FACILITIES AND OTHER RESOURCES

THE JACKSON LABORATORY

The Jackson Laboratory (JAX) is an independent, non-profit organization focusing on mammalian and human genomics research to advance human health. The mission of its *newest institute*, *The Jackson Laboratory for Genomic Medicine* (JAX-GM), is to discover the precise genomic causes of disease and develop individualized diagnostics, treatments and cures by merging the Laboratory's eight decades of research in mammalian genetics with those of the JAX-GM faculty and with the clinical expertise of Connecticut's universities and hospitals. JAX-GM has amassed a diverse array of technologies, computing capabilities, resources, core scientific services and support services to facilitate the research in the new, state-of-the art, 183,000 ft^2 genomics research facility, co-located on the University of Connecticut Health Center (UConn Health) campus in Farmington, CT.

The new facility includes ~26,300 ft^2 of wet laboratory space for molecular, cell, and genome biology-based research; ~8,600 ft^2 for interactive bioinformatics research clusters; a world-class, data center; microscopy facilities and a diverse array of scientific service core facilities led by dedicated experts in their respective fields, who provide guidance and expertise on approach, platform, experimental design and data analysis and deliver state-of-the-art research technologies, analytical tools and expertise. JAX cores house diverse cutting-edge platforms, including next-generation high-throughput sequencers; light, confocal and nano-resolution microscopes; slide imagers; flow cytometers; single-cell-based platforms and specialty support equipment. The data center houses high-performance computing equipment and high-capacity computing and storage devices.

Office and Meeting Space: All JAX investigators have the assigned private office space, furnishings and requisite workstation and laptop computers and printers to perform the proposed project. The offices for the key personnel at JAX are located in close proximity to the laboratory and computational dry spaces for their research programs. The investigators also have access to conference rooms that are equipped with real-time 'multi-point' videoconferencing platform to enable videoconference meetings with collaborators.

JAX Computational Sciences (CS) staff on both campuses works collaboratively with JAX-GM researchers in the application of advanced computational and analytical approaches to complex, data-intensive biological problems; in the development of scientific software applications that facilitate access, visualization and sharing of data and algorithms with the scientific community; and in the development of new scientific projects and platforms. Overall, CS has a staff of 26, including 15 PhD-level scientists and software engineers. The expertise ranges from analysis of all types of –OMICS data (e.g. epigenomics, genomics, transcriptomics, ribosomics, metabolomics & proteomics, ncRNAs, etc.), biophysical modeling, genomic data and model visualization, development of scientific software applications, database design & development and web interface development. CS has in-depth expertise in understanding and interpreting data in various biological domains including cancer biology, immunology, chromatin modeling, clinical genomics, metabolic disorders, and statistical genetics. CS also develops and maintains several computational platforms to enable JAX-GM researchers to efficiently analyze, query, visualize and share the scientific data. They are summarized below.

Applications Platform: Many application software packages have been installed for research use including Perl, BioPerl, Java, Python, and OpenMPI. Also several genomics relevant software modules related to Python and Perl have been deployed. They serve the development of computational and predictive genomics procedures and algorithms. The databases are supported by MySQL and file repositories.

Analytics Platform: Analytics platforms have been installed and developed to facilitate the statistical and computational inferences from the genomics projects. Many software packages that facilitate next-generation sequencing have been installed and working: BEDtools; bwa; GATK; ncbi-blast; MACS; clustalw; rsem/1.1.21; cs-util samtools; RNAshapes; cufflinkS; snpEff; RNAstructure; cufflinks; tabix; Rnall; tophat/1.3.0; SOAPaligner; fastqc; SOAPdenovo; fastx; SOAPsnp; gmap-gsnap; unafold; ViennaRNA; mctools; velvet; bismark; mfold; xenome;blat; mira; xlwt; bowtie; mirdeep;bowtie2; SMRT for PacBio technology.

Comprehensive Genome Analytics (*CGA)* is a platform that has been developed to facilitate reproducible plug-and-play analysis of genomics data, esp. to benchmark the analytical workflows and to facilitate CLIA-compliant test procedures for genomic medicine. CGA includes bioinformatics processes to identify all types of genomic variants, from targeted and whole-genome sequencing, and quantification and analysis of RNA-seq data from both direct patient samples and patient-derived xenograft (PDX) samples. CGA also includes a database and user interface to facilitate complex queries on raw data and results. CS has also developed a clinical curation knowledge base and interface to improve overall speed and performance of variant curation. CGA is an area of active method development, e.g. CS has successfully deployed a PacBio hybrid sequencing analytics platform for the analysis of isoform expression data.

As an extension of CGA, JAX is piloting studies with commercial partners to evaluate platforms for improved genomic sequencing analysis workflow management including hybrids of local HPC and cloud approaches. These projects build off of our longtime expertise in Galaxy, a public platform that facilitates construction of generic workflows and analysis of genomics data such as whole genome sequencing, exome-seq and RNA-seq.

Analytical Software: In addition to the genomics-specific analytical software packages and platforms described above, CS maintains, applies, and further develops many analytical software packages for focused data analysis. These include JMP, R, Octave, MATLAB, and Prism. Related to these analytical software activities, Computational Sciences also performs in-house development of novel software tools, e.g. our Computational Metabolomics Platform, which has been developed to efficiently and accurately process, analyze, retrieve and visualize metabolomics data.

Information Technology (IT) is the division at JAX focused on technology support for all computational projects. The infrastructure is supported by IT teams at each of the JAX-GM and Bar Harbor locations. Users can get IT support 24 hours a day. The staff of the Computational Sciences Service provides support for custom analysis and software development. The IT platforms are summarized below.

High-Performance Computing (HPC) Platform: Our HPC Platform includes 96 HP Proliant SL Series servers with 16/20 cores and 256 GB of memory per node, as well as two dedicated high-memory nodes (64 cores, 1 TB memory) for a total of 1856 compute cores. The computing infrastructure enables analysis and inference in genomics medicine projects with a wide range of complexity and computational intensities. The platform also includes a high-memory HP Proliant DL900 Series server of 2.2GHz-48Core-2TB for memory-intensive genomic computations, such as *de novo* transcriptome assembly. The operating system deployed for all servers in this platform is CentOS. These resources will be expanded as the needs grow.

Storage Platform: HPC at JAX is supported by the Isilion scale-out NAS, with 2PB of storage available for processing and analysis of scientific data. A geographically dispersed archiving system is being implemented for securing raw instrument data. These resources will be expanded as the needs grow.

Data Transfer: JAX has invested in a high-speed network architecture consisting of a 40Gb/s backbone with 10Gb/s to each server, to account for expedient transfer of large genomic data sets. In addition, we have a third-party Internet service provider that can scale beyond 1Gb as demand increases. We have an Aspera high-speed file transfer system available for the transmission of large files.

Applications Platform: Our Applications platform supports all of the standard software necessary for investigators to process and analyze their data, as well as providing the entire community with the basic blocks for them to build their own custom workflows. Database development and deployment is supported by MySQL and other database management systems.

Network Infrastructure: Our Networks platform supports both scientific and enterprise business systems, using 40Gb core switches in our server farm that delivers at least 1Gb to user devices. The environment includes wired and wireless network service and a redundant voice over IP (VOIP) system, and is protected by application firewalls. The network infrastructure for the HPC environment is comprised of a 40Gb backbone with 10Gb to each server in the cluster. Internet service is delivered by a commercial service provider that can scale beyond 1Gb as demand for data transfer increases.

Conferencing/audiovisual services: JAX-GM houses almost 10,000 ft^2 of space devoted to conferences, training and collaboration. This includes an auditorium, designed to seat 202 guests, for invited speakers and special symposia. Multiple conference rooms (4–6 per floor, 155–278 ft^2 each) are available on every floor to support laboratory meetings, discussions and communications among scientists. Meetings areas and conference rooms, including, Charles Lee's office, are equipped for real-time videoconferencing using a 'multi-point', high-definition videoconferencing compatible with Macintosh or PC computers and a variety of widely used mobile wireless devices from any site with an Internet connection, which will facilitate frequent collaborative communication with colleagues at the Yale University and McDonnell Genome Institute at Washington University in ST Louis sites.

THE MCDONNELL GENOME INSTITUTE, WASHINGTON UNIVERSITY IN ST. LOUIS

McDonnell Genome Institute

The McDonnell Genome Institute is located at 4444 Forest Park Ave., on the northeast corner of the Washington University Medical Center. Other institutions at the Medical Center include Washington University School of Medicine, Barnes-Jewish and St. Louis Children's Hospitals, Central Institute for the Deaf, Mallinckrodt Institute of Radiology and Imaging, the Goldfarb School of Nursing, and St. Louis College of Pharmacy. Currently, The McDonnell Genome Institute occupies 56,660 square feet of space for laboratory and administrative personnel. This space was designed specifically to accommodate production-sequencing activities and includes specialized equipment to maintain strict power and temperature requirements. A 900 square-foot technology development laboratory is available for testing prototype equipment and developing new hardware and biochemistry.

The CLIA licensed environment (CLE) occupies 2,412 sq. ft. of space. As a replicate of the production infrastructure necessary to achieve the expected sequencing goals, the CLE maintains all pertinent equipment within this space. This allows for the completion of all exome capture and sequence generation in a controlled environment. The access to the environment is managed by additional security provided by swipe card and keypads on all doors. The space includes laboratory benches, equipment bays, and desk cubicles for laboratory technicians and managers.

Sequencing & Data Production: The McDonnell Genome Institute operates a state-of-the-art high-throughput, high capacity, next-generation sequencing facility. Our library construction core has the capability to generate 1500 dual-indexed small insert Illumina libraries/week. Multiple quality control measures ensure precise control over insert size and library complexity. We utilize a set of 96 unique dual-indexed library adaptors to allow efficient multiplex processing of samples for hybrid capture and/or DNA/RNA sequencing. Our dual-indexed adaptors with matching 6-8 bp index sequences at both ends prevent cross talk during amplification and/or cluster identification on the Illumina instruments and ensure precise matching of DNA sequences with the sample of origin. The sequencing core currently includes 10 HiSeq Xs, 12 Illumina HiSeq 2000s, and 4 HiSeq 2500s instruments. Our Illumina HiSeq capacity is approximately 215 HiSeq flow cells per/month, which is equivalent to ~1720 human genomes with 30X sequence coverage and 475 HiSeq 2500 flow cells per/month, which is equivalent to 3,800 human exomes (Nimblegen VCRome; 38 Mb space with 80% of the target space covered at 20X depth with average mean depth of 60X coverage). We also have 4 Illumina MiSeq instruments, 3 Applied Biosystems 3730 XL instruments, a Pacific Biosystems RS2 system, and an Ion Torrent PGM sequencer. Robotic systems for sample processing include Perkin Elmer Sciclone G3s, Perkin Elmer Janus system, Eppendorf EpMotion 5075, Covaris LE220 DNA Sonicator, and Perkin Elmer LabChip XT. Sample processing is managed by a custom Oracle based laboratory information management system. The McDonnell Genome Institute has over 100 touchscreen/bar code scanner stations throughout the building which enable data entry and tracking. All sample containers are physically bar code labeled and continuously tracked to monitor progress and ensure sample integrity at every step of the workflow.

Computing Facilities: The McDonnell Genome Institute has a 15,600 sq. ft. state-of-the-art data center that is located across the street from our main building and was completed in 2010 (222 S. Newstead Ave.). The data center contains fully redundant power and cooling systems capable of housing over 100 racks of high-density network, server and storage systems in its 3,100 sq ft raised floor computer room. Electrical power to the facility is supplied by a nearby, double-ended utility power substation with a backup generator. Redundant cooling is supplied by chilled water systems, delivered under the floor. The center has office accommodations, badge secured entry, secuity cameras and a receiving dock. This data center is the first building on the School of Medicine's campus to receive the LEED Gold status by the US Green Building Council. This was a major challenge for architects and engineers who designed the building because of the energy requirements of specialized cooling systems for the computer equipment. In addition to this data center, a legacy 1,200 square-foot server room at the McDonnell Genome Institute's location (4444 Forest Park Ave.) is equipped with raised floors, redundant power and cooling is utilized for equipment with low power and cooling requirements.

The McDonnell Genome Institute maintains a 1 Gigabit external network link protected by a modern firewall and a 10 Gigabit link on the Internet 2 research network, which is protected by a central Washington University router with a whitelist of collaborating institutions such as CGHub and NCBI sequence data repositories. Within our network, the McDonnell Genome Institute has a highly scalable storage system consisting of over 16 petabytes of raw data storage spread across 23 disk controllers organized into 6 clusters based on usage patterns on our 16 Gigabit SAN network. In addition, MGI has a high performance and highly expandable tape robot managing a tape library of 5 petabytes, which allows the shuttling of data from live disk to much less expensive tape and back again on demand. The McDonnell Genome Institute's computational cluster has 469 servers, 4,774 cores, and 1.1 petabytes of RAM, and runs on average 2.5 million individual computational jobs per month equal to 131 years wall clock time. The newest computational servers have dual 10 Gigabit network links, 40 cores with hyper-threading, 384 gigabyte RAM and 3.2 terabyte local SSD storage per node, which are networked to a redundant 40 Gigabit Ethernet backplane to each storage node. These computational servers and 162 additional operational servers within our computing facilities are managed with automation tools such as 1) PXE+Kickstart for image distribution and boot 2) Puppet+Git+mcollective for server build configuration, change control, and deployment 3) Jenkins+Git for continues integration. The MGI also manages large database instances of Oracle, PostgreSql, and MySQL and utilize many monitor systems such as Zenoss, Nagios, Graphite, Logstash, RTM (LSF), Netflow Collector, OSSEC, Piwik Web Analytics, DBTuna and Google Analytics for maintaining stability, troubleshooting and tuning our systems. To insure continuity of services in the case of a disaster, we have defined service level agreements and nightly backups of critical data, which are stored monthly and retained for one year at an off-site location.

Washington University School of Medicine: The Washington University School of Medicine, consistently ranked in the top 5 medical schools in the United States by U.S. News & World Report and by funding from the National Institutes of Health, has a rich, 122-year scientific history in basic, clinical, and translational research. The Medical School is organized into 20 Departments, 14 clinical Departments and 6 basic science Departments, and includes a total of 1,874 faculty and 1,349 students. Since its founding in 1891, it has trained nearly 8,000 physicians and has contributed groundbreaking discoveries in many areas of medical research. The Medical School also has robust clinical translational infrastructure through its 30 program project or center grants funded by the National Institutes of Health. The School's faculty members are the staff physicians at Barnes-Jewish Hospital and St. Louis Children's Hospital that form the academic hub for the 5,252-bed BJC HealthCare System, the Medical School's hospital partner. The School of Medicine and these fine hospitals, which are perennially recognized for excellence in patient care by U.S. News & World Report and also provide a superb atmosphere for collaborative translational research and for training students, residents, and fellows, are the principal components of the Washington University Medical Center. The compact nature of this 230-acre academic medical center in 12 city blocks enhances the collaborative opportunities for translational research.

YALE UNIVERSITY

Yale

Gerstein Laboratory: The Gerstein laboratory is found in two connected buildings (Bass Central/Main campus). The laboratory consists of 6 rooms and comprises a total of ~1,900 sq. ft. In addition, three conference rooms that have projectors provide venues for interaction. There are 40 gigabit-ready desks, equipped with one or two 23" and 30" LCD screens. The space is properly air conditioned for supporting a large number of computers, including forty-seven working laptops in the lab, of which eighteen are recent Macbook Pro models. Mark Gerstein's office space is 178 sq. ft.

Computer Infrastructure

Laboratory Network and Storage: The lab computing infrastructure is partitioned into a private and a public network. The entire infrastructure is fully gigabit capable and is connected to the Yale backbone via gigabit optic fibre; the network architecture was designed with computing efficiency and network security in mind. The private network consists of individual laptops, desktops and workstations, as well as communal computational servers, dumb terminals, a central fileserver, a consolidated NAS, and printers. There are also servers that provide essential network services such as NIS, NFS, SMB, DHCP, monitoring and backups. The public network consists of numerous production webservers that are either real or virtual machines. The laboratory maintains its own public subnets of 128 public IP addresses and manages many of its own domains (e.g. gersteinlab.org, molmovdb.org, pseudogenes.org, and partslist.org). The lab has a full-time administrator maintaining the network.

The private and public networks obtain gigabit connectivity through four HP Procurve 5300xl switches that are mutually connected via fibre. The private network is behind a Cisco PIX 525, which is concurrently used as an IPSec VPN gateway into the private network. Within the private network are two NetApp storage appliances with 43Tb of raw space, which is configured with 27.5Tb of working space, thirty custom made 4Tb network disks with a total 120Tb capacity, a Dell NAS with a total of 30TB capacity; the NetApp appliances and Dell NAS are used for live user file space, backups of user files and backups of public production webservers. A seven-day incremental backup and a twelve-month incremental backup are currently being implemented in the lab. Wireless access is available all throughout the lab. Wireless access connects computers directly to the public network.

In total, the lab has 315u of rack space spread over eight racks. Residing in these racks are a dual CPU twelve core Opteron server with 256GB of memory, a dual CPU six core Opteron server with 128GB of memory, a dual CPU four core Opteron server with 64GB of memory, three Intel blade enclosures with 10 dual CPU Intel blades each, fourteen dual cpu 64 bit Xeons servers and six dual cpu 64 bit Opteron servers; these rack servers are in addition to the NetApp storage appliances and the Dell NAS mentioned above. The rack servers have various uses. The dual CPU Opteron servers are for hosting virtual machines, which function as web hosts. In the private network, five rack servers are for essential network services, four are storage head nodes for the Dell SAN and a few are network support or experimental machines. The rest of the rack servers are in the public network acting as webservers. The private network has seven business class color laserjet printers.
Software. A number of open source software, programs created in-house, and proprietary software is used by the lab researchers for their needs. The lab maintains a set of wiki servers for the documentation of internal information and the public dissemination of information. The lab also manages mailman servers for its mailing lists. The compute nodes are mainly used to develop and run Java and Perl code and to perform Matlab and Gromacs calculations. The public webservers are used to deploy Java, Perl, PHP and Python applications. Individual tasks are coordinated by a web group calendar. Web applications and servers are continually being monitored by a Nagios monitoring system.

Yale Life Sciences Supercomputer: The Gerstein laboratory has priority access to two of the Yale supercomputers, namely Louise and BulldogI, and regular access to six other Yale supercomputers. There are two full-time administrators maintaining the supercomputer. Louise is a cluster with 112 Dell PowerEdge R610 with (2) quad core E5620 nodes, each with 2.4 Ghz cpu cores and 48 GB RAM. They are interconnected with a Force10 network switch. There is therefore a total of 112*8 cores = 896 cores. Louise has 300 TB (raw) of BlueArc parallel file storage.

BulldogI is a cluster consisting of a head node and 170 Dell PowerEdge 1955 nodes, each containing 2 dual core 3.0 Ghz Xeon 64 bit EM64T Intel cpus, for a total of 680 cores. Each node has 16 GB RAM. The network is Gigabit ethernet. Bulldogi runs a high performance Lustre filesystem. It is managed via PBS. Three 20Tb Dell Power Vault with storage arrays are attached to BulldogI and are dedicated for Gerstein laboratory use.
The laboratory also has priority access to a SGI F1240 system. This system has 12 Xeon E5345 Quad-Core 2.33GHz CPUs (for a total of 48 processor cores), with 2 x 4M L2 cache per CPU, a 1333MHz front side bus, 96GB of memory, and 6 Raptor 150GB, 10K rpm SATA drives. It runs SUSE Linux Enterprise Server 10 as a system single image. That is, all 48 cores are managed by a single process scheduler, and the 96 GB memory is, in principle, addressable by a single process. In practice, system caches and buffers reduce the maximum

amount of memory available to any given process to about 70 GB. In many ways then, the system can be thought of as an SMP, but in terms of hardware architecture it is closer to an infiniband-connected cluster.

Core Lab: The Gerstein Lab is adjacent to the Yale Center for Structural Biology (CSB) Core laboratory. The Core laboratory resources are available to members of the Gerstein lab. The Core laboratory supports the work of all the people associated with the CSB, in total about 200 users and >200 computers. These computers include a number of high-performance graphics workstations for visualizing macromolecular structures and complex data sets. The CSB Core staff of 2 FTE provides support to the associated CSB laboratories as well as the Core computers.

Oracle Server: Yale University has an institutional site license for the Oracle database management system. As a result, many major administrative computing systems at Yale are being developed using Oracle, and Yale's ITS staff has extensive Oracle experience. Yale ITS maintains and operates several Oracle database systems at the School of Medicine, and provides access to these machines to many different projects. There are several advantages to using institutional servers. The ITS staff backs up each database on a regular schedule, typically with full backups weekly and partial backups several times a day. The ITS staff maintains the hardware of the database machine, the system software, and the Oracle software. They perform periodic upgrades when new versions of the software become available. They also handle any systems problems that occur, and are available to help troubleshoot any application problems that arise