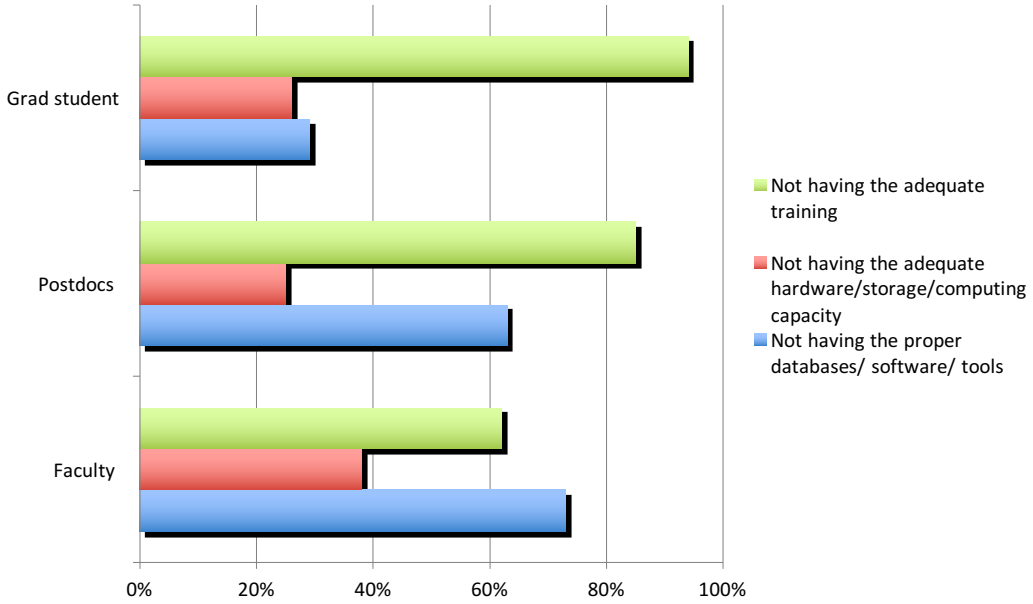


Figure S6. Main challenges. Comparison between faculty, postdocs, and graduate students.

Figure S7. Does finding, retrieving, analyzing high-throughput omics-related, or other type of data (e.g. image, text, etc.) represent a bottleneck for your research project? Total responses: 129

Comments

- ✓ I end up relying on Lynda and Coursera to teach myself necessary skills over several hours what might be able to be learned from a one hour workshop.
- ✓ Lack of training on how to best analyze high throughput data dramatically slows my research
- ✓ I am just moving, and getting all the data transferred, learning what tools are available at Yale etc. is taking time. I have not gotten to the point yet that I am

able to comment whether eventually all tools and infrastructure will turn out to be present here or not.

- ✓ We collaborate with a number of investigators on proteomics data analysis. The main bottle neck tends to be the follow-up data analysis to relate the identified protein expression to biological/functional relevant. Making available standardized software and resources for investigators to go to for these analysis will not only benefit the investigators that we collaborate with but will also greatly assist us (the core facility) and ensuring that there is a cost effective way for users to obtain these analyses.
- ✓ I have to analyze a lot of high-throughput data, and not being able to log on to IPA before 5pm really makes my progress suffer.
- ✓ As a bioinformatician, this is what I do. For me, it is not a bottleneck as much as a challenge to create pipelines to retrieve and analyze omics data.
- ✓ Not presently, but it may well be a bottleneck in the near future
- ✓ "It's difficult to determine the appropriate normalization methods to apply under different circumstances.
- ✓ It's difficult to determine read accuracy of sequencers and difficult to determine how to avoid having batch effect trends."
- ✓ In order to do novel research, I need to analyze omics data to find gene/protein of interest to design experiments as well as interpret the results with similar bioinformatics tool
- ✓ Only one person does all the analysis
- ✓ RNAseq data could be so much quicker and projects could move faster if I were better able to analyze it
- ✓ It consumes a lot of time using (and setting up) the many scripts in Unix and web-based tools for RNAseq or proteomics data to analyze.
- ✓ Using NGS methods on single cell level of rare cell populations
- ✓ High-throughput calcium imaging is widely used in our team.
- ✓ Part of learning experience for students, but problem for their PIs in general
- ✓ Yale should strive to make low cost tools amply available so we can be at the cutting edge of bioinformatics use
- ✓ Mostly the analyzing part.
- ✓ The analysis is always a bottleneck, we generate a lot of data.
- ✓ We move at our own pace and have so far been able to get help when needed form other labs that do similar work or learn on our own as we go.
- ✓ Mostly on the analysis side.
- ✓ No, but maybe I could learn how to do it more efficiently if formal training was provided
- ✓ Large amount of data and no computer clusters available
- ✓ I do use public databases and tools for analysis of my 'omics data, but I am sure my data analysis would be greatly enhanced by the types of formal training sessions suggested above.
- ✓ Not any more of a bottleneck than for others doing similar research, but this is an increasingly large component of our research overall
- ✓ We get results of our omics data but not interpretation. It would be meaningful to be able to understand the data.

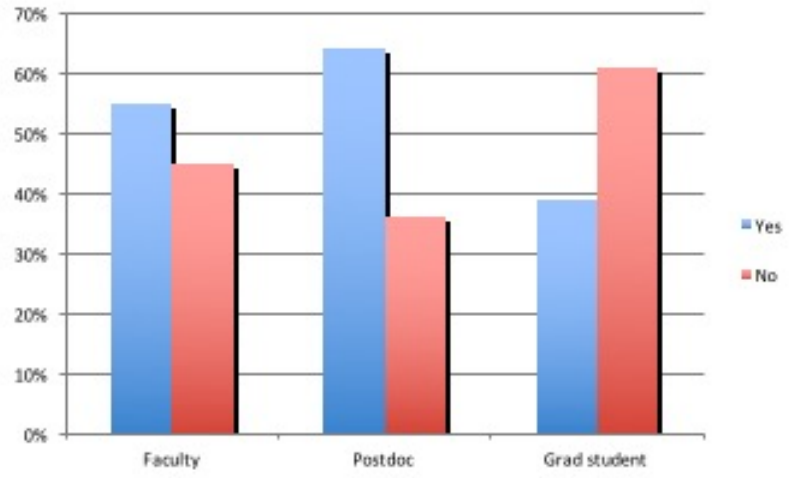


Figure S8. Finding, retrieving and analyzing data represents a bottleneck for your project. Comparison between faculty, postdocs, and graduate students

Figure S9. Do you have data that you have not been able to analyze because you lack access to the appropriate software? Please comment below. Total responses: 128

Comments

- ✓ SPSS should be a free software with training available. unfortunately that is not the case in Yale
- ✓ Not at this point in time.
- ✓ Variant Analysis
- ✓ We don't have access to sufficient software to assist in processing protein posttranslational modification (PTM) mass spectral data, additionally to carry out quantitative MS data analyses for these PTM workflow. There exist these software but it would be tremendously helpful to have these licensed software available for the investigators that utilized our resource to carry out these analyses to reduce the cost for their research.
- ✓ I mainly write my own code and build off of other software.

- ✓ I would like to be able to use a tool called PathSeq from the Broad Institute to analyze my data, but have been unable to do so because it only operates on an Amazon Cloud platform rather than the cluster…
- ✓ Could use MLWin Software
- ✓ Not exactly, bioinformaticians in our department know how to work with what we have.
- ✓ From what I understand there has yet to be any universally effective software created for non-bioinformaticians with a simple user interface that biology-background researchers can use to accurately normalize data or map sequenced reads.
- ✓ This is more of an inherent problem with the sophistication of bioinformatics and the limited understanding most biologists have in analysis. Hiring more bioinformatics specialists might be a better solution than providing alternate software suites. Free workshops on seq analysis for amateurs would be wonderful as well.
- ✓ Publicly available ChIP-seq data
- ✓ Genome seq
- ✓ I already found a good program but I have to buy it myself.
- ✓ Not sure
- ✓ But I am not sure we are using the best tools/software
- ✓ I am trying to find out what the potential genes regulated by a specific transcription factor are, based on the expression of the DNA binding site motif across the genome
- ✓ Main problem is my lack of training
- ✓ Metabolomics analyses software for profiling, quantitation, and identification.
- ✓ Due to the lack of software I have to use multiple free on-line tools with different format converters to analyze my data
- ✓ Not yet, but am preparing a proposal

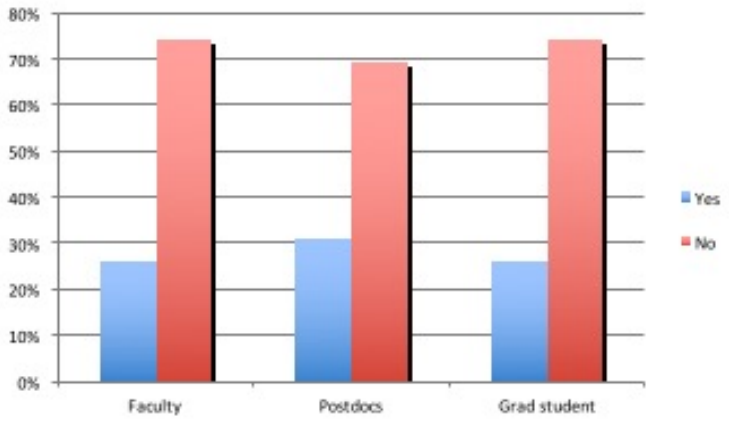


Figure S10. Do you have data that you have not been able to analyze because you lack access to the appropriate software? Comparison between faculty, postdocs, and graduate students

Figure 11. Are there commercial bioinformatics databases, software programs/tools that you would like the medical library to consider licensing? Total responses: 55

Comments

- ✓ Ingenuity VARIANT Analysis
- ✓ IPA software, RNA seq analysis
- ✓ Not at this time
- ✓ Matlab
- ✓ Golden Helix, SNP & Variation Suite
- ✓ Qiagen Ingenuity Variant Analysis tool
- ✓ Treestar Flowjo
- ✓ No, but more training in basic tools like R, python would be great
- ✓ SPSS, Prostat
- ✓ Lasergene DNAStar
- ✓ IVA, QIAGEN
- ✓ IPA
- ✓ I wish I knew enough to answer this.
- ✓ Ingenuity Variant Analysis
- ✓ MASCOT Search Engine (http://www.matrixscience.com/)
- ✓ Byonics software (http://www.proteinmetrics.com/products/byonic/)
- ✓ Proteome Discoverer (http://portal.thermo-brims.com/)
- ✓ Compound Discoverer (https://www.thermoscientific.com/en/product/compound-discoverer-software.html)"
- ✓ Almost all the analysis tools I use are open source.
- ✓ PathSeq or someway to use Amazon Cloud with the cluster without having to pay for an individual account
- ✓ Some of the software that the YCGA uses may be more broadly used, if available. I also need access to sequence analysis software such as Sequencher and SnapGene, but I am not sure if these are more appropriately provided by the Library or ITS-Med
- ✓ IVA

- ✓ Adobe Suite
- ✓ SciFinder and the link to the full text including patents.
- ✓ HGMD professional level (up to date)
- ✓ Partek Genomic Suite
- ✓ MATLAB
- ✓ XCalibur
- ✓ Integrative Genomics Viewer
- ✓ "Adobe Photoshop
- ✓ FlowJo"
- ✓ I was very appreciative of the licensing of Ingenuity, which I was able to use for my data.
- ✓ Cytobank
- ✓ Partek
- ✓ Lasergene DNAStart
- ✓ Genome Annotation
- ✓ Variant analysis - Qiagen
- ✓ Geneious R9 from Biomatters Limited
- ✓ Gene spring from Agilent
- ✓ WIll let our dept. experts answer this- I have forwarded your email about survey to them
- ✓ No, many of the analysis softwares that are well accepted in the field are open source and can be accessed and installed by anyone with appropriate hardware
- ✓ Spotfire
- ✓ "Consider licensing a platform (or finding a freeware one) that allows emulating Linux on windows given the fact that many users still use windows and this would ease the barrier of them having to purchase or only work on a Linux workstation.
- ✓ Also, some of the programs on the high performance computers need to be updated.
- ✓ Partek, Nanostring and StrandNGS
- ✓ ANYTHING FOR Rna seq -Chip seq etc.
- ✓ Vector NTI
- ✓ Progenesis QI (for metabolomics and proteomics), Mass Frontier (Thermo Fisher scientific). Having MASCOT server/search engine for proteomics licensed for Yale users.
- ✓ Yes. Geneious from Biomatters
- ✓ Most updated annotated version of the iRegulon plug-in for Cytoscape. A training session for Cytoscape would be excellent too.
- ✓ Ingenuity Variant Analysis
- ✓ IPA, Partek
- ✓ Golden Helix SVS. We absolutely depend on this software!

Figure S12 Do you have data that you have not been able to analyze because you lack sufficient training to use the analysis software/tool that is required? Please comment. Total responses: 107

Comments

- ✓ Insufficient training to analyze mass spec data.
- ✓ I really need greater training in Unix command line codes as well as statistical analysis.
- ✓ high throughput sequencing data
- ✓ As a young graduate student, I have not been trained on large data set analysis and am not sure how to acquire such training
- ✓ Currently struggling to complete a meta-analysis of microarray data
- ✓ SPSS
- ✓ Not at this point in time, but I am currently a rotation student and so the projects typically do not yet require this level of data analysis.
- ✓ RNA SEQ
- ✓ I wish I have more training in code writing so I can use Matlab efficiently to analyze my data
- ✓ Lack of programming skills
- ✓ We are struggling to handle some of our data sets, but some of these sets may be unprecedented in scale and there is no tool that will simply solve our dilemmas.
- ✓ RNAseq and ChIP-seq
- ✓ Cytof data
- ✓ I have differential expression data from RNA seq and am not proficient in producing the pathway maps
- ✓ I have the training but it is never enough with the speed of technology.
- ✓ I did analyze my data but not sure if I did it correctly
- ✓ IPA KEGG
- ✓ I believe some of my students have had to start from scratch learning things because bioinformatics personnel are tied up on projects for PI, with little time to teach students
- ✓ Time has not yet permitted ample training but training is a priority that ought be supported
- ✓ Some students have complained about this
- ✓ R, adjusting data for multinomial regression, their clustering, than correlation

- ✓ Metabolomics software.
- ✓ Yes, proteomics analysis software
- ✓ Limited capacity of bioinformatician

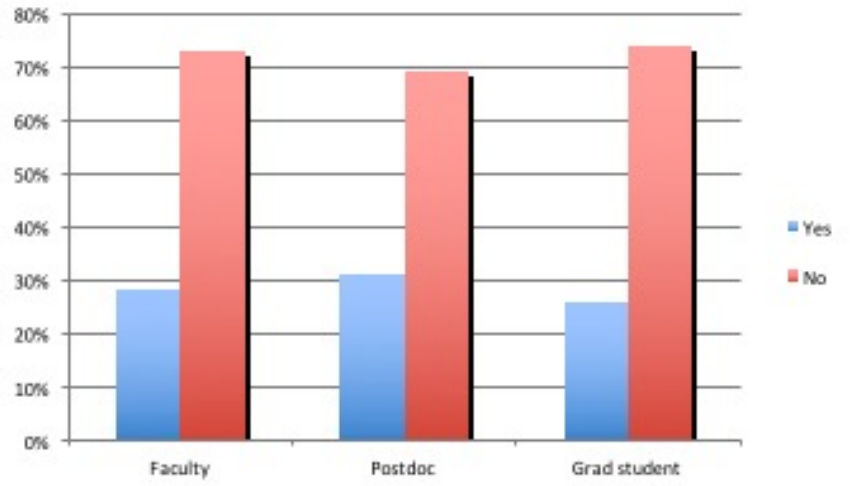


Figure S13. Do you have data that you have not been able to analyze because you lack sufficient training? Comparison between faculty, postdocs, and graduate students

Are there training, classes or how-to sessions that you would like to see offered?

- ✓ RNA seq data analysis Microarray data analysis
- ✓ Yes. I would like to see training sessions on how to process sequencing data using command line codes in Putty as well as background information about the statistical tests covered in SPSS.
- ✓ UCSC genome browser
- ✓ I can stumble through with R and python using google and books but anything to get better would be nice.
- ✓ Mass spec data analysis, sequencing analysis
- ✓ Using R Bioconductor packages, python, Unix
- ✓ Yes SPSS
- ✓ Data analysis related to Nest Generation Sequencing.
- ✓ Again, I don't know enough to answer this. You should also query my students.
- ✓ Training for the R programming language, specifically visualizations.
- ✓ Coding in Matlab, R and Python
- ✓ R programming, Phyton, Plink
- ✓ I would love a class on how to filter and map reads to the genome (such as Cufflinks, TopHat, etc)
- ✓ Training classes for the suggested software tools in the previous survey page. These training classes will be very beneficial to Yale investigators that utilize

our MS and Proteomics Resource to obtain proteomics and metabolomics data for their research.

- ✓ I would like to see more in-depth Illustrator, metacore, IPA, and DAVID training.
- ✓ Programming sessions and R sessions
- ✓ I would like to see training on using the R statistical language for biomedical data analysis.
- ✓ Statistics boot camp for genomics data (ie why/what is BH corrections or why use Bayesian stats vs classical)
- ✓ RNA-seq data analysis
- ✓ Statistics for biology
- ✓ Absolutely, especially a very general one for amateurs. I wouldn't want it to be overly-specific and catered to one form of seq from one specific type of sequencer analyzed on one specific software. Our lab uses a large swathe of sequencing and analytic methods, and the standards in the industry shift every few years.
- ✓ R course
- ✓ R with bio-statistic analysis, R with generating visualized data
- ✓ TCGA data interpretation
- ✓ Integrative Genomics Viewer
- ✓ R, Perl, Metagene graphing
- ✓ Variant analysis - Qiagen - for SOMATIC variants from paired tumor/normal samples
- ✓ YES! Proteomics data analysis.
- ✓ I think courses on understanding and analyzing types of big data would be helpful.
- ✓ How to use Galaxy, Basic training in R
- ✓ Data visualization
- ✓ How to use IPA
- ✓ SNP and RNA Seq
- ✓ Please contact Grad School!! Medical School area needs should be coordinated through the BBS program- Biological and Biomedical Sciences PhD program
- ✓ More on how to interpret omics data output
- ✓ Training on R, Linux and statistics for biologists to help them analyze RNA-Seq data
- ✓ R training, script reviews
- ✓ Introduction to R, Introduction to high performance computer usage
- ✓ Would be nice to have more basic classes for Linux and R as most analysis gets done through the servers and requires sufficient Linux knowledge and R for making graphs and analysis of some of the data after analyzed.
- ✓ Yes, metabolomics analyses.
- ✓ Programming course in python, R would make all data analysis much easier
- ✓ Yes. Proteomics analysis software
- ✓ Yes, please see above
- ✓ Programming training for non-students
- ✓ Classes in R

Table S2. Comparison between faculty, graduate students, and postdocs in terms of training needs.

Faculty	Postdocs	Graduate Students
Partially	RNA seq data analysis Microarray data analysis	Mass spec data analysis, sequencing analysis
data analysis related to Nest Generation Sequencing	Yes. I would like to see training sessions on how to process sequencing data using command line codes in Putty as well as background information about the statistical tests covered in SPSS.	I would love a class on how to filter and map reads to the genome (such as Cufflinks, TopHat, etc)
Again, I don't know enough to answer this. You should also query my students.	Ucsc genome browser	I would like to see more in-depth Illustrator, metacore, IPA, and DAVID training.
Training classes for the suggested software tools in the previous survey page. These training classes will be very beneficial to Yale investigators that utilize our MS and Proteomics Resource to obtain proteomics and metabolomics data for their research.	can stumble through with R and python using google and books but anything to get better would be nice	I would like to see training on using the R statistical language for biomedical data analysis.
RNA-seq data analysis	Using R Bioconductor packages, python, Unix	statistics boot camp for genomics data (ie why/what is BH corrections or why use Bayesian stats vs classical)
Variant analysis - Qiagen - for SOMATIC variants from paired tumor/normal samples	SPSS	Absolutely, especially a very general one for amateurs. I wouldn't want it to be overly-specific and catered to one form of seq from one specific type of sequencer analyzed on one specific software. Our lab uses a large swathe of sequencing and analytic methods, and the standards in the industry shift every few years.
I think courses on understanding and analyzing types of big data would be helpful.	Coding in Matlab, R and Python	Data visualization
Snp and RNA Seq	R programming, Phyton, Plink	
more on how to interpret omics data output	Programming sessions and R sessions	
Yes - you are doing very good job, I need to find time for it	Statistics for biology	
Introduction to R, Introduction to high performance computer usage	Classes preferred	
Yes, metabolomics analyses.	R course	
Yes, please see above	R with bio-statistical analysis R with generating visualized data	
Programming training for non-students	TCGA data interpretation	
	Integrative Genomics Viewer	
	R, perl, metagene graphing	
	YES! Proteomics data analysis.	
	How to use Galaxy, Basic training in R	
	How to use IPA	
	Training on R, Linux and statistics for biologists to help them analyze RNA-Seq data	
	R training, script reviews	
	Programming course in python, R would make all data analysis much easier	
	Yes. Proteomics analysis software	
	Classes in R	

Figure S14. Would working groups and discussion panels on the challenges and solutions to data collection and analysis be helpful? Please comment. Total responses: 107

Figure S15. Would you or someone you know, here at Yale, be willing to teach or provide training on tools for the analysis of data? If interested, please contact us. Total responses: 104

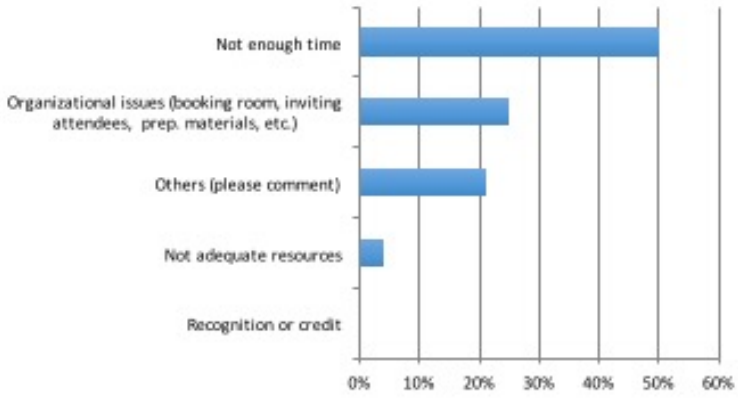


Figure S16. If you are willing to provide training, are there any barriers or challenges that impede you from doing this? Please select the main challenge. Total responses: 24

The high amount, diversity, and complexity of the data generated by high-throughput Omics-related technologies pose challenges in terms of finding, retrieving, analyzing, and sharing data. Therefore we would like to support collaborative work and networking to overcome these challenges.

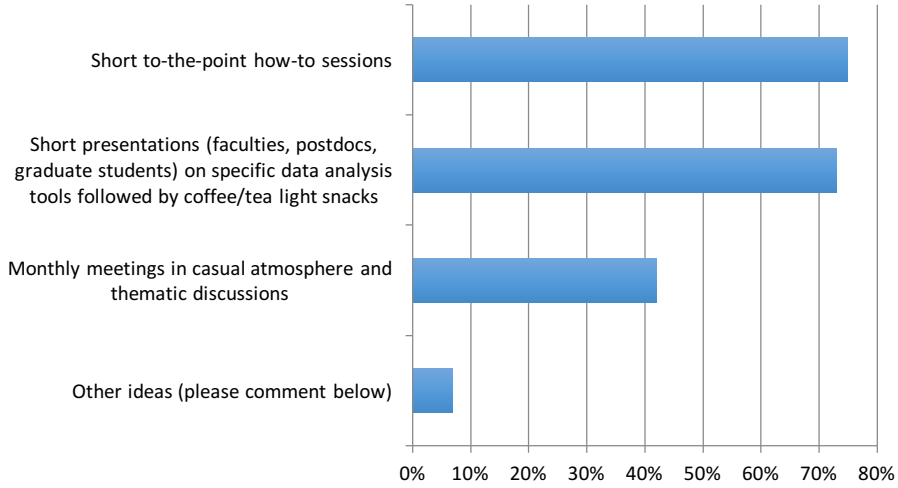


Figure S17. What format of networking and/or social event would foster the most opportunities for interaction and collaboration among Yale colleagues regarding finding, retrieving, analyzing data/information? Total responses: 107

Comments

- ✓ Online.
- ✓ Anything would be fine
- ✓ For example statistics, a series of weekly classes maybe 1 hour at a time.
- ✓ These all seem good. However there is a massive disparity in knowledge between bioinformatics specialists and biologists.
- ✓ I would recommend catering different programs and presentations for different types of researchers.
- ✓ I've sat through several of these types of presentations and they either become too boring/simple for statisticians, or too esoteric for even post-doc biologists to understand.
- ✓ 1) A YouTube channel with short to the point how to videos
  2) A web forum with Q&A about specific topics
  3) Short trainings from software vendors
- ✓ All of the above are useful for different audiences- all have advantages and disadvantages- really need to try all formats and see what works/who participates- The graduate Computational Biology and Bioinformatics program could assist/coordinate????
- ✓ Should be short talks followed by interactions
- ✓ On-line forum organized by categories with open questions and moderators

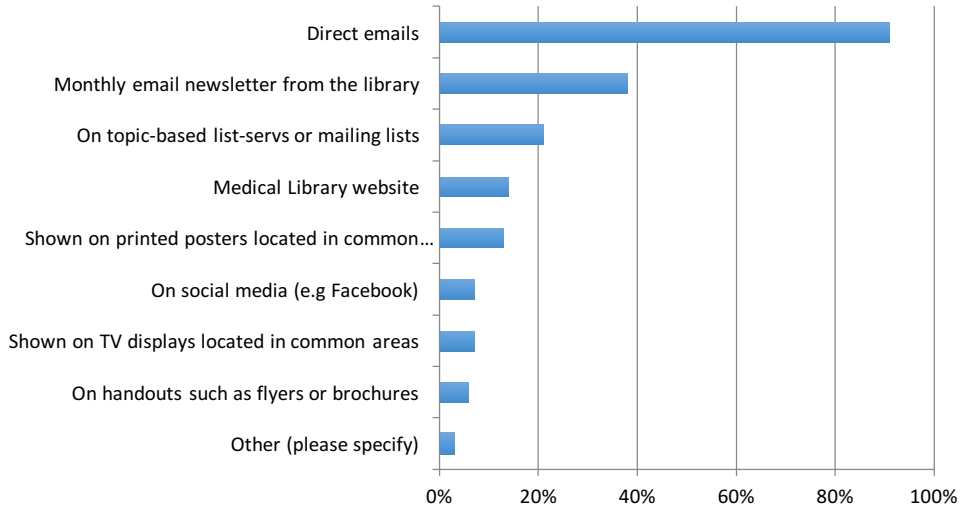


Figure S18. What is the best way to inform you about future data/information-related events (training, workshops, networking) offered by the Yale Medical Library? Check all that apply.

Conclusions and Future Steps

For two years, the Cushing/Whitney Medical Library has been providing end-user bioinformatics support to Yale biomedical researchers in the tertiary stage of the data analysis, and annotation in the form of training, consultations, and access to databases and tools (Ingenuity Pathway Analyis, MetaCore). During this time, 1329 Yale affiliates have attended (out of 2282 registered) the end-user bioinformatics training sessions organized and offered by the Medical Library. As a result, a total of six end-user commercial bioinformatics resources are currently supported by the Medical Library, and this year we began to explore the support of secondary data analysis of high throughput data with the addition of Partek Flow software (for the analysis of RNAseq data).

As we move into the third year, an evaluation of the end-user bioinformatics program is necessary in order to determine what are the gaps and services to improve.

We will continue to support the secondary data analysis stage in terms of end-user bioinformatics tools (Partek Flow, Qlucore, etc.) and training. This will allow researchers to move faster into the tertiary or downstream data analysis and use the tools already provided by the Medical Library.

Finally, we will continue to explore the role of the Medical Library in support of Precision Medicine. This will include an information/data needs assessments, as well as collaborating with Yale biomedical researchers in identifying and organizing presentations on available tools for the analysis of data and information.

References

Begley, C. G., & Ioannidis, J. P. (2015). Reproducibility in science: improving the standard for basic and preclinical research. *Circ Res, 116*(1), 116-126. doi:10.1161/CIRCRESAHA.114.303819

Burke, W., & Korngiebel, D. M. (2015). Closing the Gap between Knowledge and Clinical Application: Challenges for Genomic Translation. *PLoS Genet, 11*(2). doi:ARTN e1004978

DOI 10.1371/journal.pgen.1004978

Collins, F. S., & Varmus, H. (2015). A new initiative on precision medicine. *N Engl J Med, 372*(9), 793-795. doi:10.1056/NEJMp1500523

Dogan, R. I., Murray, G. C., Neveol, A., & Lu, Z. Y. (2009). Understanding PubMed (R) user search behavior through log analysis. *Database-the Journal of Biological Databases and Curation*. doi:ARTN bap018

10.1093/database/bap018

Gomez-Cabrero, D., Abugessaisa, I., Maier, D., Teschendorff, A., Merkenschlager, M., Gisel, A., . . . Tegner, J. (2014). Data integration in the era of omics: current and future challenges. *BMC Syst Biol, 8 Suppl 2*, I1. doi:10.1186/1752-0509-8-S2-I1

Kashef, Z. (2015). Yale launches national study of personalized medicine for metastatic melanoma. Retrieved from http://news.yale.edu/2015/04/15/yale-launches-national-study-personalized-medicine-metastatic-melanoma

Kim, T. Y., Kim, H. U., & Lee, S. Y. (2010). Data integration and analysis of biological networks. *Curr Opin Biotechnol, 21*(1), 78-84. doi:10.1016/j.copbio.2010.01.003

Kumar, S., & Dudley, J. (2007). Bioinformatics software for biologists in the genomics era. *Bioinformatics, 23*(14), 1713-1717. doi:10.1093/bioinformatics/btm239

Lynch, T. J. (2015). Precision Medicine: A Promising Future for Treating Cancer -. Retrieved from http://www.onclive.com/sap-partner/cancer-centers/yale-cancer/Precision-Medicine-A-Promising-Future-for-Treating-Cancer

Schneider, M. V., Watson, J., Attwood, T., Rother, K., Budd, A., McDowall, J., . . . Brooksbank, C. (2010). Bioinformatics training: a review of challenges, actions and support requirements. *Brief Bioinform, 11*(6), 544-551. doi:10.1093/bib/bbq021

Tan, T. W., Lim, S. J., Khan, A. M., & Ranganathan, S. (2009). A proposed minimum skill set for university graduates to meet the informatics needs and challenges of the "-omics" era. *BMC Genomics, 10 Suppl 3*, S36. doi:10.1186/1471-2164-10-S3-S36

Vicini, P., Fields, O., Lai, E., Litwack, E. D., Martin, A. M., Morgan, T. M., . . . Sogaard, M. (2016). Precision medicine in the age of big data: The present and future role of large-scale unbiased sequencing in drug discovery and development. *Clin Pharmacol Ther, 99*(2), 198-207. doi:10.1002/cpt.293

Villaveces, J. M., Koti, P., & Habermann, B. H. (2015). Tools for visualization and analysis of molecular networks, pathways, and -omics data. *Adv Appl Bioinform Chem, 8*, 11-22. doi:10.2147/AABC.S63534

Wang, J., Zuo, Y., Man, Y. G., Avital, I., Stojadinovic, A., Liu, M., . . . Ressom, H. W. (2015). Pathway and network approaches for identification of cancer signature markers from omics data. *J Cancer, 6*(1), 54-65. doi:10.7150/jca.10631

Contact Information

John Gallagher
Library Director
203-785-5352
john.gallagher@yale.edu

Janis Glover
Head of Curriculum and Research Support Department
(203) 737-2962
janis.glover@yale.edu

Rolando Garcia-Milian
Biomedical Sciences Research Support
203-785-6194
rolando.milian@yale.edu

Denise Hersey
Clinical Support Librarian
203-785-6251
denise.hersey@yale.edu

Head of Collection Development and Management
203-785-2883
nathan.rupp@yale.edu

Lei Wang
Assistant Director for Technology & Innovation
203-785-6485
lei.wang@yale.edu

Yale *Harvey Cushing / John Hay Whitney Medical Library*