OMB Number: 4040-0001
Expiration Date: 06/30/2011

APPLICATION FOR FEDERAL ASSISTANCE

SF 424 (R&R)

3. DATE RECEIVED BY STATE

State Application Identifier
ME: Maine

1. TYPE OF SUBMISSION

○ Pre-application ● Application ○ Changed/Corrected Application

4.a. Federal Identifier

b. Agency Routing Number

2. DATE SUBMITTED
2016-07-05

Application Identifier
00002279

c. Previous Grants.gov Tracking ID

5. APPLICANT INFORMATION **Organizational DUNS:**042140483

Legal Name: The Jackson Laboratory
Department: N/A Division: N/A
Street1: 600 Main Street
Street2:
City: Bar Harbor County/Parish: Hancock
State: ME: Maine Province:
Country: USA: UNITED STATES ZIP / Postal Code: 04609-1523

Person to be contacted on matters involving this application

Prefix:	First Name:	Middle Name:	Last Name:	Suffix:
	Adam	C	Carter	

Position/Title: Manager, Pre-Award Sponsored Programs
Street1: 600 Main Street
Street2:
City: Bar Harbor County/Parish: Hancock
State: ME: Maine Province:
Country: USA: UNITED STATES ZIP / Postal Code: 04609-1523
Phone Number: 207-288-6914 Fax Number: 207.288.6053 Email: adam.carter@jax.org

6. EMPLOYER IDENTIFICATION NUMBER *(EIN) or (TIN)*: 1010211513A1

7. TYPE OF APPLICANT M: Nonprofit with 501C3 IRS Status (Other than Institution of Higher Education)

Other (Specify):
Small Business Organization Type ○ Women Owned ○ Socially and Economically Disadvantaged

8. TYPE OF APPLICATION:

● New ○ Resubmission
○ Renewal ○ Continuation ○ Revision

If Revision, mark appropriate box(es).
○ A. Increase Award ○ B. Decrease Award ○ C. Increase Duration
○ D. Decrease Duration ○ E. Other *(specify)*:

Is this application being submitted to other agencies? ○Yes ●No What other Agencies?

9. NAME OF FEDERAL AGENCY:
National Institutes of Health

10. CATALOG OF FEDERAL DOMESTIC ASSISTANCE NUMBER:
TITLE: NHLBI TOPMed Program: Integrative Omics Approaches for Analysis of TOPMed Data (U01)

11. DESCRIPTIVE TITLE OF APPLICANT'S PROJECT:
Identification, Annotation and Phenotypic Association of Variants and Indels in Complex Diseases

12. PROPOSED PROJECT:

Start Date	Ending Date
04/01/2017	03/31/2020

13. CONGRESSIONAL DISTRICTS OF:
a. Applicant
CT-005

14. PROJECT DIRECTOR/PRINCIPAL INVESTIGATOR CONTACT INFORMATION

Prefix:	First Name:	Middle Name:	Last Name:	Suffix:
	Charles		Lee	

Position/Title: Scientific Director, JAX GM Organization Name: The Jackson Laboratory
Department: N/A Division: Genomic Medicine
Street1: 10 Discovery Dr.
Street2:
City: Farmington County/Parish: Hartford
State: CT: Connecticut Province:
Country: USA: UNITED STATES ZIP / Postal Code: 06032-2352
Phone Number: 860-856-2458 Fax Number: 207-288-6053 Email: charles.lee@jax.org

OMB Number: 4040-0001
Expiration Date: 06/30/2011

SF 424 (R&R) APPLICATION FOR FEDERAL ASSISTANCE Page 2

15. ESTIMATED PROJECT FUNDING

a. Total Federal Funds Requested	$2,072,221.42
b. Total Non-Federal Funds	$0.00
c. Total Federal & Non-Federal Funds	$2,072,221.42
d. Estimated Program Income	$0.00

16.IS APPLICATION SUBJECT TO REVIEW BY STATE EXECUTIVE ORDER 12372 PROCESS?

a. YES ○ THIS PREAPPLICATION/APPLICATION WAS MADE AVAILABLE TO THE STATE EXECUTIVE ORDER 12372 PROCESS FOR REVIEW ON:

DATE:

b. NO ● PROGRAM IS NOT COVERED BY E.O. 12372; OR

○ PROGRAM HAS NOT BEEN SELECTED BY STATE FOR REVIEW

17.By signing this application, I certify (1) to the statements contained in the list of certifications* and (2) that the statements herein are true, complete and accurate to the best of my knowledge. I also provide the required assurances * and agree to comply with any resulting terms if I accept an award. I am aware that any false, fictitious, or fraudulent statements or claims may subject me to criminal, civil, or administrative penalties. (U.S. Code, Title 18, Section 1001)

● I agree

** The list of certifications and assurances, or an Internet site where you may obtain this list, is contained in the announcement or agency specific instructions.*

18. SFLLL (Disclosure of Lobbying Activities) or other Explanatory Documentation

19. Authorized Representative

Prefix:	First Name:	Middle Name:	Last Name:	Suffix:
	Adam	C	Carter	

Position/Title: Interim Director, Sponsored Program
Organization Name: The Jackson Laboratory
Department: N/A
Division:
Street1: 600 Main Street
Street2:
City: Bar Harbor
County/Parish:
State: ME: Maine
Province:
Country: USA: UNITED STATES
ZIP / Postal Code: 04609-1523
Phone Number: 207-288-6914
Fax Number: 207-288-6053
Email: grants@jax.org

Signature of Authorized Representative

Date Signed

07/05/2016

20. Pre-application File Name: Mime Type:
21. Cover Letter Attachment File Name: Mime Type:

Project/Performance Site Location(s)

Project/Performance Site Primary Location

❍ I am submitting an application as an individual, and not on behalf of a company, state, local or tribal government, academia, or other type of organization.

Organization Name:	The Jackson Laboratory for Genomic Medicine
Duns Number:	042140483
Street1*:	10 Discovery Drive
Street2:	
City*:	Farmington
County:	Hartford
State*:	CT: Connecticut
Province:	
Country*:	USA: UNITED STATES
Zip / Postal Code*:	06032-2352
Project/Performance Site Congressional District*:	CT-005

Project/Performance Site Location 1

❍ I am submitting an application as an individual, and not on behalf of a company, state, local or tribal government, academia, or other type of organization.

Organization Name:	Washington University in St. Louis
DUNS Number:	068552207
Street1*:	660 South Euclid Avenue
Street2:	
City*:	St. Louis
County:	St. Louis
State*:	MO: Missouri
Province:	
Country*:	USA: UNITED STATES
Zip / Postal Code*:	63110-1093
Project/Performance Site Congressional District*:	MO-001

Project/Performance Site Location 2

❍ I am submitting an application as an individual, and not on behalf of a company, state, local or tribal government, academia, or other type of organization.

Organization Name:	Yale University
DUNS Number:	043207562
Street1*:	Grant & Contract Administration
Street2:	47 College Street - Suite 203
City*:	New Haven
County:	New Haven
State*:	CT: Connecticut
Province:	

Country*: USA: UNITED STATES

Zip / Postal Code*: 06520-8047

Project/Performance Site Congressional District*: CT-003

File Name

Additional Location(s)

OMB Number: 4040-0001
Expiration Date: 02/28/2013

RESEARCH & RELATED Other Project Information

1. Are Human Subjects Involved? ○ **Yes** ● **No**

1.a. If YES to Human Subjects

Is the Project Exempt from Federal regulations? **Yes** **No**

If YES, check appropriate exemption number: __ 1 __ 2 __ 3 __ 4 __ 5 __ 6

If NO, is the IRB review Pending? ○ **Yes** ○ **No**

IRB Approval Date:

Human Subject Assurance Number

2. Are Vertebrate Animals Used? ○ **Yes** ● **No**

2.a. If YES to Vertebrate Animals

Is the IACUC review Pending? ○ **Yes** ○ **No**

IACUC Approval Date:

Animal Welfare Assurance Number

3. Is proprietary/privileged information included in the application? ○ **Yes** ● **No**

4.a. Does this Project Have an Actual or Potential Impact - positive or negative - on the environment? ○ **Yes** ● **No**

4.b. If yes, please explain:

4.c. If this project has an actual or potential impact on the environment, has an exemption been authorized or an environmental assessment (EA) or environmental impact statement (EIS) been performed? ○ **Yes** ○ **No**

4.d. If yes, please explain:

5. Is the research performance site designated, or eligible to be designated, as a historic place? ○ **Yes** ● **No**

5.a. If yes, please explain:

6. Does this project involve activities outside the United States or partnership with international collaborators? ○ **Yes** ● **No**

6.a. If yes, identify countries:

6.b. Optional Explanation:

7. Project Summary/Abstract	M-2_Project_Summary.pdf	Mime Type: application/octet-stream
8. Project Narrative	M-3_Narrative.pdf	Mime Type: application/octet-stream
9. Bibliography & References Cited		
10. Facilities & Other Resources	M-4_Facilities.pdf	Mime Type: application/octet-stream
11. Equipment	M-5_Equipment.pdf	Mime Type: application/octet-stream

EQUIPMENT

See the **Facilities and Other Resources** Section for the description of the computing, storage and network equipment that comprises the computing infrastructure at the three participating institutions - The Jackson Laboratory, The McDonnell Genome Institute, Washington University in St. Louis and Yale University.

PUBLIC HEALTH RELEVANCE

The underlying mechanisms of many of humanity's most challenging diseases are linked to specific and oftentimes complex alterations to an individual's genome. This project will develop computational resources and tools for discovering one type of common yet powerful genomic variation, known as structural variation, and for associating specific structural variants with disease. The analytical tools and resources we develop will be broadly applicable to studies of the genomic mechanisms of disease, towards the ultimate goal of improving human health.

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

PROJECT SUMMARY

Structural variations (SVs), such as deletions, duplications, inversions and translocations, are among the most significant determinants of human genetic diversity to have been discovered, affecting far more bases than single-nucleotide variants in the genome. The TOPMed Program offers an exciting new opportunity to mine genome sequencing and other -omics datasets from a large cohort of individuals for novel SV discovery, analysis and association with common and Mendelian diseases. However, SVs are inadequately covered by current computational discovery methods, making it likely that a large proportion of variants associated with human disease remain unidentified or poorly characterized. The overarching objectives of this project are to 1) discover SVs at high resolution and large scale, 2) functionally interpret SV origin and phenotypic effects and 3) associate SVs of high functional impact with disease. We bring together a team of pioneers with a proven record of collaboration and innovation in the field of SV discovery, genotyping and large-scale functional genome analysis. We will develop and integrate novel tools for high-resolution SV discovery and use these to comprehensively profile all types of SVs, including complex SVs, in a large subset of the genomes being sequenced (Aim 1). To examine the functional impact of the identified SVs, we will develop cutting-edge methodologies for functional annotation of variants and characterization of associated biological processes through integration of –omics datasets (Aim 2), which will also enable us to prioritize subsets of SVs for association studies proposed in Aim 3. Finally, we will scale up SV detection and analysis through genotyping of all SVs detected in Aim 1 across the 100K samples of the TOPMed Program, which will provide the necessary statistical power for meaningful genotype-phenotype associations for disease-based SV association studies (Aim 3). Our deliverables will be the largest library of validated SVs discovered in humans, together with an unprecedented and broadly applicable platform of cloud-based pipelines for comprehensive, high-resolution and large-scale SV analysis.

OMB Number: 4040-0001
Expiration Date: 06/30/2016

RESEARCH & RELATED Senior/Key Person Profile (Expanded)

PROFILE - Project Director/Principal Investigator

Prefix	* First Name	Middle Name	* Last Name	Suffix
	Charles		Lee	

Position/Title: Scientific Director, JAX GM

Department: N/A

Organization Name: The Jackson Laboratory

Division: Genomic Medicine

* Street1: 10 Discovery Dr.

Street2:

* City: Farmington

County: Hartford

* State: CT: Connecticut

Province:

* Country: USA: UNITED STATES

* Zip / Postal Code: 06032-2352

*Phone Number	Fax Number	* E-Mail
860-856-2458	207-288-6053	charles.lee@jax.org

Credential, e.g., agency login: CL1234

*** Project Role:** PD/PI

Other Project Role Category:

Degree Type: Doctor of Philosophy

Degree Year: 1996

	File Name	Mime Type
***Attach Biographical Sketch**	ID-6533_BN-1_BIOSKETCH.pdf	application/octet-stream
Attach Current & Pending Support		

PROFILE - Senior/Key Person_

Prefix	* First Name	Middle Name	* Last Name	Suffix
	Mark		Gerstein	

Position/Title: Professor

Department:

Organization Name: Yale University

Division: Bioinformatics

* Street1: PO Box 208114 MBB

Street2:

* City: New Haven

County: New Haven

* State: CT: Connecticut

Province:

* Country: USA: UNITED STATES

* Zip / Postal Code: 06520-0002

*Phone Number	Fax Number	* E-Mail
203-432-6105		mark.gerstein@yale.edu

Credential, e.g., agency login: MGERSTEIN

*** Project Role:** PD/PI

Other Project Role Category:

Degree Type: Doctor of Philosophy

Degree Year: 1993

	File Name	Mime Type
Attach Biographical Sketch	ID-1512_BN-1_BIOSKETCH.pdf	application/octet-stream
Attach Current & Pending Support		

PROFILE - Senior/Key Person_

Prefix	* First Name	Middle Name	* Last Name	Suffix
	Ankit		Malhotra	

Position/Title: Associate Computational Scientist

Department: N/A

Organization Name: The Jackson Laboratory

Division: Genomic Medicine

* Street1: 10 Discovery Drive

Street2:

* City: Farmington

County: Hartford

* State: CT: Connecticut

Province:

* Country: USA: UNITED STATES

* Zip / Postal Code: 06032-2352

*Phone Number	Fax Number	* E-Mail
860-837-2432	207.288.6053	ankit.malhotra@jax.org

Credential, e.g., agency login: ANKIT.MALHOTRA

*** Project Role:** Other (Specify)

Other Project Role Category: Co-Investigator

Degree Type: Doctor of Philosophy

Degree Year: 2010

	File Name	Mime Type
Attach Biographical Sketch	ID-6518_BN-1_BIOSKETCH.pdf	application/octet-stream
Attach Current & Pending Support		

PROFILE - Senior/Key Person_

Prefix	* First Name	Middle Name	* Last Name	Suffix
	Li		Ding	

Position/Title: Associate Professor

Department:

Organization Name: Washington University in St Louis

Division: Hematology and Oncology

* Street1: 4444 Forest Park Avenue

Street2: Campus Box 8501

* City: St. Louis

County: St. Louis City

* State: MO: Missouri

Province:

* Country: USA: UNITED STATES

* Zip / Postal Code: 63108-2212

*Phone Number	Fax Number	* E-Mail
314-286-1848		lding@genome.wustl.edu

Credential, e.g., agency login: DINGLI

*** Project Role:** PD/PI

Other Project Role Category:

Degree Type: Doctor of Philosophy

Degree Year: 1998

	File Name	Mime Type
Attach Biographical Sketch	ID-1511_BN-1_BIOSKETCH.pdf	application/octet-stream
Attach Current & Pending Support		

PROFILE - Senior/Key Person_

Prefix	* First Name	Middle Name	* Last Name	Suffix
	Michael		Wendl	

Position/Title: Associate Professor of Medicine

Department:

Organization Name: Washington University in St. Louis

Division: Hemotology and Oncology

* Street1: 4444 Forest Park Ave.

Street2: Campus Box 8501

* City: St. Louis

County: St. Louis City

* State: MO: Missouri

Province:

* Country: USA: UNITED STATES

* Zip / Postal Code: 63108-2212

*Phone Number	Fax Number	* E-Mail
314-286-1848	314-286-1810	mwendle@genome.wustl.edu

Credential, e.g., agency login: MCWENDL

*** Project Role:** Other (Specify)

Other Project Role Category: Consortium Co-Investigator

Degree Type: Doctor of Science

Degree Year: 1994

	File Name	Mime Type
Attach Biographical Sketch	ID-1845_BN-1_BIOSKETCH.pdf	application/octet-stream
Attach Current & Pending Support		

OMB Number: 4040-0001
Expiration Date: 06/30/2016

RESEARCH & RELATED Senior/Key Person Profile (Expanded)

Additional Senior/Key Person Form Attachments

When submitting senior/key persons in excess of 8 individuals, please attach additional senior/key person forms here. Each additional form attached here, will provide you with the ability to identify another 8 individuals, up to a maximum of 4 attachments (32 people).

The means to obtain a supplementary form is provided here on this form, by the button below. In order to extract, fill, and attach each additional form, simply follow these steps:

- Select the "Select to Extract the R&R Additional Senior/Key Person Form" button, which appears below.
- Save the file using a descriptive name, that will help you remember the content of the supplemental form that you are creating. When assigning a name to the file, please remember to give it the extension ".xfd" (for example, "My_Senior_Key.xfd"). If you do not name your file with the ".xfd" extension you will be unable to open it later, using your PureEdge viewer software.
- Using the "Open Form" tool on your PureEdge viewer, open the new form that you have just saved.
- Enter your additional Senior/Key Person information in this supplemental form. It is essentially the same as the Senior/Key person form that you see in the main body of your application.
- When you have completed entering information in the supplemental form, save it and close it.
- Return to this "Additional Senior/Key Person Form Attachments" page.
- Attach the saved supplemental form, that you just filled in, to one of the blocks provided on this "attachments" form.

Important: Please attach additional Senior/Key Person forms, using the blocks below. Please remember that the files you attach must be Senior/Key Person Pure Edge forms, which were previously extracted using the process outlined above. Attaching any other type of file may result in the inability to submit your application to Grants.gov.

1) Please attach Attachment 1	
2) Please attach Attachment 2	
3) Please attach Attachment 3	
4) Please attach Attachment 4	

	Filename
ADDITIONAL SENIOR/KEY PERSON PROFILE(S)	
	MimeType
	Filename
Additional Biographical Sketch(es) (Senior/Key Person)	
	MimeType
	Filename
Additional Current and Pending Support(s)	
	MimeType

BIOGRAPHICAL SKETCH

Provide the following information for the Senior/key personnel and other significant contributors.
Follow this format for each person. **DO NOT EXCEED FIVE PAGES.**

NAME: Lee, Charles

eRA COMMONS USER NAME (credential, e.g., agency login): CL1234

POSITION TITLE: Scientific Director and Professor, The Jackson Laboratory for Genomic Medicine

EDUCATION/TRAINING *(Begin with baccalaureate or other initial professional education, such as nursing, include postdoctoral training and residency training if applicable. Add/delete rows as necessary.)*

INSTITUTION AND LOCATION	DEGREE *(if applicable)*	Completion Date MM/YYYY	FIELD OF STUDY
University of Alberta, Alberta, Canada	B.S.	06/1990	Genetics
University of Alberta, Alberta, Canada	M.Sc.	06/1993	Experimental Pathology
University of Alberta, Alberta, Canada	Ph.D.	06/1996	Medical Sciences
Cambridge University, Cambridge, UK	Postdoc	06/1998	Molecular Cytogenetics
Harvard Medical School, Boston, MA USA	Postdoc	06/2001	Clinical Cytogenetics

A. Personal Statement

My career in human genomics has spanned over 15 years and began with a simple premise of understanding more about the human genome and how seemingly insignificant variations in DNA could give rise to different traits and phenotypes. One of our notable discoveries came by way of the first description of widespread structural variants (SVs) – in the form of copy number variants (CNVs) – in the human genome. We demonstrated that large genomic changes (i.e., gains or losses of thousands-to-millions of nucleotides) are common, pervasive, and often hold important implications for understanding the genetic basis of human health and disease. We subsequently developed two human CNV maps that serve the basis of our understanding of common structural human genomic variants and are routinely used for differentiating pathogenic genomic imbalances from common variants in clinical genomic tests. This discovery was acknowledged as the breakthrough of the year by *Science Magazine* in 2007 as well as Thompson Reuter's selection of its 2012 Citation Laureates.

From 2000-2012, I led several research, educational and clinical genetics programs, in my capacity as Director of the Dana-Farber/Harvard Cancer Center Cytogenetics Core, Director of the Molecular Genetics Research Unit at Brigham and Women's Hospital, and appointments at Harvard Medical School and the Broad Institute. In 2013, I became the Scientific Director of The Jackson Laboratory for Genomic Medicine, where I have been able to develop and lead a team of world-renowned investigators with the collective goal of translating scientific discoveries into novel diagnostic tests and individualized treatments. At the Jackson Laboratory for Genomic Medicine, my research program encompasses three areas: 1) SVs in the human genome, 2) clinical genomic diagnostics and 3) cancer genetics and biomarkers. The current proposal will benefit from my group's expertise in leading large-scale SV projects and developing computational methods for SV discovery, as demonstrated by our participation in the 1000 Genomes Project.

B. Positions and Honors

Positions and Employment

1993	Invited Research Trainee, Johns Hopkins School of Medicine, Baltimore, MD
2000-2006	Assistant Director, Harvard Cancer Center Cytogenetics Core, Boston, MA
2001-2003	Instructor in Pathology, Harvard Medical School, Boston, MA
2001-2008	Associate Clinical Cytogeneticist, Brigham and Women's Hospital, Boston, MA
2003-2008	Assistant Professor of Pathology, Harvard Medical School, Boston, MA
2006-2013	Director, Harvard Cancer Center Cytogenetics Core, Boston, MA
2006-2013	Associate Member, Broad Institute of Harvard and MIT, Boston, MA

2008-2011	Associate Professor of Pathology, Harvard Medical School, Boston, MA
2008-2013	Clinical Cytogeneticist, Brigham and Women's Hospital, Boston, MA
2009-2013	Director, Molecular Genetics Research Unit, Brigham and Women's Hospital, Boston, MA
2013-Present	Scientific Director and Professor, The Jackson Laboratory for Genomic Medicine, Farmington, CT

Awards and Honors

1994	75th Anniversary Faculty of Medicine Scholarship, University of Alberta, Canada
1994-1996	PhD Studentship, Alberta Heritage Foundation for Medical Research, Canada
1996	MRC Postdoctoral Fellowship, Medical Research Council of Canada
1996-1998	NSERC Postdoctoral Fellowship, Natural Sciences and Engineering Research Council of Canada
2002	Stanley L. Robbins Research Award, Department of Pathology, Brigham and Women's Hospital, Boston, MA
2008	Ho-Am Prize in Medicine, Ho Am Foundation, Seoul, South Korea
2008	C. Thomas Caskey Award, University of South Carolina, SC
2010	George W. Brumley, Jr., M.D. Memorial Award, Duke University, Durham, NC
2012	Chen Global Investigator Award, Human Genome Organization (HUGO)
2012	Fellow, American Association for the Advancement of Science (AAAS)
2012	Vandenberghe Visiting Chair, Center for Human Genetics, Catholic University of Leuven, Belgium
2013-2015	Distinguished Visiting Professor, Seoul National University School of Medicine, Korea
2014	Citation Laureate, Thompson Reuter, USA
2015-Present	Distinguished EWHA University Visiting Professor, EWHA Womans University, Korea

Other Professional Activities

2002-Present	Diplomate, American Board of Medical Genetics
2004	Ad hoc reviewer, Genome Canada Grant Competition
2005	Ad hoc reviewer, NSF Peer Review Committee
2006-2008	Member, NIH/ NHGRI Structural Variation Steering Committee
2007-2010	Member, American Society of Human Genetics Program Committee
2008-2012	Faculty, Program in Quantitative Genomics at Harvard School of Public Health
2008-2012	Scientific Advisory Board, Yale Center for Excellence in Genome Sciences
2008-Present	Co-chair, 1000 genomes project (www.1000genomes.org) - SV Analysis Group
2009	Chair, NHGRI Cytogenetics Core Advisory Board
2009-2011	Associate Editor, *American Journal of Human Genetics*
2009-2012	Director, Cancer Cytogenomics Microarray Consortium Board of Directors
2015-Present	Associate Editor, *Genomic Medicine*
2015-Present	Editorial Board, *Human Genomics*

C. Contribution to Science

1. **Structural and Copy Number Variation in the Human Genome and the genomes of model organisms.** Our research in human structural genomic variation aims to accurately identify and characterize deletions, duplications and balanced chromosomal rearrangements in the genomes of humans and model organisms and understand the biological implications of these variants. We originally described the widespread presence of structural genomic variants in the human genome and have extended our studies to specific world populations. Over the past five years, we have led the structural variation group of the 1000 Genomes Project (an international collaboration aimed at identifying and cataloging all genetic variants occurring at a frequency of at least 1% in 26 world populations, http://www.1000genomes.org/) and have developed methods for identifying human structural genomic variants at higher resolution from whole genome sequence analyses.
    a. Iafrate AJ, Feuk L, Rivera MN, Listewnik ML, Donahoe PK, Qi Y, Scherer SW, **Lee C**. Detection of large-scale variation in the human genome. *Nature Genetics*. 2004; 36(9):949-51. PMID: 15286789
    b. Park HS, Kim JI, Ju YS, Gokcumen O, Mills, RE, … , Darvishi K, Yang SJ, Yang KS, Kim HT, Hurles ME, Scherer SW, Carter NP, Tyler-Smith C, Seo JS, **Lee C**. Absolute quantification of common Asian copy number variants (CNVs) using an integrated approach of high resolution array CGH and massively parallel DNA sequencing. *Nat Genet*. 2010; 42: 400-5. PMCID: PMC3329635

c. Mills RE, Walter K, Stewart C, Handsaker RE, Chen K, Alkan C, … , Eichler EE*, Gerstein MB*, Hurles ME*, **Lee C***, McCarroll SA*, Korbel JO*. Mapping copy number variation by population-scale genome sequencing. *Nature*. 2011; 470(7332):59-65. PMCID: PMC3077050 **co-senior author*

d. Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, … , Mills RE*, Gerstein M*, Bashir A*, Stegle O*, Devine SE*, **Lee C***, Eichler EE*, Korbel JO*. An integrated map of structural variation in 2,504 human genomes. *Nature*. 2015 Oct 1;526(7571):75-81.**co-senior author* PMCID: PMC4617611 **co-senior author*

2. **Structural and Copy Number Variation in the genomes of model organisms.** Understanding the biological implications of specific structural genomic variants requires the accurate identification and genotyping of these variants in cell lines and model organisms. To optimize genomic studies in model organisms and more accurately understand the contributions of specific variants to human diseases, we have developed the first structural genomic maps for several non-human genomes, including the zebrafish, the Chimpanzee, and the Rhesus Macaque. We have further shown that most of these structural variants lie outside of genes but yet can dramatically influence cellular transcriptional profiles.

a. Perry GH, Tchinda J, McGrath SD, Zhang J, Picker SR, Caceres AM, Iafrate AJ, Tyler-Smith C, Scherer SW, Eichler EE, Stone AC, **Lee C**. Hotspots for copy number variation in chimpanzees and humans. *Proc Natl Acad Sci USA*. 2006; 103: 8006-11. PMCID: PMC1472420

b. Lee AS, Gutierrez-Arcelus M, Perry GH, Palacios R, Vallender EJ, Johnson WE, Miller GM, Korbel JO, **Lee C.** Analysis of copy number variation in the rhesus macaque genome identified candidate loci for evolutionary and human disease studies. *Hum Mol Genet*. 2008; 17: 1127-36.

c. Brown, KH, Dobrinski KP, Lee AS, Gokcumen O, Mills RE, Shi X, Chong WW, Chen JY, Yoo P, David S, Peterson SM, Raj T, Choy KW, Stranger B, Williamson RE, Zon LI, Freeman JL, **Lee C**. Extensive genetic diversity and sub-structuring among zebrafish strains revealed through copy number variant analysis. *Proc Natl Acad Sci USA*. 2012; 109: 529-534. PMCID: PMC3258620

d. Iskow RC, Gokcumen O, Abyzov A, Malukiewicz J, Zhu Q, Sukumar AT, Pai AA, Mills RE, Habegger L, Cusanovich DA, Rubel MA, Perry GH, Gerstein M, Stone AC, Gilad Y, **Lee C**. Regulatory element copy number differences shape primate expression profiles. *Proc Natl Acad Sci USA*. 2012; 109: 12656-61.

3. **Clinical Genomic Diagnostics.** Combining our laboratory's interest in advanced molecular technologies, the search for new biomarkers and clinical expertise as a board-certified clinical cytogeneticist, our laboratory is well poised to develop new clinical genomic diagnostic tests. Some of our newly developed assays include (1) the *TMPRSS2* and *ETS* transcription factor in aggressive prostate cancer, and (2) *MET* amplification in gefitinib-resistance non small cell lung carcinoma.

a. **Lee C**, Gisselsson D, Jin C, Nordgren A, Ferguson DO, Blennow E, Fletcher JA, Morton CC. Limitations of chromosome classification by multicolor karyotyping. *Am J Hum Genet*. 2001; 68: 1043-7.

b. **Lee C**, Iafrate AJ, Brothman AR. Copy number variations and clinical cytogenetic diagnosis of constitutional disorders. *Nat Genet*. 2007; 39: S48-S54.

c. Tomlins SA, Rhodes DR, Perner S, Dhanasekaran SM Mehra R, Sun X-W, Varambally S, Cao X, Tchinda J, Kuefer R, **Lee C**, Montie JE, Shah RB, Pienta KJ, Rubin MA, Chinnaiyan AM. Recurrent fusion of *TMPRSS2* and ETS transcription factor genes in prostate cancer. *Science.* 2005; 310: 644-8.

d. Engelman JA, Zejnullahu K, Mitsudomi T, Song Y, Hyland C, Park JO, Lindeman N, Gale C-M, Zhao X, Christensen J, Kosaka T, Holmes AJ, Rogers AM, Cappuzzo F, Mok T, **Lee C**, Johnson BE, Cantley LC, Janne PA. *MET* amplification leads to gefitinib resistance via ERBB3 in *EGFR* mutant lung cancer. *Science*. 2007; 316: 1039-43.

4. **Cancer Genomics.** Our research has been focused on the study of human cancer genome sequences and structures to provide insights into cancer biology, diagnosis and therapy. Thus, we have undertaken a number of studies directed towards molecular cytogenetics of cancer. More recently, we have initiated a large-scale project using patient-derived xenograft (PDX) tumor models established in NSG (Nod-SCID-IL2RKO) immunodeficient mice bearing the human immune system. This serves as a personalized animal model of a patient's tumor, which can be used in both co-clinical trials for drug efficacy as well as the development of databases for genomic profiles and clinical outcomes.

a. Garraway LA, Widlund HR, Rubin MA, Berger AJ, Sridhar R, Chen F, Beroukhim R, Getz G, Milner DA, Granter SR, Du J, **Lee C**, Wagner SN, Li C, Golub TR, Rimm DL, Meyerson M, Fisher DE, Sellers WR. Integrative genomic analysis identify *MITF* as a lineage survival oncogene amplified in malignant melanoma. *Nature*. 2005; 436:117-22.

b. Demichelis F, Setlur SR, Banerjee S, Chakravarty D, Chen JY, Chen CX, Huang J, Beltran H, Oldridge DA, Kitabayashi N, Stenzel B, Schaefer G, Horinger W, Bektic J, Chinnaiyan AM, Goldenberg S, Siddiqui J, Regan M, Kearney M, Soong TD, Rickman DS, Elemento O, Wei JT, Scherr DS, Sanda MA, Bartsch G, Klocker H*, Rubin MA*, **Lee C*.** Identification of functionally active, low frequency copy number variants at 15q21.3 and 12q21.31 associated with prostate cancer risk. *PNAS*. 2012; 109(17):6686-91. PMCID: PMC3340033 **co-senior author*

c. Chen Z, Cheng K, Walton Z, Wang Y, Ebi H, … , **Lee C**, … , Engelman JA, Wong KK. A murine lung cancer co-clinical trial identifies genetic modifiers of therapeutic response. *Nature*. 2012; 483(7391):613-7. PMCID: PMC3385933

d. Yang L, Luquette LJ, Gehlenborg N, Xi R, Haseley PS, Hsieh C-H, Zhang C, Ren X, Protopopov A, Chin L, Kucherlapati R, **Lee C***, Park PJ*. Diverse mechanisms of somatic structural variations in human cancer genomes. *Cell*. 2013; 153(4):919-29. PMCID: PMC3704973 **co-senior author*

A complete list of published work can be found in My Bibliography:
http://www.ncbi.nlm.nih.gov/myncbi/browse/collection/41142830/?sort=date&direction=ascending

D. Research Support

Ongoing Research Support

U41 HG007497 Lee (PI) 08/01/13-06/30/17
NIH/NHGRI
An integrative analysis of structural variation for the 1000 Genomes Project
The major goal of this project is to develop and assess new methods for accurately identifying structural genomic variants in next generation DNA sequencing datasets.
Role: Program Director

P01 HD068250 Donahoe (PI) 07/01/11-06/30/16
NIH/NICHD
Gene Mutation and rescue in Congenital Diaphragmatic Hernia
The major goal of this project is to explore and compare the impact of using WGS in clinical conditions.
Role: Principal Investigator of Project 2

U41 HG006834 Ledbetter, Martin, Mitchell, Nussbaum, Rehm (PI) 10/01/13-07/31/16
NIH/NHGRI
A Unified Clinical Genomics Database
The goal of this project is to collect and organize genome-wide structural and sequence-level variation data from many sources into a free and publically accessible environment and enable expert curation of that data for use in improving healthcare and biomedical research.
Role: Consortium Principal Investigator

Completed Research Support

R01 A1089246 Simon (PI) 06/01/10-05/31/15
NIH/NIAID
Genomic Determinants of Intrinsic Antiviral host Defenses
The major goal of this project was to identify copy number variants associated with increased susceptibility to infectious disease.
Role: Co-Investigator

U01 HG005725 Lee (PD) 08/18/10 – 05/31/13
NIH/NHGRI
Analysis of Patterns of Structural Variation in the 1000 Genomes Data Set
The major goal of this project is to characterize structural genomic variation in the 1000 Genomes next generation whole-genome DNA sequencing datasets and bioinformatically predict function.
Role: Program Director

BIOGRAPHICAL SKETCH

Provide the following information for the Senior/key personnel and other significant contributors.
Follow this format for each person. DO NOT EXCEED FIVE PAGES.

NAME: Wendl, Michael

eRA COMMONS USER NAME (credential, e.g., agency login): MCWENDL

POSITION TITLE: Assistant Professor

EDUCATION/TRAINING *(Begin with baccalaureate or other initial professional education, such as nursing, include postdoctoral training and residency training if applicable.)*

INSTITUTION AND LOCATION	DEGREE (if applicable)	Completion Date MM/YYYY	FIELD OF STUDY
Washington University, St. Louis, MO	BS	05/1989	
Washington University, St. Louis, MO	MS	05/1990	
Washington University, St. Louis, MO	SCD	05/1994	

A. Personal Statement

This proposal outlines an ambitious project whose end-goal is the discovery of a large class of structural variants (translocations, complex SVs, etc) and establishment of their associations with various complex human diseases. Given the enormous data to which investigators will have access and the coupling with the advanced analytical/statistical methods, the work outlined in this proposal is very likely to ultimately yield important, clinically-relevant advancements in human disease. A substantial portion of this project will depend upon sophisticated mathematical, statistical, and computer science methodologies applied to analysis of genomic data, including SV detection, association/hypothesis testing, pathway/network analysis, and information systems. My primary role will be to contribute to and to help manage this aspect of the work. My abilities on a technical level to ensure its success in this regard are affirmed by mathematical and statistical contributions over a broad spectrum: in (1) bioinformatics, e.g. Phred, BreakDancer, MuSiC, PathScan, etc. (2) applied mathematics and statistics, e.g. coverage theory, and (3) pure mathematics, e.g. combinatorial probability, random variable collisions, and differential equations. Coupled with the expertise of the other investigators, we believe we can fully deliver on what we are proposing, and furthermore, that this work will indeed be an important foundation for the future of medicine and the diagnosis and treatment of complex human disease.

1. Xie M, Lu C, Wang J, McLellan MD, Johnson KJ, Wendl MC, McMichael JF, Schmidt HK, Yellapantula V, Miller CA, Ozenberger BA, Welch JS, Link DC, Walter MJ, Mardis ER, Dipersio JF, Chen F, Wilson RK, Ley TJ, Ding L. Age-related mutations associated with clonal hematopoietic expansion and malignancies. Nat Med. 2014 Dec;20(12):1472-8. PubMed PMID: 25326804; PubMed Central PMCID: PMC4313872.
2. Ding L, Raphael BJ, Chen F, Wendl MC. Advances for studying clonal evolution in cancer. Cancer Lett. 2013 Nov 1;340(2):212-9. PubMed PMID: 23353056; PubMed Central PMCID: PMC3783624.
3. Dees ND, Zhang Q, Kandoth C, Wendl MC, Schierding W, Koboldt DC, Mooney TB, Callaway MB, Dooling D, Mardis ER, Wilson RK, Ding L. MuSiC: identifying mutational significance in cancer genomes. Genome Res. 2012 Aug;22(8):1589-98. PubMed PMID: 22759861; PubMed Central PMCID: PMC3409272.
4. Wendl MC, Korf I, Chinwalla AT, Hillier LW. Automated processing of raw DNA sequence data. IEEE Eng Med Biol Mag. 2001 Jul-Aug;20(4):41-8. PubMed PMID: 11494768.

B. Positions and Honors

Positions and Employment

1994 - 2001	Research Associate, Washington University McDonnell Genome Institute, St. Louis, MO
2001 - 2005	Research Instructor, Washington University McDonnell Genome Institute, St. Louis, MO
2005 - 2014	Research Assistant Professor, Washington University McDonnell Genome Institute, St. Louis,

	MO
2010 -	Assistant Professor, Washington University Mathematics Dept, St. Louis, MO
2014 -	Assistant Professor, Washington University McDonnell Genome Institute, St. Louis, MO

Other Experience and Professional Memberships

Honors

1992	Graduate Student Fellow Award, American Institute of Aeronautics and Astronautics

C. Contribution to Science

1. I have contributed a number of bioinformatic and statistical tools/techniques that have become widely-used and well-established in the genomic sciences. My initial work was with Phil Green on PHRED, the analog-to-digital converter (Fast Fourier Transform and numerical optimization) that became the standard base-caller for Sanger sequencing. I have since contributed to processing tools (e.g. MuSiC), detectors (e.g. PolyScan, BreakDancer), and analysis tools (e.g. PathScan). Many of the other specialized statistical tests I have developed, for example, for allelic imbalance analysis, have appeared as part of a larger biological publication, rather than a standalone paper.

    a. Ewing B, Hillier L, Wendl MC, Green P. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res. 1998 Mar;8(3):175-85. PubMed PMID: 9521921.

    b. Chen K, McLellan MD, Ding L, Wendl MC, Kasai Y, Wilson RK, Mardis ER. PolyScan: an automatic indel and SNP detection approach to the analysis of human resequencing data. Genome Res. 2007 May;17(5):659-66. PubMed PMID: 17416743; PubMed Central PMCID: PMC1855178.

    c. Wendl MC, Wallis JW, Lin L, Kandoth C, Mardis ER, Wilson RK, Ding L. PathScan: a tool for discerning mutational significance in groups of putative cancer genes. Bioinformatics. 2011 Jun 15;27(12):1595-602. PubMed PMID: 21498403; PubMed Central PMCID: PMC3106187.

    d. Xie M, Lu C, Wang J, McLellan MD, Johnson KJ, Wendl MC, McMichael JF, Schmidt HK, Yellapantula V, Miller CA, Ozenberger BA, Welch JS, Link DC, Walter MJ, Mardis ER, Dipersio JF, Chen F, Wilson RK, Ley TJ, Ding L. Age-related mutations associated with clonal hematopoietic expansion and malignancies. Nat Med. 2014 Dec;20(12):1472-8. PubMed PMID: 25326804; PubMed Central PMCID: PMC4313872.

2. Much of my applied mathematics work has focused on quantifying various molecular biology processes, e.g. mapping and sequencing DNA, requirements and power analysis for detecting various kinds of features, etc. This work revealed some non-intuitive surprises in designing post-HGP projects that subsequent experience has confirmed, for example that sequencing genomes of an individual for the purpose of detecting variation would need to be substantially higher than the 10X HGP standard (e.g. >30X, which is now standard), but that sequencing large populations to find rare variants is optimized for much lower depths, around 4x (a design subsequently used by the 1000 Genomes Project). Additional work has described optimization for many other scenarios, e.g. for detecting indels and for metagenomic sequencing of microbe communities.

    a. Wendl MC. A general coverage theory for shotgun DNA sequencing. J Comput Biol. 2006 Jul-Aug;13(6):1177-96. PubMed PMID: 16901236.

    b. Wendl MC, Wilson RK. Aspects of coverage in medical DNA sequencing. BMC Bioinformatics. 2008 May 16;9:239. PubMed PMID: 18485222; PubMed Central PMCID: PMC2430974.

    c. Wendl MC, Wilson RK. The theory of discovering rare variants via DNA sequencing. BMC Genomics. 2009 Oct 20;10:485. PubMed PMID: 19843339; PubMed Central PMCID: PMC2778663.

    d. Wendl MC, Kota K, Weinstock GM, Mitreva M. Coverage theories for metagenomic DNA sequencing based on a generalization of Stevens' theorem. J Math Biol. 2013 Nov;67(5):1141-61. PubMed PMID: 22965653; PubMed Central PMCID: PMC3795925.

3. My work in pure mathematics has been in 2 specific areas: partial differential equations (PDEs) and combinatorial probability. In the former, I have addressed some long-standing PDEs governed primarily by

the linear 2nd-order diffusion operator, including the Couette problem first discussed by GI Taylor in 1923. These analyses are closely applicable to fluid physics, heat transfer, etc. given the connection of the operator to physical diffusion processes. For the latter, I have focused on combinatorial problems that depend upon set-partitioning, including the collisions of random variables. This class of problems is widely applicable to peer-to-peer searching, physical collision and coverage processes, etc.

a. Wendl MC. General solution for the Couette flow profile. Phys Rev E Stat Phys Plasmas Fluids Relat Interdiscip Topics. 1999 Nov;60(5 Pt B):6192-4. PubMed PMID: 11970531.
b. Wendl MC. Mathematical analysis of coaxial disk cellular shear loading devices. The Review of scientific instruments. 2001; 72(11):4212-4217.
c. Wendl MC. Collision probability between sets of random variables. Statistics & probability letters. 2003; 64(3):249-254.

D. Additional Information: Research Support and/or Scholastic Performance

Ongoing Research Support

1R01CA178383-01A1, NCI
Ding, Li (PI)
09/08/14-08/31/17
Virus discovery and characterization in large-scale cancer sequencing data
Role: KP

CA172652-02, NCI
Chen, Ken (PI)
04/01/13-03/31/17
Delineating Heterogeneous Structural Complexity in Cancer Genomes
Role: KP

Completed Research Support

5R01HG005690-04, NHGRI
Raphael, Ben (PI)
01/01/11-12/31/15
Computational approaches for structural variation studies in genomes
Role: KP

BIOGRAPHICAL SKETCH

NAME: Ding, Li

eRA COMMONS USER NAME (agency login): DINGLI

POSITION TITLE: Associate Professor

EDUCATION/TRAINING

INSTITUTION AND LOCATION	DEGREE (if applicable)	Completion Date MM/YYYY	FIELD OF STUDY
Fudan University, Shanghai	BS	07/1991	Biology
University of Utah School of Medicine, Salt Lake City, UT	PHD	08/1998	Biochemistry
Stanford University, Palo Alto, CA	Postdoctoral Fellow	05/2000	Biochemistry

A. Personal Statement

The long term goal of my research is to combine my strengths in biology and bioinformatics to answer fundamental questions in biological science and human diseases. My current research focuses on understanding the genetic basis of human diseases, through innovative computational and experimental approaches. My lab has developed important bioinformatics tools, including VarScan, BreakDancer, TIGRA-SV, Pindel-C, MSIsensor, HotSpot3D, and MuSiC, which are widely used for human genetics and cancer proteogenomic studies. We have aggressively pursued technological innovation and have been actively developing enabling tools in proteomics using methodologies that increasingly combine the power of genomics and proteomics to discover novel disease markers. I have been at the forefront of human genetics and cancer genomics research and with a long record of significant discovery: conducted and led the very first large-scale lung cancer genomics study (2006-2008), pioneered drug-resistant subclone/mutation discovery during tumor metastasis/relapse (2009-2012), used the Pan-cancer approach to gain statistical power for rapid, massive cancer gene discovery (starting from 2012), developed integrative approaches to study combined effects of germline and somatic mutations simultaneously (Starting from 2012), and utilized global proteomic and phosphoproteomic data for druggable protein discovery (Starting from 2012). Innovative use of blood sample sequencing data uncovered pre-existing leukemic mutations in important leukemia-associated genes in individuals without hematological diseases (starting from 2013). In summary, I have developed an outstanding program to advance biology and treatment, as demonstrated by a track record of innovative approaches and ground-breaking discoveries. Relevant publications selected from over 120 total publications are listed below.

1. Kandoth C, McLellan MD, Vandin F, Ye K, Niu B, Lu C, Xie M, Zhang Q, McMichael JF, Wyczalkowski MA, Leiserson MD, Miller CA, Welch JS, Walter MJ, Wendl MC, Ley TJ, Wilson RK, Raphael BJ, **Ding L.** Mutational landscape and significance across 12 major cancer types. **Nature.** 2013 Oct 17;502(7471):333-9. PMCID: PMC3927368.
2. Xie M, Lu C, Wang J, McLellan MD, Johnson KJ, Wendl MC, McMichael JF, Schmidt HK, Yellapantula V, Miller CA, Ozenberger BA, Welch JS, Link DC, Walter MJ, Mardis ER, Dipersio JF, Chen F, Wilson RK, Ley TJ, **Ding L.** Age-related mutations associated with clonal hematopoietic expansion and malignancies. **Nat Med.** 2014 Dec;20(12):1472-8. PMCID: PMC4313872.
3. Lu C, Xie M, Wendl MC, Wang J, McLellan MD, Leiserson MD, Huang KL, Wyczalkowski MA, Jayasinghe R, Banerjee T, Ning J, Tripathi P, Zhang Q, Niu B, Ye K, Schmidt HK, Fulton RS, McMichael JF, Batra P, Kandoth C, Bharadwaj M, Koboldt DC, Miller CA, Kanchi KL, Eldred JM, Larson DE, Welch JS, You M, Ozenberger BA, Govindan R, Walter MJ, Ellis MJ, Mardis ER, Graubert TA, DiPersio JF, Ley TJ, Wilson RK, Goodfellow PJ, Raphael BJ, Chen F, Johnson KJ, Parvin JD, **Ding L**. Patterns and functional implications of rare germline variants across 12 cancer types. **Nat Commun.** 2015 Dec 22;6:10086 PMCID: PMC4703835
4. Niu B, Scott AD, Sengupta S, Bailey MH, Batra P, Ning J, Wyczalkowski MA, Liang WW, Zhang Q, McLellan MD, Sun SQ, Tripathi P, Lou C, Ye K, Mashl RJ, Wallis J, Wendl MC, Chen F, **Ding L.** Protein-structure-guided discovery of functional mutations across 19 cancer types. **Nat Genet**. 2016 Jun. PMID: 27294619

B. Positions and Honors

Positions and Employment

2005 - 2012	Group leader, Medical Genomics Group, The Genome Institute, Washington University in St. Louis, MO
2008 - present	Assistant Director, McDonnell Genome Institute, Washington University in St. Louis, MO
2012 - 2015	Assistant Professor, Department of Medicine, Department of Genetics, St. Louis, MO
2015 - present	Associate Professor, Department of Medicine, Washington University School of Medicine, St. Louis, MO

Other Experience and Professional Memberships

Member, American Association for the Advancement of Science (AAAS)
Member, American Association for Cancer Research (AACR)
Member, American Society of Human Genetics (ASHG)
Research Member, Siteman Cancer Center
Reviewer, Cancer Research UK
Review Panel, UK-ICGC program (prostate and oesophagus projects)
Source Evaluation Group, NCI Cancer Genomics Data Commons
Review Panel/Reviewer, NIH Cancer Genetic Study Section
Review Panel/Reviewer, NIH microbiome sciences and the associated informatics SEP
Co-chair, ICGC Pan Cancer Mutation Calling Group
Co-chair, TCGA Pan Cancer Atlas Oncogenic Process Group
Co-chair, TCGA Sarcoma Analysis Working Group
Steering Committee Member, Clinical Proteomic Tumor Analysis Consortium (CPTAC)
Member, CPTAC Data Analysis Working Group
Co-chair, Pan Cancer Atlas Germline Group
Co-chair, Pan Cancer Atlas Driver Group
Steering Committee Member, NCI Genomic Data Commons (GDC)

Honors

2008	Tomorrow's PI, Genome Technology
2010	Chair for Functional and Cancer Genomics Session Session, The Biology of Genomes
2011	Chair for Population and Personal Genomics Session, Genome Informatics
2013	The Hottest Scientific Researchers of 2012 , Thomson Reuters
2014	Chair for Genomic Alterations of Tumors Session, American Society of Human Genetics
2014	The World’s Most Influential Scientific Minds 2014 , Thomson Reuters
2015	Chair for TCGA Fourth Annual Scientific Symposium, NIH

C. Contribution to Science

1. **Develop and Refine Computational Tools for Genome Research and Clinical Applications.** We have developed an enterprise system for analyzing genomics and proteomics data, with particular emphasis on detection and interpretation. It handles tracking, support, etc. and includes powerful detection and interpretation tools like VarScan, BreakDancer, SomaticSniper, HotSpot3D, Pindel-C, MSIsensor, MuSiC and others, which are widely used for both individual projects and large-scale collaborations, e.g. TCGA, CPTAC, and ICGC. We have further developed these tools under the banner of the “Turnkey Variant Analysis Project” (TVAP) (http://tvap.genome.wustl.edu/). All TVAP programs are publically available through GitHub or SourceForge. We recently summarized the broad landscape of such tools in several reviews.
    a. Koboldt DC, Chen K, Wylie T, Larson DE, McLellan MD, Mardis ER, Weinstock GM, Wilson RK, **Ding L.** VarScan: variant detection in massively parallel sequencing of individual and pooled samples. **Bioinformatics.** 2009 Sep 1;25(17):2283-5. PMCID: PMC2734323.

b. Dees ND, Zhang Q, Kandoth C, Wendl MC, Schierding W, Koboldt DC, Mooney TB, Callaway MB, Dooling D, Mardis ER, Wilson RK, **Ding L.** MuSiC: identifying mutational significance in cancer genomes. **Genome Res**. 2012 Aug;22(8):1589-98. PMCID: PMC3409272.

c. **Ding L,** Wendl MC, McMichael JF, Raphael BJ. Expanding the computational toolbox for mining cancer genomes. **Nat Rev Genet**. 2014 Aug;15(8):556-70. PMCID: PMC4168012.

d. Ye K, Wang J, Jayasinghe R, Lameijer EW, McMichael JF, Ning J, McLellan MD, Xie M, Cao S, Yellapantula V, Huang KL, Scott A, Foltz S, Niu B, Johnson KJ, Moed M, Slagboom PE, Chen F, Wendl MC, **Ding L**. Systematic discovery of complex insertions and deletions in human cancers. **Nat Med** 2015 Dec 14; PMID: 26657142

2. **Employ the Latest Genomics/Proteomics Technologies to Understand DNA/RNA/protein Interactions and Conduct Protein Biomarker Discovery.** I have been working with Drs. David Fenyö and Ben Raphael's labs to develop computational approaches for integrating genomics and proteomics data for identifying altered networks and pathways in ovarian, colorectal, and breast cancer cohorts. There is of great potential for learning the significant, cancer-specific molecular alterations that have biological and clinical implications. In addition, we have been investigating phosphoproteomics data to evaluate phosphorylation status at the locus and gene levels in tumor samples. One ongoing effort is to discern activated pathways in human tumors using proteomics data and discover druggable targets using protein/phosphoprotein outlier analysis and test treatment responses using patient-derived xenograft models (manuscript under review).

a. Zhang B, Wang J, Wang X, Zhu J, Liu Q, Shi Z, Chambers MC, Zimmerman LJ, Shaddox KF, Kim S, Davies SR, Wang S, Wang P, Kinsinger CR, Rivers RC, Rodriguez H, Townsend RR, Ellis MJ, Carr SA, Tabb DL, Coffey RJ, Slebos RJ, Liebler DC. Proteogenomic characterization of human colon and rectal cancer. **Nature.** 2014 Sep 18;513(7518):382-7. PMCID: PMC4249766. (**Ding L** is a CPTAC consortium member)

b. Leiserson MD, Vandin F, Wu HT, Dobson JR, Eldridge JV, Thomas JL, Papoutsaki A, Kim Y, Niu B, McLellan M, Lawrence MS, Gonzalez-Perez A, Tamborero D, Cheng Y, Ryslik GA, Lopez-Bigas N, Getz G, **Ding L**, Raphael BJ. Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. **Nat Genet**. 2015 Feb;47(2):106-14. PMCID:PMC4444046

c. Ruggles KV, Tang Z, Wang X, Grover H, Askenazi M, Teubl J, Cao S, McLellan MD, Clauser KR, Tabb DL, Mertins P, Slebos R, Erdmann-Gilmore P, Li S, Gunawardena HP, Xie L, Liu T, Zhou JY, Sun S, Hoadley KA, Perou CM, Chen X, Davies SR, Maher CA, Kinsinger CR, Rodland KD, Zhang H, Zhang Z, **Ding L**, Townsend RR, Rodriguez H, Chan D, Smith RD, Liebler DC, Carr SA, Payne S, Ellis MJ, Fenyo D. An analysis of the sensitivity of proteogenomic mapping of somatic mutations and novel splicing events in cancer. **Mol Cell Proteomics** 2015 Dec 2; PMCID:PMC4813688

d. Mertins P, Mani DR, Ruggles KV, Gillette MA, Clauser KR, Wang P, Wang X, Qiao JW, Cao S, Petralia F, Kawaler E, Mundt F, Krug K, Tu Z, Lei JT, Gatza ML, Wilkerson M, Perou CM, Yellapantula V, Huang KL, Lin C, McLellan MD, Yan P, Davies SR, Townsend RR, Skates SJ, Wang J, Zhang B, Kinsinger CR, Mesri M, Rodriguez H, **Ding L**, Paulovich AG, Fenyo D, Ellis MJ, Carr SA & the NCI CPTAC. Proteogenomics connects somatic mutations to signaling in breast cancer. **Nature 2016 June;** 534,55–62

3. **Reveal Molecular Processes and Components Underpinning Tumor Initiation, Clonal Evolution, and Metastasis/Relapse.** 1) Somatic variants associated with cancers initiation and progression: At the launch of the Tumor Sequencing Project (TSP), circa 2005, most institutes were using ABI instruments to investigate a few candidate genes in a handful of cancer samples. TSP pushed the envelope by systematically characterizing 188 lung adenocarcinoma genomes. With colleagues from 19 different institutes, I led the sequencing and analysis of 623 selected candidate genes between 2006 and 2008. This is considered a pilot for TCGA. 2) Relapse and metastasis: Relapse and metastasis are signatures of malignancy and are the most common causes of cancer-related death, yet the genetic changes underlying these phenomena are poorly understood. I formed clinical collaborations for some of their earliest genomic studies, including the first observations of subclonal mutations in heterogeneous breast tumors and demonstration of clonal evolution and selection of resistant subclones by chemotherapy treatments in AML. 3) Driver genes and mutations contributing to individual and multiple cancer types: My lab developed a suite of tools collectively called MuSiC for cancer driver mutation/gene discovery and for revealing clinical

associations using large-scale cancer data sets. Utilizing MuSiC, we discovered 127 significantly mutated cancer genes in over 3,000 tumors from 12 major cancer types. Due to its novelty and broad implications, this study was widely covered in both scientific literature (Skipper, Nature Review Genetics, 2013, Ashworth and Hudson, Nature, 2013) and mainstream media (Wall Street Journal, Oct. 16th, 2013; Bloomberg, Oct. 16th, 2013; The Economist, Jan. 14th, 2014). It is representative of my commitment to collaborate with investigators world-wide to answer critical questions in biomedical research.

a. **Ding L,** Getz G, Wheeler DA, Mardis ER, McLellan MD, Cibulskis K, Sougnez C, Greulich H, Muzny DM, Morgan MB, *et al.* Somatic mutations affect key pathways in lung adenocarcinoma. **Nature.** 2008 Oct 23;455(7216):1069-75. PMCID: PMC2694412.

b. **Ding L**, Ellis MJ, Li S, Larson DE, Chen K, Wallis JW, Harris CC, McLellan MD, Fulton RS, Fulton LL, Abbott RM, Hoog J, Dooling DJ, Koboldt DC, *et al.* Genome remodelling in a basal-like breast cancer metastasis and xenograft. **Nature.** 2010 Apr 15;464(7291):999-1005. PMCID: PMC2872544.

c. **Ding L**, Ley TJ, Larson DE, Miller CA, Koboldt DC, Welch JS, Ritchey JK, Young MA, Lamprecht T, McLellan MD, McMichael JF, Wallis JW, Lu C, Shen D, Harris CC, Dooling DJ, Fulton RS, Fulton LL, Chen K, Schmidt H, Kalicki-Veizer J, Magrini VJ, Cook L, McGrath SD, Vickery TL, Wendl MC, Heath S, Watson MA, Link DC, Tomasson MH, Shannon WD, Payton JE, Kulkarni S, Westervelt P, Walter MJ, Graubert TA, Mardis ER, Wilson RK, DiPersio JF. Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. **Nature.** 2012 Jan 11;481(7382):506-10. PMCID: PMC3267864.

d. Kandoth C, McLellan MD, Vandin F, Ye K, Niu B, Lu C, Xie M, Zhang Q, McMichael JF, Wyczalkowski MA, Leiserson MD, Miller CA, Welch JS, Walter MJ, Wendl MC, Ley TJ, Wilson RK, Raphael BJ, **Ding L.** Mutational landscape and significance across 12 major cancer types. **Nature.** 2013 Oct 17;502(7471):333-9. PMCID: PMC3927368.

4. **Establish Interactions between Germline and Somatic genomes to Reveal Joint Contributions to Cancer Predisposition, Initiation, and Progression.** Major advancements have been made in cataloging somatic variations in cancer genomes, but companion analysis for germline changes remains challenging. Recently, we analyzed germline and tumor sequence data from 4,034 samples representing 12 cancer types. Our study, for the first time, revealed a large number of rare germline mutations enriched in the tumors across all 12 cancer types. Further, we found that thirteen genes had a significantly elevated burden of mutations across all 12 cancer types, including BRCA1, BRCA2, ATM, BRIP1, and PALB2 and an additional 21 genes had suggestive evidence of an increased burden, comprising 8.3% of total cancer cases studied. We also investigated pre-existing mutations in hematopoietic stem cells to understand their relevance to cancer mutations and development. We further demonstrated that age-related hematopoietic clonal mosaicism. The implication is elevated incidence of hematological malignancy from the expansion of such mutant clones could occur as life expectancy increases.

a. Kanchi KL, Johnson KJ, Lu C, McLellan MD, Leiserson MD, Wendl MC, Zhang Q, Koboldt DC, Xie M, Kandoth C, McMichael JF, Wyczalkowski MA, Larson DE, Schmidt HK, Miller CA, Fulton RS, Spellman PT, Mardis ER, Druley TE, Graubert TA, Goodfellow PJ, Raphael BJ, Wilson RK, **Ding L.** Integrated analysis of germline and somatic variants in ovarian cancer. **Nat Commun**. 2014;5:3156. PMCID: PMC4025965.

b. Xie M, Lu C, Wang J, McLellan MD, Johnson KJ, Wendl MC, McMichael JF, Schmidt HK, Yellapantula V, Miller CA, Ozenberger BA, Welch JS, Link DC, Walter MJ, Mardis ER, Dipersio JF, Chen F, Wilson RK, Ley TJ, **Ding L**. Age-related mutations associated with clonal hematopoietic expansion and malignancies. **Nat Med.** 2014 Dec;20(12):1472-8. PMCID: PMC4313872.

c. Zhang J, Walsh MF, Wu G, Edmonson MN, Gruber TA, Easton J, Hedges D, Ma X, Zhou X, Yergeau DA, Wilkinson MR, Vadodaria B, Chen X, McGee RB, Hines-Dowell S, Nuccio R, Quinn E, Shurtleff SA, Rusch M, Patel A, Becksfort JB, Wang S, Weaver MS, **Ding L**, Mardis ER, Wilson RK, Gajjar A, Ellison DW, Pappo AS, Pui CH, Nichols KE, Downing JR. Germline mutations in predisposition genes in pediatric cancer. **N Engl J Med** 2015 Dec 10;373(24):2336-2346 PMC Journal – In Process

d. Lu C, Xie M, Wendl MC, Wang J, McLellan MD, Leiserson MD, Huang KL, Wyczalkowski MA, Jayasinghe R, Banerjee T, Ning J, Tripathi P, Zhang Q, Niu B, Ye K, Schmidt HK, Fulton RS, McMichael JF, Batra P, Kandoth C, Bharadwaj M, Koboldt DC, Miller CA, Kanchi KL, Eldred JM, Larson DE, Welch JS, You M, Ozenberger BA, Govindan R, Walter MJ, Ellis MJ, Mardis ER, Graubert TA, DiPersio JF, Ley TJ, Wilson RK, Goodfellow PJ, Raphael BJ, Chen F, Johnson KJ, Parvin JD, **Ding L**. Patterns and functional implications of rare germline variants across 12 cancer types. **Nat Commun** 2015 Dec 22;6:10086 PMC Journal – In Process PMCID: PMC4703835

D. Research Support

Ongoing Research Support

U01 HG006517-04, NHGRI - Ding, Li 02/01/2012-12/31/2016
A Turnkey System for High-throughput Variant Discovery and Interpretation (NCE)
The goal of this project is to make the analysis tools and next-generation pipelines currently in place in large genome centers available to the wider community, both individually and as part of a complete informatics solution.
Role: PI

R01 CA178383-01A1, NCI - Ding, Li 09/08/2014-08/31/2017
Virus Discovery and Characterization In Large-Scale Cancer Sequencing Data
We propose to develop a set of computational methods and analysis strategies for systematic discovery of integrated and episomal, DNA and RNA oncoviruses. We will perform simultaneous analysis of viruses and somatic/germline alterations in the host genome.
Role: PI

R01 CA180006-03, NCI - Ding, Li 02/012013-01/31/2017
Cancer Susceptibility Variant Discovery In High Throughput Sequencing Data
We will develop a computational pipeline for the identification and interpretation of germline alterations in cancer including single nucleotide variants, insertions and deletions (indels), copy number variations, and structural variants. This pipeline will be initially used to systematically analyze whole genome, exome, and RNA-sequencing data from over 5,000 cancer cases already generated by several major efforts and individual research groups and additional cases that will be made publicly available in the next several years.
Role: PI

5U24CA16003502, NIH-NCI (Chen/Ellis/Giddings/Townsend) 4/01/2011-07/31/2016
Cancer Proteomic Center at Washington University and University of North Carolina
The goal of this project is to assist the CPTAC in discovering new biomarkers, verifying their clinical applicability, and ultimately, helping translate selected biomarkers into clinical practice to reduce mortality from cancer.
Role: Co-Investigator

1U41HG007497, NIH - Lee 09/20/2013-08/31/2016
An Integrative Analysis of Structural Variation for the 1000 Genomes Project
We propose to pool expertise from various research groups to provide an integrative analysis of SVs by combining rigorous computational algorithmic development with extensive experimental validation.
Role: Co-investigator

1R01CA172652 , NIH - Chen, K 04/01/2013-03/31/2017
Delineating Heterogeneous Structural Complexity in Cancer Genomes
To fully harness the power of NGS and to facilitate advances toward personalized medicine, we propose to develop a set of novel computational tools for detecting structural variants in heterogeneous cancer genomes.
Role: Co-Investigator

2P01CA101937, NIH-NCI -Ley 09/19/2003-03/31/2018
Genomics of Acute Myelogenous Leukemia
The primary goal in this project is to utilize high throughput genomics technologies to define the commonly mutated target genes in AML that are relevant to clinical outcome.
Role: Co-Investigator

BIOGRAPHICAL SKETCH

Provide the following information for the Senior/key personnel and other significant contributors.
Follow this format for each person. DO NOT EXCEED FIVE PAGES.

NAME: Mark Gerstein

eRA COMMONS USER NAME (credential, e.g., agency login): MGERSTEIN

POSITION TITLE: Albert L. Williams Professor of Biomedical Informatics

EDUCATION/TRAINING *(Begin with baccalaureate or other initial professional education, such as nursing, include postdoctoral training and residency training if applicable. Add/delete rows as necessary.)*

INSTITUTION AND LOCATION	DEGREE *(if applicable)*	MM/YY	FIELD OF STUDY
Harvard College, Cambridge, MA	AB	06/1989	Physics
Cambridge University, Cambridge, UK	PhD	05/1993	Bioinformatics/Chemistry
Stanford University, Palo Alto, CA	post-doc	09/1996	Bioinformatics

A. Personal Statement

This proposal involves work in bioinformatics and computational genomics. Prof Gerstein is a leader in these fields and thus well-suited to be part of this the proposal. He has many peer-reviewed publications (as of '16, >500 total with an H-index via Google scholar of >135). He has served as the "computational" lead on many previous NIH-funded projects (e.g. ENCODE & 1000 Genomes). Most recently, he has developed quantitative approaches and practical tools for the processing of next-generation sequencing data, including those related to chIP-seq, RNA-seq and the detection of DNA structural variation. He has also developed approaches to analyze molecular networks and perform integrative data mining in a wide variety of contexts.

J Chen, J Rozowsky, T Galeev, A Harmanci, R Kitchen, J Bedford, A Abyzov, Y Kong, L Regan, **M Gerstein** (2016). "A uniform survey of allele-specific binding and expression over 1000-Genomes-Project individuals." *Nat Commun* 7: 11101 [PMC4837449]

C Cheng, E Andrews, K Yan, M Ung, D Wang, **M Gerstein** (2015). “An approach for determining and measuring network hierarchy applied to comparing the phosphorylome and the regulome.” *Genome Biol* 16: 63. [PMC4404648]

L Lochovsky, J Zhang, Y Fu, E Khurana, **M Gerstein** (2015). "LARVA: an integrative framework for large-scale analysis of recurrent variants in noncoding annotations." *NAR* 43: 8123. [PMC4787796]

A Abyzov, S Li, D Kim, M Mohiyuddin, A Stutz, N Parrish, X Mu, W Clark, K Chen, M Hurles, JO Korbel, H Lam, C Lee, **M Gerstein** (2015). "Analysis of deletion breakpoints from 1,092 humans reveals details of mutation mechanisms." *Nat Commun* 6: 7256. [PMC4451611]

B. Positions and Honors

Positions and Employment

2006-	AL Williams Prof. Biomedical Informatics, Yale
2002-	Co-Director Yale Computational Biology and Bioinformatics Program
1999-	Prof. of Computer Science, Yale (asst., '99-'01; assoc. '01-'06)
1997-	Prof. Molecular Biophysics & Biochemistry, Yale (asst., '97-'01; assoc '01-'06)

Honors

1989-1993	Herchel-Smith Scholarship funding for PhD at Cambridge
1993-1996	Damon Runyon-Walter Winchell post-doctoral Fellowship
1997-2001	Young Investigator Awards from Navy & IBM, PhRMA, Donaghue, & Keck foundations
2009	AAAS Fellow
2015	ISCB Fellow

Other Experience and Professional Memberships

Editorial boards: Genome Res., MSB, J Struc Func Gen., PLoS Comp Bio, GenomeBiology

Analysis Working Group co-chair: modENCODE ('07-'14), exRNA Consortium ('13-), 1000 Genomes Functional Interpretation Group ('10-'15), PsychENCODE Consortium (’14-), Pan-Cancer Analysis Working Group #2 (regulatory drivers)('14-)

C. Contribution to Science

Human Genome Annotation & Interpretation of Variants

The Gerstein lab has made a number of contributions to developing large-scale human genome annotation, ranging from noncoding RNAs to enhancers to pseudogenes, and using these annotations to interpret variants in personal genomes in a functional context. Our tools address both germline and somatic variants. Our interpretation scheme ranks these variants in relation to their deleteriousness in causing disease, and also interprets potential functional effects.

Y Fu, Z Liu, S Lou, J Bedford, X Mu, KY Yip, E Khurana, **M Gerstein** (2014). "FunSeq2: A framework for prioritizing noncoding regulatory variants in cancer." *Genome Biol* 15: 480. [PMC4203974]

E Khurana, Y Fu, V Colonna, XJ Mu... (42 authors)... H Yu, MA Rubin, C Tyler-Smith, **M Gerstein** (2013). "Integrative Annotation of Variants from 1092 Humans: Application to Cancer Genomics." *Science* 342: 1235587 [PMC3947637].

E Khurana, Y Fu, J Chen, **M Gerstein** (2013). "Interpretation of genomic variants using a unified biological network approach." *PLoS Comput Biol* 9: e1002886. [PMC3591262]

E Khurana, Y Fu, D Chakravaty, F Demichelis, MA Rubin, **M Gerstein** (2016). “Role of non-coding sequence variants in cancer.” *Nat Rev Genet.* 2:93-98. [PMID26781813]

Personal Genomics & Privacy

The unique character of each individual’s genome has potential impacts ranging from disease propensity to physical appearance to intelligence. We have developed tools to build personal genomes from DNA-sequencing data and to link molecular phenotypes such as gene expression to differences in parental alleles. We have also developed tools to address the critical question of whether it is possible to share molecular data without compromising the identities or the highly personal genetic information of sample donors.

J Rozowsky, A Abyzov, J Wang, P Alves, D Raha, A Harmanci, J Leng, R Bjornson, Y Kong, N Kitabayashi, N Bhardwaj, M Rubin, M Snyder, **M Gerstein** (2011). "AlleleSeq: analysis of allele-specific expression and binding in a network framework." *Mol Syst Biol* 7: 522. [PMC3208341]

L Habegger, A Sboner, TA Gianoulis, J Rozowsky, A Agarwal, M Snyder, **M Gerstein** (2011). RSEQtools: a modular framework to analyze RNA-Seq data using compact, anonymized data summaries. *Bioinformatics* 27: 281. [PMC3018817]

D Greenbaum, A Sboner, X J Mu, **M Gerstein** (2011). “Genomics and Privacy: Implications of the New Reality of Closed Data for the Field” *PLoS Comput Biol* 7: e1002278 [PMC3228779]

A Harmanci, **M Gerstein** (2016). “Quantification of private information leakage from phenotype-genotype data: linking attacks.” *Nat Methods* 13:251. [PMC4834871]

Comparative & Integrative Genomics

We have developed a number of approaches for comparing the human genome to the genomes of model organisms. Our comparative analyses, particularly for the transcriptome, have yielded conserved principles of regulation. We have also developed integrative models that relate the transcriptome to the epigenome, and for combining these together to improve regulatory region annotation.

M Gerstein, J Rozowsky, KK Yan, D Wang...(89 authors)... TR Gingeras, R Waterston (2014). "Comparative analysis of the transcriptome across distant species." *Nature* 512: 445. [PMC4155737]

C Sisu, B Pei, J Leng, A Frankish, Y Zhang, S Balasubramanian, R Harte, D Wang, M Rutenberg-Schoenberg, W Clark, M Diekhans, J Rozowsky, T Hubbard, J Harrow, **M Gerstein** (2014). "Comparative analysis of pseudogenes across three phyla." *PNAS* 111: 13361. [PMC4169933]

KK Yan, D Wang, J Rozowsky, H Zheng, C Cheng, **M Gerstein** (2014). "OrthoClust: an orthology-based network framework for clustering data across multiple species." *Genome Biology* 15:R100 [PMC4289247]

M Gerstein, ZJ Lu... (128 authors)... L Stein, JD Lieb, RH Waterston (2010). "Integrative analysis of the Caenorhabditis elegans genome by the modENCODE project." *Science* 330: 1775. [PMC3142569].

Analysis of Diverse Networks

Network representations can be applied consistently to many different types of biological data. We have developed tools to build and analyze regulatory networks, protein-protein interactions and metabolic pathways, identifying key nodes such as hubs and bottlenecks. Moreover, we have integrated networks with dynamic gene-expression data (identifying transient hubs), 3D-protein structures, and other regulatory data to find large-scale regulatory principles for biological systems.

M Gerstein, A Kundaje... (50 authors)... R Myers, S Weissman, M Snyder (2012). " Architecture of the human regulatory network derived from ENCODE data." *Nature* 489: 91 [PMC4154057]

D Wang, KK Yan, C Sisu, C Cheng, J Rozowsky, W Meyerson, **M Gerstein** (2015). “Loregic: a method to characterize the cooperative logic of regulatory factors.” *PLoS Comput Biol* 11: e1004132. [PMC4401777]

PM Kim, LJ Lu, Y Xia, **M Gerstein** (2006). “Relating three-dimensional structures to protein networks provides evolutionary insights.” *Science* 314:1938-41. [PMID17185604]

K Yan, G Fang, N Bhardwaj, R Alexander, **M Gerstein** (2010). "Comparing genomes to computer operating systems in terms of the topology and evolution of their regulatory control networks." *PNAS* 107: 9186. [PMC2889091]

Tools for Processing Next-Gen Sequencing Data

Next-gen sequencing has been one of the most exciting advances in the biological sciences, producing data on an unprecedented scale. This has given rise to the need to create new tool sets that can process very large-scale data very efficiently. We have developed tool sets that address a wide range of biological problems from sequencing data, including calling structural genetic variants and annotating specific regions of biological activity.

KY Yip, C Cheng, N Bhardwaj, JB Brown, J Leng, A Kundaje, J Rozowsky, E Birney, P Bickel, M Snyder, **M Gerstein** (2012). "Classification of human genomic regions based on experimentally determined binding sites of more than 100 transcription-related factors." *Genome Biol* 13: R48. [PMC3491392]

A Abyzov, AE Urban, M Snyder, **M Gerstein** (2011). "CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing." *Genome Res* 21: 974-84. [PMC3106330]

A Harmanci, J Rozowsky, **M Gerstein** (2014). "MUSIC: Identification of Enriched Regions in ChIP-Seq Experiments using a Mappability-Corrected Multiscale Signal Processing Framework." *Genome Biol* 15: 474. [PMC4234855]

A Abyzov, R Iskow, O Gokcumen, DW Radke, S Balasubramanian, B Pei, L Habegger, The 1000 Genomes Project Consortium, C Lee, **M Gerstein** (2013). "Analysis of variable retroduplications in human populations suggests coupling of retrotransposition to cell division." *Genome Res* 23: 2042 [PMC3847774]

Complete List of Publications

http://www.ncbi.nlm.nih.gov/sites/myncbi/mark.gerstein.1/bibliography/44005333/public

D. Research Support

Ongoing Research Support

U01 MH103365-01 - 06/15/14-05/31/17 - NIH
Gene regulatory elements and transcriptome in iPSCs and embryonic human cortex
- Role: Multi-PI (PIs: Vaccarino, Gerstein, Weissman) This application is in response to FOA [MH-14-020 ("PsychENCODE")]. The major aims are to provide a comprehensive catalogue of all types of RNA and their regulatory elements in the cerebral cortex of the mid-gestational human fetal brain as compared to induced pluripotent stem cells (iPSCs) derived from the same fetuses.

R01 DA030976-05 - 09/30/10-05/31/16 - University of North Carolina - NIH (NCE)
Deep sequencing studies for cannabis and stimulant dependence
- Role: Co-I (PI: Wilhelmsen) The major goals of this project are to determine structural variants in the genome from deep sequencing studies for cannabis and stimulant dependence.

DE-AC02-98CH10886 - 10/17/11-07/31/16 - Brookhaven National Laboratory - DOE
Kbase: An Integrated Knowledgebase for Predictive Biology and Environmental Research
- Role: Co-PI (PI: Maslov) The major goal of this project is to assist in the construction of the DOE Knowledgebase. Our role is to provide support to the plant and microbial subcomponents.

U41 HG007000-03 - 09/21/12-07/31/16 - University of Massachusetts - NIH
EDAC: ENCODE Data Analysis Center
- Role: Co-I (PI: Weng) The major goal of this project is to perform global and integrative data analysis for the ENCODE project.

U41 HG007355-02 - 09/20/13-07/31/17 - University of Washington - NIH
Creating Comprehensive Maps of Worm and Fly Transcription Factor Binding Sites
- Role: Co-I (PI: Waterston) Our role on the project is the determination of binding sites for transcription factors in worm and the fly. We will analyze large-scale Chip-Seq experiments to identify regions in the genome that are bound by transcription factors.

U43 DA036134-01 - 08/01/13-07/31/18 - Baylor College of Medicine - NIH
Data Management and Resource Repository for the exRNA Atlas
- Role: Multi-PI (PIs: Gerstein, Galas, Milosavljevic) Our role on the project is administering the DIAC (data integration and analysis center) for ex-RNA data.

U41 HG007234-02 - 04/01/13-03/31/17 - Wellcome Trust - NIH
GENCODE: Comprehensive gene annotation for human and mouse
- Role: Co-I (PI: Harrow) Our role in the project is to identify pseudogenes comprehensively in human and mouse genomes and provide a systematic annotation of them.

U41 HG007497-02 - 09/20/13-08/31/16 - Jackson Laboratory - NIH
An Integrative Analysis of Structural Variation for the 1000 Genomes Project
- Role: Co-I (PI: Lee) Our role on the project is analyzing the 1000 genomes data set to determine structural variation on a large scale.

R01 GM108663-02 - 02/01/14-01/31/18 - NIH
Deciphering mechanisms governing functional partitioning of the C. elegans genome
- Role: Co-I (PI: Reinke) Our role on the project is assisting the PI with bioinformatic analyses related to the worm genome.

R01 MH100914-01A1 - 01/01/14-12/31/18 - NIH
Genomic mosaicism in developing human brain
- Role: Multi-PI (PIs: Vaccarino, Sestan, Gerstein) Our role on the project is to analyze somatic variations in the human genome.

U01 HL126495-01 - 08/01/14-04/30/19 - U Massachusetts - NIH
Racial and Ethnic Diversity in Human Extracellular RNA
- Role: Multi-PI (PIs: Freedman, Gerstein, Mukamal, O'Donnell) The primary goal of this proposal is the generation of exRNA profiles in healthy individuals in two large and well-defined cohorts, the Framingham Heart Study and the Multi-Ethnic Study of Atherosclerosis, to be used as a reference to facilitate disease diagnosis and discovery.

P50 MH106934-01 - 09/19/14-07/31/19 - NIH
Functional Genomics of Human Brain Development
- Role: Co-I (PI: Sestan) This grant will apply functional genomics to study human brain development. Our role is do analyses of these datasets.

3P50 MH106934-02S1 - 08/01/15 - 07/31/17 - NIH (concurrent with parent)
-Role: Co-I (PI: Sestan) Functional Genomics of Human Brain Development
Supplement to P50 MH106934 (above), effort on parent grant.

P30 DA018343-11A1 - 07/01/15-05/31/20 - Yale/NIDA Neuroproteomics Research Center - NIH
- Role: Co-I (PI: Williams) The overall grant is funding for a neuroproteomics center at Yale. The Gerstein lab contribution is to develop approaches for comparing and correlating protein abundance and mRNA levels in relation to neuroproteomics.

2UM1HG006504-05 - 01/14/16 - 11/30/19 - NIH
Yale Center for Mendelian Genomics
- Role: Multi PI (PIs: Lifton, Gerstein, Gunel, Mane) The majority of genomic variation in Mendelian disorders is due to variation in protein coding regions in the genome. Sequencing of these regions allow for rapid identification of disease causing mutations, which will allow us to understand and dissect the biology of these disorders leading to better diagnostic and therapeutic tools.

Completed Research Support in the Last Three Years

5U54 HG004558-05 - 01/01/10-06/30/13
Production Center for Global Mapping of Regulatory Elements
- Role: Co-I (PI: Snyder) The major goal of this project is to comprehensively probe transcription factor binding throughout the human genome.

5U54 HG004555-04S1 - 09/27/07-03/31/13
Integrated human genome annotation: generation of a reference gene set
- Role: Co-I (PI: Hubbard) The major goal of this project is to construct a pseudogene annotation of the human genome.

5U01 HG004267-05S1 - 01/01/10-03/31/13
Global Identification of Transcription Factor Binding Sites in *C. elegans*
- Role: Co-I (PI: Snyder) The major goal of this project is to build a genome-wide map of the binding sites for every *C. elegans* transcription factor.

5U01 HG005718-02 - 09/13/10-06/30/14 - NIH
Loss-of-function variants in the 1000 genomes data set and implications to GWAS
- Role: Co-I (PI: Zhao). The goals of this project are to survey loss-of-function (LOF) variants in the 1000 genomes data set and make this analysis available to the community as a useful resource.

5R01CA152057-03 - 08/01/11-07/31/15 - Weill Medical College of Cornell - NIH
Comprehensive Prostate Cancer Characterization by Genomic and Transcriptomic Profiling
- Role: Co-I (PI: Rubin) The major goal of this project is to identify biomarkers for prostate cancer through analysis of various types of 'omics data.

DE-SC0004856 - 08/15/10-08/14/15 - DOE
Tools and Models for Integrating Multiple Cellular Network
- Role: PI The major goals of this project are the development of tools for the analysis of network and pathways in micro-organisms for the Systems Biology Knowledgebase proposed by the DOE.

5U54HG006504-03 - 12/05/11-11/30/15 - NIH
Yale Center for Mendelian Disorders
- Role: Multi-PI (PIs: Gerstein, Lifton, Gunel, Mane) The major goal of this project is to develop informatics approaches to characterize rare variants in the framework of the Centers for Mendelian Genomics.

OMB No. 0925-0001/0002 (Rev. 08/12 Approved Through 8/31/2015)

BIOGRAPHICAL SKETCH

NAME: Malhotra, Ankit

eRA COMMONS USER NAME: ANKIT.MALHOTRA

POSITION TITLE: Associate Computational Scientist

EDUCATION/TRAINING

INSTITUTION AND LOCATION	DEGREE	Completion Date MM/YYYY	FIELD OF STUDY
Hans Raj College, University of Delhi, Delhi, India	B.S.	05/2003	Computer Science
University of Virginia, Charlottesville, VA	M.S.	01/2007	Computer Science
University of Virginia, Charlottesville, VA	Ph.D.	08/2010	Biochemistry

A. Personal Statement

The work described in the project would involve creating methods for discovering complex structural variation and INDELs from whole genome sequence datasets being generated by the various projects of the TOPMed program. The identified variants would be used in cross-program analyses to gain novel biological insight in human biology and disease mechanisms.

I have been working in the field of genetic variation for about eight years and have authored several important publications during my doctoral studies (in the lab of Dr. Anindya Dutta) as well as a postdoctoral researcher in the lab of Dr. Ira M Hall at the University of Virginia. My basic training (B.S. and M.S.) was in the field of Computer Science, and I earned my Ph.D. from the Department of Biochemistry at University of Virginia. This cross-disciplinary training has enabled me to bring sophisticated algorithms and methods from the field of computer science and apply them to important questions in biology. I was one of the first people to analyze data from high-throughput sequencing machines and developed important algorithms to discover and characterize structural variations in yeast and subsequently in human genomes. During my postdoctoral training, I worked on meta-analysis of complex structural variants from a large cohort of cancer datasets from the TCGA consortium and made very interesting and impactful observations. More recently, as part of the 1000 Genomes project, I have been involved in developing new methods and analysis to help generate the largest cohort of germline SVs from 2,504 individuals across 27 different populations. In total, I have authored/coauthored 18 peer reviewed publications on various aspects of genome analysis.

For this project, I will provide my expertise in the field of structural variation analysis. I will be involved in novel method development and analysis of sequencing datasets from the different TOPMed cohorts. As an Associate Computational Scientist at The Jackson Laboratory, I will lend my skills and experience in bioinformatics, cancer biology and genomic data analysis to this project.

B. Positions and Honors

2002-2003	Research Assistant, Dr. Harmeet Kaur, Hans Raj College, University of Delhi, Delhi, India
2003-2003	Research Assistant, Dr. P.R. Panda, Indian Institute of Technology, Delhi
2003-2010	Graduate Research Assistant, Dr. Anindya Dutta, University of Virginia, Charlottesville, VA
2010-2013	Research Associate, Dr. Ira M Hall, Department of Biochemistry and Molecular Genetics, University of Virginia, Charlottesville, VA
2013-Present	Associate Computational Scientist, Computational Sciences, The Jackson Laboratory for Genomic Medicine, Farmington, CT

C. Contribution to Science

My primary contributions are in field of cancer genomics, more specifically the study of structural variation and its impact on human health and disease. Over the last 8 years, I have contributed methods for discovering structural variation using next generation sequencing data. In 2013, we published the first meta-analysis of complex variations in multiple cancer genomes sequenced as part of the TCGA (The Cancer Genome Atlas) consortium. This study highlighted the importance of studying complex events, as they could be the initiating events in a patient's history. We also studied mechanisms that gave rise to such events and concluded that non-homologous repair of concurrently arising DNA double-strand breaks is the predominant mechanism underlying complex cancer genome rearrangements. I served as the primary investigator / analyst for all of these studies. More recently, as part of the Phase 3 of the 1000 Genomes project, I have also been involved in generating the largest set of known germline SVs from 2,504 individuals across 27 different world populations.

1. Shibata, Y., Malhotra, A. & Dutta, A. Detection of DNA fusion junctions for BCR-ABL translocations by Anchored ChromPET. *Genome Med* **2,** 70 (2010).
2. Malhotra, A. *et al.* Breakpoint profiling of 64 cancer genomes reveals numerous complex rearrangements spawned by homology-independent mechanisms. *Genome Res.* **23,** 762–776 (2013).
3. Malhotra, A. *et al.* Ploidy-Seq: inferring mutational chronology by sequencing polyploid tumor subpopulations. *Genome Med* **7,** 6–6 (2015).
4. Sudmant, P et al. An integrated map of structural variation in 2,504 human genomes. *Nature 2015; 526(7571):75-81*

My earliest publications were as part of the pilot phase of the ENCyclopedia Of DNA Elements (ENCODE) project. I was part of a two-person team that worked on the analysis of the genomic data from the cell lines to produce a time of replication across the 1% of the genome. These resulted in the following publications:

1. ENCODE Project Consortium *et al.* Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447,** 799–816 (2007).
2. Karnani, N., Taylor, C., Malhotra, A. & Dutta, A. Pan-S replication patterns and chromosomal domains defined by genome-tiling arrays of ENCODE genomic areas. *Genome Res.* **17,** 865–876 (2007).
3. Karnani, N., Taylor, C. M., Malhotra, A. & Dutta, A. Genomic study of replication initiation in human chromosomes reveals the influence of transcription regulation and chromatin structure on origin selection. *Mol. Biol. Cell* **21,** 393–404 (2010).

Link to Publications:
(http://www.ncbi.nlm.nih.gov/sites/myncbi/ankit.malhotra.1/bibliography/48094753/public/?sort=date&direction=ascending)

D. Research Support

Current Research Support

U41 HG007497 Lee (PI) 08/01/13-06/30/17
NIH/NHGRI
An integrative analysis of structural variation for the 1000 Genomes Project
The major goal of this project is to develop and assess new methods for accurately identifying structural genomic variants in next generation DNA sequencing datasets.
Role: Co-Investigator, Computational Scientist

Completed Research Support

2011-2013: **Department of Defense Breast Cancer Postdoctoral Fellowship Award**, for proposal titled "*Role and mechanism of structural variation in progression of breast cancer*", awarded by Office of the Congressionally Directed Medical Research Programs (CDMRP), US Dept. of Defense
Role: PI

RESEARCH & RELATED BUDGET - SECTION A & B, BUDGET PERIOD 1

*** ORGANIZATIONAL DUNS:** 042140483

*** Budget Type:** ● Project ○ Subaward/Consortium

Enter name of Organization: The Jackson Laboratory

*** Start Date:** 04-01-2017 *** End Date:** 03-31-2018 **Budget Period: 1**

A. Senior/Key Person

Prefix	* First Name	Middle Name	* Last Name	Suffix	* Project Role	Base Salary ($)	Calendar Months	Academic Months	Summer Months	* Requested Salary ($)	* Fringe Benefits ($)	* Funds Requested ($)
1.	Charles		Lee		PD/PI	185,100.00	1.20	0	0	18,510.00	5,553.00	24,063.00
2.	Ankit		Malhotra		Co-Investigator	104,110.00	3.00	0	0	26,027.50	7,808.25	33,835.75

Total Funds Requested for all Senior Key Persons in the attached file

Additional Senior Key Persons: File Name: **Total Senior/Key Person** **57,898.75**

Mime Type:

B. Other Personnel

* Number of Personnel	* Project Role	Cal. Months	Acad. Months	Sum. Months	* Requested Salary ($)	* Fringe Benefits	* Funds Requested ($)
1	Post Doctoral Associates	12	0	0	55,000.00	16,500.00	71,500.00
0	Graduate Students	0	0	0	0.00	0.00	0.00
0	Undergraduate Students	0	0	0	0.00	0.00	0.00
0	Secretarial/Clerical	0	0	0	0.00	0.00	0.00
0	Other	0	0	0	0.00	0.00	0.00
0	Other Professionals	0	0	0	0.00	0.00	0.00
0	Allocated Admin Support	0	0	0	0.00	0.00	0.00
1	**Total Number Other Personnel**					**Total Other Personnel**	**71,500.00**
						Total Salary, Wages and Fringe Benefits (A+B)	**129,398.75**

RESEARCH & RELATED Budget {A-B} (Funds Requested)

RESEARCH & RELATED BUDGET - SECTION C, D, & E, BUDGET PERIOD 1

*** ORGANIZATIONAL DUNS:** 042140483

*** Budget Type:** ● Project ○ Subaward/Consortium

Enter name of Organization: The Jackson Laboratory

*** Start Date:** 04-01-2017 *** End Date:** 03-31-2018 **Budget Period: 1**

C. Equipment Description

List items and dollar amount for each item exceeding $5,000

Equipment Item	*** Funds Requested ($)**
Total funds requested for all equipment listed in the attached file	
Total Equipment	

Additional Equipment:

File Name: Mime Type:

D. Travel	**Funds Requested ($)**
1. Domestic Travel Costs (Incl. Canada, Mexico, and U.S. Possessions)	3,000.00
2. Foreign Travel Costs	0.00
Total Travel Cost	**3,000.00**

E. Participant/Trainee Support Costs		**Funds Requested ($)**
1. Tuition/Fees/Health Insurance		0.00
2. Stipends		0.00
3. Travel		0.00
4. Subsistence		0.00
5. Other: Other		0.00
0 Number of Participants/Trainees	**Total Participant Trainee Support Costs**	**0.00**

RESEARCH & RELATED Budget {C-E} (Funds Requested)

RESEARCH & RELATED BUDGET - SECTIONS F-K, BUDGET PERIOD 1

*** ORGANIZATIONAL DUNS:** 042140483

*** Budget Type:** ● Project ○ Subaward/Consortium

Enter name of Organization: The Jackson Laboratory

*** Start Date:** 04-01-2017 *** End Date:** 03-31-2018 **Budget Period: 1**

F. Other Direct Costs	**Funds Requested ($)**
1. Materials and Supplies	934.25
2. Publication Costs	0.00
3. Consultant Services	0.00
4. ADP/Computer Services	0.00
5. Subawards/Consortium/Contractual Costs	419,407.77
6. Equipment or Facility Rental/User Fees	0.00
7. Alterations and Renovations	0.00
8. Other Direct Costs	0.00
9. All Other Costs	0.00
Total Other Direct Costs	**420,342.02**

G. Direct Costs	**Funds Requested ($)**
Total Direct Costs (A thru F)	**552,740.77**

H. Indirect Costs

Indirect Cost Type	**Indirect Cost Rate (%)**	**Indirect Cost Base ($)**	*** Funds Requested ($)**
1. MTDC	89	183,333.00	163,166.37
		Total Indirect Costs	**163,166.37**

Cognizant Federal Agency HHS, Ryan McCarthy 212-264-2069

(Agency Name, POC Name, and POC Phone Number)

I. Total Direct and Indirect Costs	**Funds Requested ($)**
Total Direct and Indirect Institutional Costs (G + H)	**715,907.14**

J. Fee	**Funds Requested ($)**
	0.00

K. * Budget Justification File Name: M-12_S2S_Budget_Justification.pdf Mime Type: application/octet-stream

(Only attach one file.)

RESEARCH & RELATED Budget {F-K} (Funds Requested)

RESEARCH & RELATED BUDGET - SECTION A & B, BUDGET PERIOD 2

*** ORGANIZATIONAL DUNS:** 042140483

*** Budget Type:** ● Project ○ Subaward/Consortium

Enter name of Organization: The Jackson Laboratory

*** Start Date:** 04-01-2018 *** End Date:** 03-31-2019 **Budget Period: 2**

A. Senior/Key Person

	Prefix	*** First Name**	**Middle Name**	*** Last Name**	**Suffix**	*** Project Role**	**Base Salary ($)**	**Calendar Months**	**Academic Months**	**Summer Months**	*** Requested Salary ($)**	*** Fringe Benefits ($)**	*** Funds Requested ($)**
1.		Charles		Lee		PD/PI	190,653.00	1.20	0	0	19,065.30	5,719.59	24,784.89
2.		Ankit		Malhotra		Co-Investigator	107,233.00	2.40	0	0	21,446.66	6,434.00	27,880.66

Total Funds Requested for all Senior Key Persons in the attached file

Additional Senior Key Persons: File Name: **Total Senior/Key Person** **52,665.55**

Mime Type:

B. Other Personnel

*** Number of Personnel**	*** Project Role**	**Cal. Months**	**Acad. Months**	**Sum. Months**	*** Requested Salary ($)**	*** Fringe Benefits**	*** Funds Requested ($)**
1	Post Doctoral Associates	12	0	0	56,650.00	16,995.00	73,645.00
0	Graduate Students	0	0	0	0.00	0.00	0.00
0	Undergraduate Students	0	0	0	0.00	0.00	0.00
0	Secretarial/Clerical	0	0	0	0.00	0.00	0.00
0	Other	0	0	0	0.00	0.00	0.00
0	Other Professionals	0	0	0	0.00	0.00	0.00
0	Allocated Admin Support	0	0	0	0.00	0.00	0.00
1	**Total Number Other Personnel**					**Total Other Personnel**	**73,645.00**
						Total Salary, Wages and Fringe Benefits (A+B)	**126,310.55**

RESEARCH & RELATED Budget {A-B} (Funds Requested)

RESEARCH & RELATED BUDGET - SECTION C, D, & E, BUDGET PERIOD 2

*** ORGANIZATIONAL DUNS:** 042140483

*** Budget Type:** ● Project ○ Subaward/Consortium

Enter name of Organization: The Jackson Laboratory

*** Start Date:** 04-01-2018 *** End Date:** 03-31-2019 **Budget Period: 2**

C. Equipment Description

List items and dollar amount for each item exceeding $5,000

Equipment Item	*** Funds Requested ($)**
Total funds requested for all equipment listed in the attached file	
Total Equipment	

Additional Equipment:

File Name: Mime Type:

D. Travel	**Funds Requested ($)**
1. Domestic Travel Costs (Incl. Canada, Mexico, and U.S. Possessions)	3,000.00
2. Foreign Travel Costs	0.00
Total Travel Cost	**3,000.00**

E. Participant/Trainee Support Costs		**Funds Requested ($)**
1. Tuition/Fees/Health Insurance		0.00
2. Stipends		0.00
3. Travel		0.00
4. Subsistence		0.00
5. Other: Other		0.00
0 Number of Participants/Trainees	**Total Participant Trainee Support Costs**	**0.00**

RESEARCH & RELATED Budget {C-E} (Funds Requested)

RESEARCH & RELATED BUDGET - SECTIONS F-K, BUDGET PERIOD 2

*** ORGANIZATIONAL DUNS:** 042140483

*** Budget Type:** ● Project ❍ Subaward/Consortium

Enter name of Organization: The Jackson Laboratory

*** Start Date:** 04-01-2018 *** End Date:** 03-31-2019 **Budget Period: 2**

F. Other Direct Costs	**Funds Requested ($)**
1. Materials and Supplies	4,022.45
2. Publication Costs	0.00
3. Consultant Services	0.00
4. ADP/Computer Services	0.00
5. Subawards/Consortium/Contractual Costs	426,157.77
6. Equipment or Facility Rental/User Fees	0.00
7. Alterations and Renovations	0.00
8. Other Direct Costs	0.00
9. All Other Costs	0.00
Total Other Direct Costs	**430,180.22**

G. Direct Costs	**Funds Requested ($)**
Total Direct Costs (A thru F)	**559,490.77**

H. Indirect Costs

Indirect Cost Type	**Indirect Cost Rate (%)**	**Indirect Cost Base ($)**	*** Funds Requested ($)**
1. MTDC	89	133,333.00	118,666.37
		Total Indirect Costs	**118,666.37**

Cognizant Federal Agency HHS, Ryan McCarthy 212-264-2069

(Agency Name, POC Name, and POC Phone Number)

I. Total Direct and Indirect Costs	**Funds Requested ($)**
Total Direct and Indirect Institutional Costs (G + H)	**678,157.14**

J. Fee	**Funds Requested ($)**
	0.00

K. * Budget Justification File Name: M-12_S2S_Budget_Justification.pdf Mime Type: application/octet-stream

(Only attach one file.)

RESEARCH & RELATED BUDGET - SECTION A & B, BUDGET PERIOD 3

*** ORGANIZATIONAL DUNS:** 042140483

*** Budget Type:** ● Project ○ Subaward/Consortium

Enter name of Organization: The Jackson Laboratory

*** Start Date:** 04-01-2019 *** End Date:** 03-31-2020 **Budget Period: 3**

A. Senior/Key Person

	Prefix	* First Name	Middle Name	* Last Name	Suffix	* Project Role	Base Salary ($)	Calendar Months	Academic Months	Summer Months	* Requested Salary ($)	* Fringe Benefits ($)	* Funds Requested ($)
1.		Charles		Lee		PD/PI	196,373.00	1.20	0	0	19,637.26	5,891.18	25,528.44
2.		Ankit		Malhotra		Co-Investigator	110,450.00	2.40	0	0	22,090.06	6,627.02	28,717.08

Total Funds Requested for all Senior Key Persons in the attached file

Additional Senior Key Persons: File Name: **Total Senior/Key Person** **54,245.52**

Mime Type:

B. Other Personnel

* Number of Personnel	* Project Role	Cal. Months	Acad. Months	Sum. Months	* Requested Salary ($)	* Fringe Benefits	* Funds Requested ($)
1	Post Doctoral Associates	12	0	0	58,349.50	17,504.85	75,854.35
0	Graduate Students	0	0	0	0.00	0.00	0.00
0	Undergraduate Students	0	0	0	0.00	0.00	0.00
0	Secretarial/Clerical	0	0	0	0.00	0.00	0.00
0	Other	0	0	0	0.00	0.00	0.00
0	Other Professionals	0	0	0	0.00	0.00	0.00
0	Allocated Admin Support	0	0	0	0.00	0.00	0.00
1	**Total Number Other Personnel**					**Total Other Personnel**	**75,854.35**
						Total Salary, Wages and Fringe Benefits (A+B)	**130,099.87**

RESEARCH & RELATED Budget {A-B} (Funds Requested)

RESEARCH & RELATED BUDGET - SECTION C, D, & E, BUDGET PERIOD 3

*** ORGANIZATIONAL DUNS:** 042140483

*** Budget Type:** ● Project ❍ Subaward/Consortium

Enter name of Organization: The Jackson Laboratory

*** Start Date:** 04-01-2019 *** End Date:** 03-31-2020 **Budget Period: 3**

C. Equipment Description

List items and dollar amount for each item exceeding $5,000

Equipment Item	*** Funds Requested ($)**
Total funds requested for all equipment listed in the attached file	
Total Equipment	

Additional Equipment:

File Name: Mime Type:

D. Travel	**Funds Requested ($)**
1. Domestic Travel Costs (Incl. Canada, Mexico, and U.S. Possessions)	3,000.00
2. Foreign Travel Costs	0.00
Total Travel Cost	**3,000.00**

E. Participant/Trainee Support Costs		**Funds Requested ($)**
1. Tuition/Fees/Health Insurance		0.00
2. Stipends		0.00
3. Travel		0.00
4. Subsistence		0.00
5. Other: Other		0.00
0 Number of Participants/Trainees	**Total Participant Trainee Support Costs**	**0.00**

RESEARCH & RELATED Budget {C-E} (Funds Requested)

RESEARCH & RELATED BUDGET - SECTIONS F-K, BUDGET PERIOD 3

*** ORGANIZATIONAL DUNS:** 042140483

*** Budget Type:** ● Project ❍ Subaward/Consortium

Enter name of Organization: The Jackson Laboratory

*** Start Date:** 04-01-2019 *** End Date:** 03-31-2020 **Budget Period: 3**

F. Other Direct Costs	**Funds Requested ($)**
1. Materials and Supplies	233.13
2. Publication Costs	0.00
3. Consultant Services	0.00
4. ADP/Computer Services	0.00
5. Subawards/Consortium/Contractual Costs	426,157.77
6. Equipment or Facility Rental/User Fees	0.00
7. Alterations and Renovations	0.00
8. Other Direct Costs	0.00
9. All Other Costs	0.00
Total Other Direct Costs	**426,390.90**

G. Direct Costs	**Funds Requested ($)**
Total Direct Costs (A thru F)	**559,490.77**

H. Indirect Costs

Indirect Cost Type	**Indirect Cost Rate (%)**	**Indirect Cost Base ($)**	*** Funds Requested ($)**
1. MTDC	89	133,333.00	118,666.37
		Total Indirect Costs	**118,666.37**

Cognizant Federal Agency HHS, Ryan McCarthy 212-264-2069

(Agency Name, POC Name, and POC Phone Number)

I. Total Direct and Indirect Costs	**Funds Requested ($)**
Total Direct and Indirect Institutional Costs (G + H)	**678,157.14**

J. Fee	**Funds Requested ($)**
	0.00

K. * Budget Justification File Name: M-12_S2S_Budget_Justification.pdf Mime Type: application/octet-stream

(Only attach one file.)

RESEARCH & RELATED Budget {F-K} (Funds Requested)

RESEARCH & RELATED BUDGET - Cumulative Budget

	Totals ($)	
Section A, Senior/Key Person		**164,809.82**
Section B, Other Personnel		**220,999.35**
Total Number Other Personnel	3	
Total Salary, Wages and Fringe Benefits (A+B)		**385,809.17**
Section C, Equipment		**0.00**
Section D, Travel		**9,000.00**
1. Domestic	9,000.00	
2. Foreign	0.00	
Section E, Participant/Trainee Support Costs		**0.00**
1. Tuition/Fees/Health Insurance	0.00	
2. Stipends	0.00	
3. Travel	0.00	
4. Subsistence	0.00	
5. Other	0.00	
6. Number of Participants/Trainees	0	
Section F, Other Direct Costs		**1,276,913.14**
1. Materials and Supplies	5,189.83	
2. Publication Costs	0.00	
3. Consultant Services	0.00	
4. ADP/Computer Services	0.00	
5. Subawards/Consortium/Contractual Costs	1,271,723.31	
6. Equipment or Facility Rental/User Fees	0.00	
7. Alterations and Renovations	0.00	
8. Other 1	0.00	
9. Other 2		
10. Other 3		
Section G, Direct Costs (A thru F)		**1,671,722.31**
Section H, Indirect Costs		**400,499.11**
Section I, Total Direct and Indirect Costs (G + H)		**2,072,221.42**
Section J, Fee		**0.00**

BUDGET JUSTIFICATION – The Jackson Laboratory for Genomic Medicine

SENIOR/KEY PERSONNEL

Charles Lee, Ph.D., Principal Investigator (Contact PI) (1.2 calendar months), is Scientific Director and Professor at The Jackson Laboratory for Genomic Medicine (JAX-GM) will serve as Principal Investigator and Contact PI. As Contact PI, he will coordinate communication among the PIs, key personnel, and NIH scientific and program officials, and will be responsible for coordinating the preparation and submission of annual progress reports. Dr. Lee has over 15 years experience successfully leading both R01- and collaborative U-mechanism projects, with a particular emphasis on the detection of genetic variants using state-of-the-art technologies. For this TOPMed, he will be responsible for the overall scientific direction of the development of an integrative pipeline of tools for complex structural variation (SV) discovery and *in silico* validation of SV events discovered in the analysis of genome sequence datasets being generated at the Centers for Common Disease Genomics and the Centers for Mendelian Genomics.

Ankit Malhotra, Ph.D., Co-Investigator (3.0 calendar months year 1, 2.4 calendar months years 2-3) and Associate Computational Scientist at The Jackson Laboratory, has extensive expertise in genome analysis, especially in the field of structural variations (SV) in the human genome. As a postdoc, he was recipient of a prestigious Department of Defense (DoD) Breast Cancer Research Program (BCRP) fellowship for his work on characterizing structural variations in breast cancer patients. He also lead one of the first meta analysis of complex structural variations in patient datasets from The Cancer Genome Atlas (TCGA) consortium and has developed tools to discover complex SV from whole genome sequencing datasets. He will be involved in novel method development and analysis of sequencing datasets from the different centers producing the genome sequence datasets being generated at the Centers for Common Disease Genomics and the Centers for Mendelian Genomics. Working together with Dr. Lee, he would also be responsible for mentoring the two postdocs being funded for their work on this project.

OTHER PERSONNEL

Support is requested for To-Be-Named project-driven postdoctoral level **Computational Scientist(s) (12 calendar months total per year).** The postdoc(s) will direct effort towards developing the methods for discovery and *in-silico* validation of the whole genome sequencing datasets being generated at the Centers for Common Disease Genomics and the Centers for Mendelian Genomics. They will also be involved in the meta analysis of population structure and recombination rates using the identified SVs.

FRINGE BENEFITS

Fringe benefit costs have been calculated at 30% of salary in accordance with the Rate agreement negotiated between The Jackson Laboratory and DHHS at the time of proposal submission.

TRAVEL

$3,000 per/year is requested to accommodate one trip for each of the key investigators at JAX-GM to attend and participate in an in-person meeting with NIH U01 TOPMed Project representatives. The estimated cost of $750 for each key individuals includes coach airfare ($250) to Washington, one night lodging ($250) and two days per diem expenses and ground transportation ($250). In addition, the remaining $1,500 is requested to support the costs of two researchers to attend and participate in an annual scientific meeting.

OTHER EXPENSES

Materials and Supplies

Lab supplies Funds are requested for office supplies, disposables, and other materials needed for the in house computing work taking place at Jax-GM.

SUBAWARDS/CONSORTIUM/CONTRACTUAL COSTS

A letter of intent to enter into a formal written consortium agreement for meeting the scientific, administrative, financial, and reporting requirements of the proposed project has been executed with both Yale University and the McDonnell Genome Institute at Washington University in St. Louis. The proposed total costs for Yale are $663,248 and the proposed total costs for the McDonnell Genome Institute is $608,475.

FACILITIES AND ADMINISTRATIVE COSTS

The Facilities and Administrative costs have been calculated at 89% of MTDC for the Farmington, CT campus in accordance with the Rate Agreement negotiated between The Jackson Laboratory and DHHS at the time of proposal submission.

R&R SUBAWARD BUDGET ATTACHMENT(S) FORM

Instructions: On this form, you will attach the R&R Subaward Budget files for your grant application. Complete the subawardee budget(s) in accordance with the R&R budget instructions. Please remember that any files you attach must be a PDF document.

Important: Please attach your subawardee budget file(s) with the file name of the subawardee organization. Each file name must be unique.

1) Please attach Attachment 1	Yale0000227912
2) Please attach Attachment 2	WashU0000227913

RESEARCH & RELATED BUDGET - SECTION A & B, BUDGET PERIOD 1

*** ORGANIZATIONAL DUNS:** 0432095620000

*** Budget Type:** ❍ Project ● Subaward/Consortium

Enter name of Organization: Yale University

*** Start Date:** 04-01-2017 *** End Date:** 03-31-2018 **Budget Period: 1**

A. Senior/Key Person

	Prefix	* First Name	Middle Name	* Last Name	Suffix	* Project Role	Base Salary ($)	Calendar Months	Academic Months	Summer Months	* Requested Salary ($)	* Fringe Benefits ($)	* Funds Requested ($)
1.		Mark		Gerstein	PhD	PI	138,825.00			0.45	6,941.25	2,158.73	9,099.98

Total Funds Requested for all Senior Key Persons in the attached file

Additional Senior Key Persons: File Name: **Total Senior/Key Person** **9,099.98**

Mime Type:

B. Other Personnel

* Number of Personnel	* Project Role	Cal. Months	Acad. Months	Sum. Months	* Requested Salary ($)	* Fringe Benefits	* Funds Requested ($)
1	Post Doctoral Associates	7.00			27,416.67	8,526.58	35,943.25
	Graduate Students						
	Undergraduate Students						
	Secretarial/Clerical						
1	Associate Research Scientist	12.00			55,000.00	17,105.00	72,105.00
2	**Total Number Other Personnel**					**Total Other Personnel**	**108,048.25**
						Total Salary, Wages and Fringe Benefits (A+B)	**117,148.23**

RESEARCH & RELATED Budget {A-B} (Funds Requested)

RESEARCH & RELATED BUDGET - SECTION C, D, & E, BUDGET PERIOD 1

*** ORGANIZATIONAL DUNS:** 0432095620000

*** Budget Type:** ○ Project ● Subaward/Consortium

Enter name of Organization: Yale University

*** Start Date:** 04-01-2017 *** End Date:** 03-31-2018 **Budget Period: 1**

C. Equipment Description

List items and dollar amount for each item exceeding $5,000

Equipment Item	*** Funds Requested ($)**
1. Dell	10,000.00
Total funds requested for all equipment listed in the attached file	
Total Equipment	**10,000.00**

Additional Equipment:

File Name: Mime Type:

D. Travel	**Funds Requested ($)**
1. Domestic Travel Costs (Incl. Canada, Mexico, and U.S. Possessions)	2,000.00
2. Foreign Travel Costs	
Total Travel Cost	**2,000.00**

E. Participant/Trainee Support Costs	**Funds Requested ($)**
1. Tuition/Fees/Health Insurance	
2. Stipends	
3. Travel	
4. Subsistence	
5. Other:	

Number of Participants/Trainees **Total Participant Trainee Support Costs**

RESEARCH & RELATED Budget {C-E} (Funds Requested)

RESEARCH & RELATED BUDGET - SECTIONS F-K, BUDGET PERIOD 1

*** ORGANIZATIONAL DUNS:** 0432095620000

*** Budget Type:** ○ Project ● Subaward/Consortium

Enter name of Organization: Yale University

*** Start Date:** 04-01-2017 *** End Date:** 03-31-2018 **Budget Period: 1**

F. Other Direct Costs	**Funds Requested ($)**
1. Materials and Supplies	4,184.77
2. Publication Costs	
3. Consultant Services	
4. ADP/Computer Services	
5. Subawards/Consortium/Contractual Costs	
6. Equipment or Facility Rental/User Fees	
7. Alterations and Renovations	
Total Other Direct Costs	**4,184.77**

G. Direct Costs	**Funds Requested ($)**
Total Direct Costs (A thru F)	**133,333.00**

H. Indirect Costs

Indirect Cost Type	**Indirect Cost Rate (%)**	**Indirect Cost Base ($)**	*** Funds Requested ($)**
1. Modified Total Direct Cost	67.50	123,333.00	83,249.77
		Total Indirect Costs	**83,249.77**

Cognizant Federal Agency DHHS, Ryan McCarthy 212-264-2069

(Agency Name, POC Name, and POC Phone Number)

I. Total Direct and Indirect Costs	**Funds Requested ($)**
Total Direct and Indirect Institutional Costs (G + H)	**216,582.77**

J. Fee	**Funds Requested ($)**

K. * Budget Justification File Name: BudgetJustification_TOPMED_MG.pdf Mime Type: application/pdf

(Only attach one file.)

RESEARCH & RELATED BUDGET - SECTION A & B, BUDGET PERIOD 2

*** ORGANIZATIONAL DUNS:** 0432095620000

*** Budget Type:** ○ Project ● Subaward/Consortium

Enter name of Organization: Yale University

*** Start Date:** 04-01-2018 *** End Date:** 03-31-2019 **Budget Period: 2**

A. Senior/Key Person

	Prefix	* First Name	Middle Name	* Last Name	Suffix	* Project Role	Base Salary ($)	Calendar Months	Academic Months	Summer Months	* Requested Salary ($)	* Fringe Benefits ($)	* Funds Requested ($)
1.		Mark		Gerstein	PhD	PI	142,989.75			0.45	7,149.49	2,223.49	9,372.98

Total Funds Requested for all Senior Key Persons in the attached file

Additional Senior Key Persons: File Name: **Total Senior/Key Person** **9,372.98**

Mime Type:

B. Other Personnel

* Number of Personnel	* Project Role	Cal. Months	Acad. Months	Sum. Months	* Requested Salary ($)	* Fringe Benefits	* Funds Requested ($)
1	Post Doctoral Associates	8.00			32,273.33	10,037.01	42,310.34
	Graduate Students						
	Undergraduate Students						
	Secretarial/Clerical						
1	Associate Research Scientist	12.00			56,650.00	17,618.15	74,268.15
2	**Total Number Other Personnel**					**Total Other Personnel**	**116,578.49**
						Total Salary, Wages and Fringe Benefits (A+B)	**125,951.47**

RESEARCH & RELATED Budget {A-B} (Funds Requested)

RESEARCH & RELATED BUDGET - SECTION C, D, & E, BUDGET PERIOD 2

*** ORGANIZATIONAL DUNS:** 0432095620000

*** Budget Type:** ○ Project ● Subaward/Consortium

Enter name of Organization: Yale University

*** Start Date:** 04-01-2018 *** End Date:** 03-31-2019 **Budget Period: 2**

C. Equipment Description

List items and dollar amount for each item exceeding $5,000

Equipment Item	* Funds Requested ($)
Total funds requested for all equipment listed in the attached file	
Total Equipment	

Additional Equipment:

File Name: Mime Type:

D. Travel	Funds Requested ($)
1. Domestic Travel Costs (Incl. Canada, Mexico, and U.S. Possessions)	2,000.00
2. Foreign Travel Costs	
Total Travel Cost	**2,000.00**

E. Participant/Trainee Support Costs	Funds Requested ($)
1. Tuition/Fees/Health Insurance	
2. Stipends	
3. Travel	
4. Subsistence	
5. Other:	
Number of Participants/Trainees **Total Participant Trainee Support Costs**	

RESEARCH & RELATED Budget {C-E} (Funds Requested)

RESEARCH & RELATED BUDGET - SECTIONS F-K, BUDGET PERIOD 2

*** ORGANIZATIONAL DUNS:** 0432095620000

*** Budget Type:** ○ Project ● Subaward/Consortium

Enter name of Organization: Yale University

*** Start Date:** 04-01-2018 *** End Date:** 03-31-2019 **Budget Period: 2**

F. Other Direct Costs	**Funds Requested ($)**
1. Materials and Supplies	5,381.53
2. Publication Costs	
3. Consultant Services	
4. ADP/Computer Services	
5. Subawards/Consortium/Contractual Costs	
6. Equipment or Facility Rental/User Fees	
7. Alterations and Renovations	
Total Other Direct Costs	**5,381.53**

G. Direct Costs	**Funds Requested ($)**
Total Direct Costs (A thru F)	**133,333.00**

H. Indirect Costs

Indirect Cost Type	**Indirect Cost Rate (%)**	**Indirect Cost Base ($)**	*** Funds Requested ($)**
1. Modified Total Direct Cost	67.50	133,333.00	89,999.77
		Total Indirect Costs	**89,999.77**

Cognizant Federal Agency DHHS, Ryan McCarthy 212-264-2069

(Agency Name, POC Name, and POC Phone Number)

I. Total Direct and Indirect Costs	**Funds Requested ($)**
Total Direct and Indirect Institutional Costs (G + H)	**223,332.77**

J. Fee	**Funds Requested ($)**

K. * Budget Justification File Name: BudgetJustification_TOPMED_MG.pdf Mime Type: application/pdf

(Only attach one file.)

RESEARCH & RELATED Budget {F-K} (Funds Requested)

RESEARCH & RELATED BUDGET - SECTION A & B, BUDGET PERIOD 3

*** ORGANIZATIONAL DUNS:** 0432095620000

*** Budget Type:** ○ Project ● Subaward/Consortium

Enter name of Organization: Yale University

*** Start Date:** 04-01-2019 *** End Date:** 03-31-2020 **Budget Period: 3**

A. Senior/Key Person

	Prefix	* First Name	Middle Name	* Last Name	Suffix	* Project	Base	Calendar	Academic	Summer	* Requested	* Fringe	* Funds
						Role	Salary ($)	Months	Months	Months	Salary ($)	Benefits ($)	Requested ($)
1.		Mark		Gerstein	PhD	PI	147,279.44			0.45	7,363.97	2,290.20	9,654.17

Total Funds Requested for all Senior Key Persons in the attached file

Additional Senior Key Persons: File Name: **Total Senior/Key Person** **9,654.17**

Mime Type:

B. Other Personnel

* Number of	* Project Role	Cal. Months	Acad. Months	Sum. Months	* Requested	* Fringe Benefits	* Funds Requested
Personnel					Salary ($)		($)
1	Post Doctoral Associates	7.00			29,086.34	9,045.85	38,132.19
	Graduate Students						
	Undergraduate Students						
	Secretarial/Clerical						
1	Associate Research Scientist	12.00			58,349.50	18,146.69	76,496.19
2	**Total Number Other Personnel**					**Total Other Personnel**	**114,628.38**
						Total Salary, Wages and Fringe Benefits (A+B)	**124,282.55**

RESEARCH & RELATED Budget {A-B} (Funds Requested)

RESEARCH & RELATED BUDGET - SECTION C, D, & E, BUDGET PERIOD 3

*** ORGANIZATIONAL DUNS:** 0432095620000

*** Budget Type:** ○ Project ● Subaward/Consortium

Enter name of Organization: Yale University

*** Start Date:** 04-01-2019 *** End Date:** 03-31-2020 **Budget Period: 3**

C. Equipment Description

List items and dollar amount for each item exceeding $5,000

Equipment Item	*** Funds Requested ($)**
Total funds requested for all equipment listed in the attached file	
Total Equipment	

Additional Equipment:

File Name: Mime Type:

D. Travel	**Funds Requested ($)**
1. Domestic Travel Costs (Incl. Canada, Mexico, and U.S. Possessions)	2,000.00
2. Foreign Travel Costs	
Total Travel Cost	**2,000.00**

E. Participant/Trainee Support Costs	**Funds Requested ($)**
1. Tuition/Fees/Health Insurance	
2. Stipends	
3. Travel	
4. Subsistence	
5. Other:	
Number of Participants/Trainees **Total Participant Trainee Support Costs**	

RESEARCH & RELATED BUDGET - SECTIONS F-K, BUDGET PERIOD 3

*** ORGANIZATIONAL DUNS:** 0432095620000

*** Budget Type:** ○ Project ● Subaward/Consortium

Enter name of Organization: Yale University

*** Start Date:** 04-01-2019 *** End Date:** 03-31-2020 **Budget Period: 3**

F. Other Direct Costs	**Funds Requested ($)**
1. Materials and Supplies	7,050.45
2. Publication Costs	
3. Consultant Services	
4. ADP/Computer Services	
5. Subawards/Consortium/Contractual Costs	
6. Equipment or Facility Rental/User Fees	
7. Alterations and Renovations	
Total Other Direct Costs	**7,050.45**

G. Direct Costs	**Funds Requested ($)**
Total Direct Costs (A thru F)	**133,333.00**

H. Indirect Costs

Indirect Cost Type	**Indirect Cost Rate (%)**	**Indirect Cost Base ($)**	*** Funds Requested ($)**
1. Modified Total Direct Cost	67.50	133,333.00	89,999.77
		Total Indirect Costs	**89,999.77**

Cognizant Federal Agency DHHS, Ryan McCarthy 212-264-2069

(Agency Name, POC Name, and POC Phone Number)

I. Total Direct and Indirect Costs	**Funds Requested ($)**
Total Direct and Indirect Institutional Costs (G + H)	**223,332.77**

J. Fee	**Funds Requested ($)**

K. * Budget Justification File Name: BudgetJustification_TOPMED_MG.pdf Mime Type: application/pdf

(Only attach one file.)

RESEARCH & RELATED Budget {F-K} (Funds Requested)

RESEARCH & RELATED BUDGET - Cumulative Budget

	Totals ($)	
Section A, Senior/Key Person		**28,127.13**
Section B, Other Personnel		**339,255.12**
Total Number Other Personnel	6	
Total Salary, Wages and Fringe Benefits (A+B)		**367,382.25**
Section C, Equipment		**10,000.00**
Section D, Travel		**6,000.00**
1. Domestic	6,000.00	
2. Foreign		
Section E, Participant/Trainee Support Costs		
1. Tuition/Fees/Health Insurance		
2. Stipends		
3. Travel		
4. Subsistence		
5. Other		
6. Number of Participants/Trainees		
Section F, Other Direct Costs		**16,616.75**
1. Materials and Supplies	16,616.75	
2. Publication Costs		
3. Consultant Services		
4. ADP/Computer Services		
5. Subawards/Consortium/Contractual Costs		
6. Equipment or Facility Rental/User Fees		
7. Alterations and Renovations		
8. Other 1		
9. Other 2		
10. Other 3		
Section G, Direct Costs (A thru F)		**399,999.00**
Section H, Indirect Costs		**263,249.31**
Section I, Total Direct and Indirect Costs (G + H)		**663,248.31**
Section J, Fee		

RESEARCH & RELATED BUDGET - SECTION A & B, BUDGET PERIOD 1

*** ORGANIZATIONAL DUNS:** 0685522070000

*** Budget Type:** ○ Project ● Subaward/Consortium

Enter name of Organization: Washington University

*** Start Date:** 04-01-2017 *** End Date:** 03-31-2018 **Budget Period: 1**

A. Senior/Key Person

	Prefix	* First Name	Middle Name	* Last Name	Suffix	* Project Role	Base Salary ($)	Calendar Months	Academic Months	Summer Months	* Requested Salary ($)	* Fringe Benefits ($)	* Funds Requested ($)
1.		Li		Ding		PD/PI		1.80			27,765.00	5,096.00	32,861.00
2.		Wendl		Michael		Consortium Co-Investigator		1.20			8,761.00	2,638.00	11,399.00

Total Funds Requested for all Senior Key Persons in the attached file

Additional Senior Key Persons: File Name:

Mime Type:

Total Senior/Key Person **44,260.00**

B. Other Personnel

* Number of Personnel	* Project Role	Cal. Months	Acad. Months	Sum. Months	* Requested Salary ($)	* Fringe Benefits	* Funds Requested ($)
1	Post Doctoral Associates	2.40			12,675.00	3,778.00	16,453.00
	Graduate Students						
	Undergraduate Students						
	Secretarial/Clerical						
1	Analyst	9.60			37,134.00	10,168.00	47,302.00
2	**Total Number Other Personnel**					**Total Other Personnel**	**63,755.00**
						Total Salary, Wages and Fringe Benefits (A+B)	**108,015.00**

RESEARCH & RELATED Budget {A-B} (Funds Requested)

RESEARCH & RELATED BUDGET - SECTION C, D, & E, BUDGET PERIOD 1

*** ORGANIZATIONAL DUNS:** 0685522070000

*** Budget Type:** ○ Project ● Subaward/Consortium

Enter name of Organization: Washington University

*** Start Date:** 04-01-2017 *** End Date:** 03-31-2018 **Budget Period: 1**

C. Equipment Description

List items and dollar amount for each item exceeding $5,000

Equipment Item	*** Funds Requested ($)**
Total funds requested for all equipment listed in the attached file	
Total Equipment	

Additional Equipment:

File Name: Mime Type:

D. Travel	**Funds Requested ($)**
1. Domestic Travel Costs (Incl. Canada, Mexico, and U.S. Possessions)	3,000.00
2. Foreign Travel Costs	
Total Travel Cost	**3,000.00**

E. Participant/Trainee Support Costs	**Funds Requested ($)**
1. Tuition/Fees/Health Insurance	
2. Stipends	
3. Travel	
4. Subsistence	
5. Other:	
Number of Participants/Trainees **Total Participant Trainee Support Costs**	

RESEARCH & RELATED Budget {C-E} (Funds Requested)

RESEARCH & RELATED BUDGET - SECTIONS F-K, BUDGET PERIOD 1

*** ORGANIZATIONAL DUNS:** 0685522070000

*** Budget Type:** ○ Project ● Subaward/Consortium

Enter name of Organization: Washington University

*** Start Date:** 04-01-2017 *** End Date:** 03-31-2018 **Budget Period: 1**

F. Other Direct Costs	**Funds Requested ($)**
1. Materials and Supplies	6,000.00
2. Publication Costs	3,000.00
3. Consultant Services	
4. ADP/Computer Services	
5. Subawards/Consortium/Contractual Costs	
6. Equipment or Facility Rental/User Fees	
7. Alterations and Renovations	
8. Computing	12,985.00
Total Other Direct Costs	**21,985.00**

G. Direct Costs	**Funds Requested ($)**
Total Direct Costs (A thru F)	**133,000.00**

H. Indirect Costs

Indirect Cost Type	**Indirect Cost Rate (%)**	**Indirect Cost Base ($)**	*** Funds Requested ($)**
1. MTDC	52.50	133,000.00	69,825.00
		Total Indirect Costs	**69,825.00**

Cognizant Federal Agency DHHS, Division of Cost Allocation, 1301 Young Street, Dallas, TX, 75202, 214-767-3261

(Agency Name, POC Name, and POC Phone Number)

I. Total Direct and Indirect Costs	**Funds Requested ($)**
Total Direct and Indirect Institutional Costs (G + H)	**202,825.00**

J. Fee	**Funds Requested ($)**

K. * Budget Justification File Name: 8. Budget Justification.pdf Mime Type: application/pdf

(Only attach one file.)

RESEARCH & RELATED BUDGET - SECTION A & B, BUDGET PERIOD 2

*** ORGANIZATIONAL DUNS:** 0685522070000

*** Budget Type:** ○ Project ● Subaward/Consortium

Enter name of Organization: Washington University

*** Start Date:** 04-01-2018 *** End Date:** 03-31-2019 **Budget Period: 2**

A. Senior/Key Person

	Prefix	* First Name	Middle Name	* Last Name	Suffix	* Project Role	Base Salary ($)	Calendar Months	Academic Months	Summer Months	* Requested Salary ($)	* Fringe Benefits ($)	* Funds Requested ($)
1.		Li		Ding		PD/PI		1.80			27,765.00	5,120.00	32,885.00
2.		Wendl		Michael		Consortium Co-Investigator		1.20			9,024.00	2,749.00	11,773.00

Total Funds Requested for all Senior Key Persons in the attached file

Additional Senior Key Persons: File Name: **Total Senior/Key Person** **44,658.00**

Mime Type:

B. Other Personnel

* Number of Personnel	* Project Role	Cal. Months	Acad. Months	Sum. Months	* Requested Salary ($)	* Fringe Benefits	* Funds Requested ($)
1	Post Doctoral Associates	2.40			13,056.00	3,999.00	17,055.00
	Graduate Students						
	Undergraduate Students						
	Secretarial/Clerical						
1	Analyst	9.60			38,250.00	10,726.00	48,976.00
2	**Total Number Other Personnel**					**Total Other Personnel**	**66,031.00**
						Total Salary, Wages and Fringe Benefits (A+B)	**110,689.00**

RESEARCH & RELATED Budget {A-B} (Funds Requested)

RESEARCH & RELATED BUDGET - SECTION C, D, & E, BUDGET PERIOD 2

*** ORGANIZATIONAL DUNS:** 0685522070000

*** Budget Type:** ○ Project ● Subaward/Consortium

Enter name of Organization: Washington University

*** Start Date:** 04-01-2018 *** End Date:** 03-31-2019 **Budget Period: 2**

C. Equipment Description

List items and dollar amount for each item exceeding $5,000

Equipment Item	*** Funds Requested ($)**
Total funds requested for all equipment listed in the attached file	
Total Equipment	

Additional Equipment:

File Name: Mime Type:

D. Travel	**Funds Requested ($)**
1. Domestic Travel Costs (Incl. Canada, Mexico, and U.S. Possessions)	3,000.00
2. Foreign Travel Costs	
Total Travel Cost	**3,000.00**

E. Participant/Trainee Support Costs	**Funds Requested ($)**
1. Tuition/Fees/Health Insurance	
2. Stipends	
3. Travel	
4. Subsistence	
5. Other:	
Number of Participants/Trainees **Total Participant Trainee Support Costs**	

RESEARCH & RELATED Budget {C-E} (Funds Requested)

RESEARCH & RELATED BUDGET - SECTIONS F-K, BUDGET PERIOD 2

*** ORGANIZATIONAL DUNS:** 0685522070000

*** Budget Type:** ○ Project ● Subaward/Consortium

Enter name of Organization: Washington University

*** Start Date:** 04-01-2018 *** End Date:** 03-31-2019 **Budget Period: 2**

F. Other Direct Costs	**Funds Requested ($)**
1. Materials and Supplies	6,000.00
2. Publication Costs	3,311.00
3. Consultant Services	
4. ADP/Computer Services	
5. Subawards/Consortium/Contractual Costs	
6. Equipment or Facility Rental/User Fees	
7. Alterations and Renovations	
8. Computing	10,000.00
Total Other Direct Costs	**19,311.00**

G. Direct Costs	**Funds Requested ($)**
Total Direct Costs (A thru F)	**133,000.00**

H. Indirect Costs

Indirect Cost Type	**Indirect Cost Rate (%)**	**Indirect Cost Base ($)**	*** Funds Requested ($)**
1. MTDC	52.50	133,000.00	69,825.00
		Total Indirect Costs	**69,825.00**

Cognizant Federal Agency DHHS, Division of Cost Allocation, 1301 Young Street, Dallas, TX, 75202, 214-767-3261

(Agency Name, POC Name, and POC Phone Number)

I. Total Direct and Indirect Costs	**Funds Requested ($)**
Total Direct and Indirect Institutional Costs (G + H)	**202,825.00**

J. Fee	**Funds Requested ($)**

K. * Budget Justification File Name: 8. Budget Justification.pdf Mime Type: application/pdf

(Only attach one file.)

RESEARCH & RELATED BUDGET - SECTION A & B, BUDGET PERIOD 3

*** ORGANIZATIONAL DUNS:** 0685522070000

*** Budget Type:** ○ Project ● Subaward/Consortium

Enter name of Organization: Washington University

*** Start Date:** 04-01-2019 *** End Date:** 03-31-2020 **Budget Period: 3**

A. Senior/Key Person

	Prefix	* First Name	Middle Name	* Last Name	Suffix	* Project Role	Base Salary ($)	Calendar Months	Academic Months	Summer Months	* Requested Salary ($)	* Fringe Benefits ($)	* Funds Requested ($)
1.		Li		Ding		PD/PI		1.80			27,765.00	5,150.00	32,915.00
2.		Wendl		Michael		Consortium Co-Investigator		1.20			9,295.00	2,866.00	12,161.00

Total Funds Requested for all Senior Key Persons in the attached file

Additional Senior Key Persons: File Name:

Mime Type:

Total Senior/Key Person **45,076.00**

B. Other Personnel

* Number of Personnel	* Project Role	Cal. Months	Acad. Months	Sum. Months	* Requested Salary ($)	* Fringe Benefits	* Funds Requested ($)
1	Post Doctoral Associates	2.40			13,448.00	4,343.00	17,791.00
	Graduate Students						
	Undergraduate Students						
	Secretarial/Clerical						
1	Analyst	9.60			39,397.00	11,330.00	50,727.00
2	**Total Number Other Personnel**					**Total Other Personnel**	**68,518.00**
						Total Salary, Wages and Fringe Benefits (A+B)	**113,594.00**

RESEARCH & RELATED Budget {A-B} (Funds Requested)

RESEARCH & RELATED BUDGET - SECTION C, D, & E, BUDGET PERIOD 3

*** ORGANIZATIONAL DUNS:** 0685522070000

*** Budget Type:** ○ Project ● Subaward/Consortium

Enter name of Organization: Washington University

*** Start Date:** 04-01-2019 *** End Date:** 03-31-2020 **Budget Period: 3**

C. Equipment Description

List items and dollar amount for each item exceeding $5,000

Equipment Item	*** Funds Requested ($)**
Total funds requested for all equipment listed in the attached file	
Total Equipment	

Additional Equipment:

File Name: Mime Type:

D. Travel	**Funds Requested ($)**
1. Domestic Travel Costs (Incl. Canada, Mexico, and U.S. Possessions)	3,000.00
2. Foreign Travel Costs	
Total Travel Cost	**3,000.00**

E. Participant/Trainee Support Costs	**Funds Requested ($)**
1. Tuition/Fees/Health Insurance	
2. Stipends	
3. Travel	
4. Subsistence	
5. Other:	
Number of Participants/Trainees **Total Participant Trainee Support Costs**	

RESEARCH & RELATED Budget {C-E} (Funds Requested)

RESEARCH & RELATED BUDGET - SECTIONS F-K, BUDGET PERIOD 3

*** ORGANIZATIONAL DUNS:** 0685522070000

*** Budget Type:** ○ Project ● Subaward/Consortium

Enter name of Organization: Washington University

*** Start Date:** 04-01-2019 *** End Date:** 03-31-2020 **Budget Period: 3**

F. Other Direct Costs	**Funds Requested ($)**
1. Materials and Supplies	6,000.00
2. Publication Costs	3,000.00
3. Consultant Services	
4. ADP/Computer Services	
5. Subawards/Consortium/Contractual Costs	
6. Equipment or Facility Rental/User Fees	
7. Alterations and Renovations	
8. Computing	7,406.00
Total Other Direct Costs	**16,406.00**

G. Direct Costs	**Funds Requested ($)**
Total Direct Costs (A thru F)	**133,000.00**

H. Indirect Costs

Indirect Cost Type	**Indirect Cost Rate (%)**	**Indirect Cost Base ($)**	*** Funds Requested ($)**
1. MTDC	52.50	133,000.00	69,825.00
		Total Indirect Costs	**69,825.00**

Cognizant Federal Agency DHHS, Division of Cost Allocation, 1301 Young Street, Dallas, TX, 75202, 214-767-3261

(Agency Name, POC Name, and POC Phone Number)

I. Total Direct and Indirect Costs	**Funds Requested ($)**
Total Direct and Indirect Institutional Costs (G + H)	**202,825.00**

J. Fee	**Funds Requested ($)**

K. * Budget Justification File Name: 8. Budget Justification.pdf Mime Type: application/pdf

(Only attach one file.)

RESEARCH & RELATED Budget {F-K} (Funds Requested)

RESEARCH & RELATED BUDGET - Cumulative Budget

	Totals ($)	
Section A, Senior/Key Person		**133,994.00**
Section B, Other Personnel		**198,304.00**
Total Number Other Personnel	6	
Total Salary, Wages and Fringe Benefits (A+B)		**332,298.00**
Section C, Equipment		
Section D, Travel		**9,000.00**
1. Domestic	9,000.00	
2. Foreign		
Section E, Participant/Trainee Support Costs		
1. Tuition/Fees/Health Insurance		
2. Stipends		
3. Travel		
4. Subsistence		
5. Other		
6. Number of Participants/Trainees		
Section F, Other Direct Costs		**57,702.00**
1. Materials and Supplies	18,000.00	
2. Publication Costs	9,311.00	
3. Consultant Services		
4. ADP/Computer Services		
5. Subawards/Consortium/Contractual Costs		
6. Equipment or Facility Rental/User Fees		
7. Alterations and Renovations		
8. Other 1	30,391.00	
9. Other 2		
10. Other 3		
Section G, Direct Costs (A thru F)		**399,000.00**
Section H, Indirect Costs		**209,475.00**
Section I, Total Direct and Indirect Costs (G + H)		**608,475.00**
Section J, Fee		

Budget Justification

PERSONNEL:

Mark Gerstein, Ph.D. PI (.45 summer months). Dr. Gerstein is the Albert Williams Professor of Biomedical Informatics. His lab (http://gersteinlab.org) was one of the first to perform integrated data mining on functional genomics data and to do genome-wide surveys. His tools for analyzing motions and packing are widely used. Most recently, he has designed and developed a wide array of databases and computational tools to mine genome data in humans, as well as in many other organisms. He has worked extensively in the 1000 genomes project in the SV and FIG groups. He also worked in the ENCODE pilot project and currently works extensively in the ENCODE and modENCODE production projects. He is also a co-PI in DOE KBase and the leader of the Data Analysis Center for the NIH exRNA consortium. In these roles Dr. Gerstein has designed and developed a wide array of databases and computational tools to mine genomic data in humans as well as in many other organisms. He will be directly supervising Dr. Harmanci and Dr. Navarro on this project.

Dr. Arif Harmanci, Ph.D., Assoc. Research Scientist (12 calendar months). Dr. Harmanci has extensive experience with bioinformatic approaches to genome-wide analysis and a strong background in scientific computation. As part of his PhD thesis, he developed advanced methods for RNA secondary structure prediction. In the Gerstein laboratory, he has developed new algorithms to identify transcription factor binding peaks from ChIP-Seq data. He is currently working on transcriptome analysis of several RNA-Seq datasets that include the Geuvadis dataset (RNA-Seq on 500 individuals). He will work on the analysis proposed in the grant under the direction of Dr Gerstein. He will supervise the involved laboratories during the integration of structural variation detection methods and also integrate the of RNA-seq dataset into the functional prioritization of SVs.

Fábio Navarro, Ph.D., Postdoctoral Assoc. (7 calendar months years 1 and 3, 8 calendar months year 2). Dr. Navarro has a strong background in scientific computation and biochemistry. He graduated from Universidade Federal de São Carlos in Computer Engineering and obtained his Ph.D. at the Biochemistry program from the Universidade de São Paulo. His background is development of software solutions to biological problems involving analysis and mining of genomic and trascriptomic data. He used to work with the description of genetic variation in human populations and human diseases such as cancer and Xeroderma Pigmentosa. He is now working with the expression profile of repetitive elements in the healthy tissues and investigating the functional impact of point mutation and structural variation in the human genome. Dr. Navarro will integrate and further develop tools to detect structural variations based on whole genome sequencing data and will extend the methods to prioritize SVs based on protein coding annotation, non-coding regulatory regions and eQTL analysis.

Fringe Benefits:

Fringe benefits are calculated at 31.1% for the PI, ARS and PDA.

EQUIPMENT:

In year 1, we need a Dell Poweredge R815 server with 160GB of memory and four AMD Opteron processors. This will be used for processing our data and digital visualization. This is needed to complete the proposed research and will solely benefit this project.

TRAVEL:

We are budgeting incremental funds for airfare, lodging and meal expenses to attend a scientific meeting annually that benefit the project.

SUPPLIES:

We are budgeting an incremental amount of supplies for the individuals named above. This supplies budget will be used to cover computer supplies for them, covering such expenses as: diskettes, tapes, and other miscellaneous computer parts (e.g. replacing worn out surge suppressors), software upgrades, web hosting and "cloud computing" fees, and reprint charges. These items are needed to complete the proposed research and will solely benefit this project.

Indirect costs are calculated at Yale’s federally negotiated rate of 67.5% of modified total direct costs. DHHS agreement dated 02/23/2016.

BUDGET JUSTIFICATION – Washington University

Senior Personnel:

Li Ding, Ph.D. (PD/PI; 1.80 calendar, 15% effort):

Dr. Ding is an Associate Professor of Medicine and Genetics at Washington University School of Medicine, Director of Computational Biology in Oncology, and Assistant Director of The McDonnell Genome Institute. She has a unique combination of in-depth understanding of software development, integrated data analysis, and biological science. Her research team has developed a collection of computational tools, including VarScan, SomaticSniper, SciClone, BreakDancer, BreakFusion, MSIsensor, Pindel-C, GenomeVIP, HotSpot3D, PathScan, and MuSiC, all widely used by the research community for analyzing high-throughput sequencing data. In particular, her team established a robust analysis pipeline that has been used for 1000 Genomes Project (including 1000G SV project), the Cancer Genome Atlas (TCGA) project, Clinical Proteomic Tumor Analysis Consortium (CPTAC), and other medical genomics projects. Dr. Ding plays significant roles in TCGA and ICGC, co-chairing TCGA PanCanAtlas Oncogenic Process group, Sarcoma AWG, and ICGC mutation calling group (PCAWG-1). Dr. Ding also serves on the Steering Committee for NCI Genomic Data Commons (GDC). Dr. Ding has successfully led many large-scale, multi-institute studies on the genomics of lung adenocarcinomas, AML, and breast cancer. Building on this foundation, her lab has produced a series of seminal publications in the fields of cancer genomics research and cancer biology, including 1) the discovery of 127 cancer genes across over 3,000 tumors from 12 major cancer types; 2) the report of 13 significant germline susceptibility genes in over 4,000 cancer cases; 2) the identification of pre-existing mutations in 19 leukemia and/or lymphoma-associated genes in people without overt hematological malignancies; and 4) the development of advanced computational tools for detecting druggable complex indels often missed by the existing approaches in cancer patients; and 5) the discovery of druggable targets validated experimentally using modern global proteomics approaches. Dr. Ding will be responsible for providing leadership on computational tool and pipeline development as well as association studies for this application.

Michael Wendl, Ph.D. (Co-Investigator; 1.20 calendar, 10% effort):

Dr. Michael Wendl is a mathematical biologist whose expertise is in statistical data analysis, statistical genomics and pathway/network analysis, project design and optimization, databases, and algorithms. Dr. Wendl will work with Dr. Ding to direct the development of association pipeline and association analysis in Aim 3.

Other Personnel:

Adam Scott, Ph.D. (Analyst; 9.60 calendar, 80% effort):

Dr. Adam Scott is a computational scientist and physicist whose expertise is in clustering and data modeling. Dr. Scott will be responsible for developing burden analysis approaches by integrating impact scores.

Matthew Wyczalkowski, Ph.D. (Research Associate; 2.40 calendar, 20% effort):

Dr. Wyczalkowski is a computational scientist whose expertise is in software development and data analysis. Dr. Wyczalkowski will be responsible for association analysis.

Travel: ($9,000)

We are requesting funding of $3,000 per year for 2 project members to participate in annual scientific conferences and meetings.

Consumable Supplies:

Computer Supplies: ($18,000)

We are requesting fund of $6000 per year for purchasing 1 laptop and 1 workstation.

Other Expenses:

Computing: ($30,391)

We are requesting $12,985 in year 1, $10,000 in year 2, and $7,406 in year 3 for cloud computing and cloud pipeline development.

Publication: ($9,311)
We plan to publish 2 papers a year with the cost of $3000 in year 1, $3,311 in year 2, and $3,000 in year 3.

PHS 398 Cover Page Supplement

OMB Number: 0925-0001

Expiration Date: 10/31/2018

1. Human Subjects Section

Clinical Trial? ☑ No □ Yes

*Agency-Defined Phase III Clinical Trial?

2. Vertebrate Animals Section

Are vertebrate animals euthanized?

If "Yes" to euthanasia

Is method consistent with American Veterinary Medical Association (AVMA) guidelines?

If "No" to AVMA guidelines, describe method and provide scientific justification

3. *Program Income

*Is program income anticipated during the periods for which the grant support is requested?

□ Yes ☑ No

If you checked "yes" above (indicating that program income is anticipated), then use the format below to reflect the amount and source(s). Otherwise, leave this section blank.

*Budget Period	*Anticipated Amount($)	*Source(s)

PHS 398 Cover Page Supplement

OMB Number: 0925-0001

Expiration Date: 10/31/2018

4. Human Embryonic Stem Cells Section

*Does the proposed project involve human embryonic stem cells? ☑ No □ Yes

If the proposed project involves human embryonic stem cells, list below the registration number of the specific cell line(s) from the following list:http://stemcells.nih.gov/research/registry/. Or, if a specific stem cell line cannot be referenced at this time, please check the box indicating that one from the registry will be used:

Cell Line(s): Specific stem cell line cannot be referenced at this time. One from the registry will be used.

5. Inventions and Patents Section (RENEWAL)

*Inventions and Patents:

If the answer is "Yes" then please answer the following:

*Previously Reported:

PHS 398 Cover Page Supplement

OMB Number: 0925-0001

Expiration Date: 10/31/2018

7. Change of Investigator / Change of Institution Section

□ Change of principal investigator / program director

Name of former principal investigator / program director:

Prefix:

*First Name:

Middle Name:

*Last Name

Suffix:

□ Change of Grantee Institution

*Name of former institution:

PHS 398 Research Plan

OMB Number: 0925-0001
Expiration Date: 10/31/2018

Introduction

1. Introduction to Application
(Resubmission and Revision)

Research Plan Section

2. Specific Aims	
3. Research Strategy*	M-7_PHS_ResearchPlan_ResearchStrategy.docx
4. Progress Report Publication List	

Human Subjects Section

5. Protection of Human Subjects

6. Data Safety Monitoring Plan

7. Inclusion of Women and Minorities

8. Inclusion of Children

Other Research Plan Section

9. Vertebrate Animals	
10. Select Agent Research	
11. Multiple PD/PI Leadership Plan	M-14_PHS_ResearchPlan_MultiplePILeadershipPlan.pdf
12. Consortium/Contractual Arrangements	M-11_PHS_ResearchPlan_ConsortiumContractualArrangements.pdf
13. Letters of Support	M-10_PHS_ResearchPlan_LettersOfSupport.pdf
14. Resource Sharing Plan(s)	M-9_PHS_ResearchPlan_ResourceSharingPlans.pdf
15. Authentication of Key Biological and/or Chemical Resources	

Appendix

16. Appendix

CONSORTIUM/CONTRACTUAL ARRANGEMENTS

Consortium arrangements are proposed with Yale University, New Haven, CT, Dr. Mark Gerstein, PI, and with The Washington University, St. Louis, MO, Dr. Li Ding, PI.

At Yale University, Dr. Gerstein's role on the project is to develop tools to examine the functional impact of the identified SVs and to develop a novel pipeline of methodologies for functional annotation of variants and characterization of associated biological processes. Dr. Gerstein will communicate regularly with the PIs, about the aims, results and analyses of this project, and will attend several formal meetings throughout the year to present findings to the other members of the research team.

At Washington University in St. Louis, Dr. Ding's role on the project will be to lead the efforts to develop the SV detection tool and cloud pipeline tailored to finding adequately powered SVs. Dr. Ding will communicate regularly with the PIs, about the aims, results and analyses of this project, and will attend several formal meetings throughout the year to present findings to the other members of the research team.

The appropriate program and administrative personnel of each organization involved in this grant application for this project are aware of the NIH consortium and cooperative agreement grant policies and are prepared to establish the necessary inter-organizational agreements consistent with those policies.

RESOURCE SHARING PLANS

Research tools and resources will be made available in full accordance with the NIH Grants Policy Statement, the Principles and Guidelines for Recipients of NIH Research Grants and Contracts, and the NIH Genomic Data Sharing Policy. Each of the Principal Investigators has significant experience participating in large-scale genome sequencing projects, and has an established record of sharing unique resources generated with NIH support with the academic research community, through publications and submission to data repositories without any license or research restrictions.

Genomic Data Sharing Plan

All genomic data (large-scale data as defined in the NIH Genomic Data Sharing Policy) resulting from the proposed project, as well as relevant phenotype-associated data, will be made available in a timely manner in full accordance with the NIH Final Genomic Data Sharing (GDS) Policy as well as any distributions plans established by the TOPMed Program. In particular, the catalog of structural variations identified from the individuals sequenced as part of the NHLBI-supported TOPMed Program by project investigators will be made publicly available within the timelines required by the GDS Policy and Supplemental Information. Project investigators that download unrestricted-access data from NIH-designated data repositories will acknowledge the specific data sets or applicable accession numbers(s) and the NIH-designated data repository. It is anticipated that genomic data made available from the TOPMed Program will be de-identified, and therefore, the derived data will meet the standards set forth in the HHS Regulations for the Protection of Human Subjects. In addition, this project will be registered in the Database of Genotypes and Phenotypes (dbGaP) http://www.ncbi.nih.gov/gap and the human genomic data will be submitted to the relevant NIH-designated public data archives advised by the steering committee of the TOPMed program.

Software Sharing Plan

This proposal describes three major software pipelines for SV analysis: an extended fusorSV framework for discovery, SVIM for functional annotation and impact assessment, and SV2Pheno for determining genotype-phenotype associations. As this proposal encompasses substantial work in several different areas, we plan on semi-regular code releases over the duration of the funding period (see timeline below). The release process will follow conventional practice: source code and documentation (including description, installation, use cases, etc.) will be packaged and available from a public distribution site, e.g. GitHub. Software images for use on cloud computing resources will also be disseminated to the broader community as appropriate. Each pipeline will be accompanied by journal publication describing the new enabled capabilities along with any new biological findings.

1. Software Licensing Plan. We will release the pipeline software developed as part of this grant under the Apache License, Version 2.0 (Apache-2.0). This license is a permissive free, open-source license ensuring the software is freely available to everyone, including biomedical researchers and educators in the non-profit sector, such as institutions of education, research institutions, and government laboratories. This license also allows everyone, including external researchers, collaborators, and students to modify the source code and to share these modifications with the broader community. This attribute ensures that code maintenance and development can continue under the auspices of others in the community in the very unlikely event that we are unwilling or unable to do so. Derivative works are not required also to be distributed under Apache-2.0, allowing dissemination and commercialization of enhanced or customized versions of the software and the incorporation of the software or pieces of it into other software packages.

2. Software Release Strategy. We will release official versions of the pipelines with major improvements on a semi-regular basis and minor updates as needed to ensure properly functioning software. All releases will be numbered and made available on GitHub. Improvements in the external software tools on which our pipelines depend will be reviewed for compatibility with our software, and we will issue numbered software releases as appropriate. We will also work to ensure new releases maintain compatibility across supported local data center and cloud platforms. Given typical development cycles, we estimate major releases perhaps a few times a year, minor releases on the cycle of months, and ad hoc patches on the cycle of weeks. Finally, we will publish companion scientific articles and application notes when appropriate. Announcements of updates will be made through the regular communication channels described in the proposal, e.g. Biostars community message board. We believe this overall strategy will be highly responsive to the needs of the user community.

The Jackson Laboratory

Leading the search for tomorrow's cures

Edison T. Liu, M.D.
President and Chief Executive Officer
207.288.6041 *t* | 207.288.6044 *f* | edison.liu@jax.org

June 28, 2016

Weiniu Gan, Ph.D.
Division of Lung Sciences
National Heart, Lung, and Blood Institute
National Institutes of Health

Re: NHLBI TOPMed Program: Integrative Omics Approaches for Analysis of TOPMed Data U01 (RFA-HL-17-011)(Charles Lee, Ph.D., PI)—Letter of Institutional Support

Dear Dr. Gan and Members of the Review Committee:

It is with great pleasure that I confirm our institution's highest possible support for Dr. Lee in the present application and our commitment to this TOPMed Program project. This project will meet a growing need for analytical tools, methodologies and expertise towards the discovery of meaningful associations between genome sequence and common human diseases. This mission aligns perfectly with JAX's overall goal to discover the precise genomic causes of disease and to empower the global biomedical community through the provision of unique genetic and genomic resources. JAX supports these efforts through significant institutional investments designed to ensure its position as a globally renowned institution for genetics and genomics research spanning the basic, translational and clinical research spectrum.

JAX Genomic Medicine (JAX-GM) is ideally poised to support the objectives and goals of this project. JAX-GM uses a human-centered, computationally oriented genomics approach towards research and discovery into the complex mechanisms of human disease. It is home to some of the world's leading scientists in bioinformatics, genetics, clinical genomics and computational biology. JAX-GM is rapidly hiring faculty and we plan for one-third of the faculty at full capacity to be computationally focused. This is in addition to the rich intellectual expertise already present at JAX through our Computational Sciences team of highly trained bioinformaticians and computational biologists, and our Information Technology staff who provide round-the-clock support to all computational projects. JAX-GM researchers enjoy state-of-the-art facilities, technologies and core services required to advance genomic medicine. Relevant to the current proposal is our extensive data storage, sharing and analysis capabilities. JAX is deeply committed to development of our compute infrastructure and expertise. Since JAX-GM was founded in 2012, we have devoted resources to build high-performance computing capabilities with nearly ~2000 cores and 2 PB of storage space, and these resources will be scaled up further over the next few years. In addition to the computational resources available on site, JAX-GM computational biologists and staff are developing cloud-ready software pipelines to take full advantage of the rapid growth of cloud computing in genomics. This U01 application is therefore perfectly synergistic with the direction of JAX-GM.

I am particularly enthusiastic about the Center's focus on structural variation, a powerful form of genomic variation for which the project PIs, Drs. Charles Lee, Merk Gerstein and Li Ding, have unmatched blend of expertise. Together they bring the requisite experience in SV discovery and functional analysis necessary for the proposed program, particuarly within the framework of large-scale consortia such as the TOPMed Program.

In summary, I am very enthusiastic about the potential for the this project to make significant inroads in our understanding of human genomic diseases and am committed to investments that will support and enable this project to achieve its goals.

Thank you for considering this proposal.

Sincerely,

Edison T. Liu, M.D.

Then you MULTIPLE PI LEADERSHIP PLAN

Rationale:
This TOPMed project represents an inter-institutional effort that leverages existing, strong collaborative relationships among three leaders in the field of structural variant (SV) detection and functional interpretation—Drs. Charles Lee, Mark Gerstein and Li Ding. Each PI brings significant experience in leading and participating in large-scale genome sequencing consortia, which will be enable successful execution of project goals and facilitate integration within the broader NHLBI TOPMed Program. The unique and complementary expertise that each PI brings to the program justifies a multi-PI approach. The scientific success of the proposed project is to provide a platform of pipelines for comprehensive, high-resolution and large-scale SV analysis.

Leadership Roles and Responsibilities:
Dr. Charles Lee is Professor and Scientific Director of The Jackson Laboratory for Genomic Medicine. He is well-recognized for his contributions to the genomics field among the first to define copy number variants as a highly abundant form of natural human genetic variation and as a pioneer in the development of high-resolution methods for identifying human structural genomic variants. He has also been highly successful as leader of the 1000 Genomes Project SV group, in which he oversaw the comprehensive identification of the full repertoire of SV events in 2,500 healthy genomes and defined the methodology for identifying SVs from whole genome sequencing datasets. For this project, he will lead efforts to develop and integrate tools for high-resolution SV discovery and to build a reference database of complex SVs from TOPMed cohort data as more fully described in the Research Strategy as Aim 1. As contact PI, Dr. Lee will be responsible for communications with NHLBI and the TOPMed Program, including the submission of annual progress reports.

Dr. Mark Gerstein is the AL Williams Professor of Biomedical Informatics at Yale University and Co-Director of the Yale Computational Biology Program, and he is a leader in the field of computational genomics. He has designed and developed a wide array of databases and computational tools to mine genomic data in humans, as well as in many other organisms. He has developed quantitative approaches and practical tools for the processing of next-generation sequencing data, including those related to chIP-seq, RNA-seq and the detection of DNA structural variation. He has also served as a computational lead on numerous NIH-funded genomics projects (e.g., ENCODE & 1000 Genomes Project), and is co-leading the International Cancer Genome Consortium's analysis of mutations in regulatory regions group. For this project, Dr. Gerstein will lead efforts to develop tools to examine the functional impact of the identified SVs and to develop a novel pipeline of methodologies for functional annotation of variants and characterization of associated biological processes (described in Aim 2).

Dr. Li Ding is Associate Professor of Medicine at Washington University School of Medicine with recognized research expertise that includes the development of computational tools for analyzing NGS data, with a particular emphasis on variant detection and interpretation. Dr. Ding has led several landmark multi-center studies and produced a number of seminal publications in the field of genomics. Dr. Ding is currently co-leading the ICGC mutation calling group and she also serves as a co-chair for the oncogenic process group for the TCGA PanCanAtlas project. For this project, Dr. Ding will lead efforts to develop the cloud pipeline for genotyping high-impact SVs across a large number of samples in the TOPMed Program as described in Aim 3.

Communication:
The PIs are committed to regular and productive communication between each other and all relevant staff to facilitate the goals and objectives of the project. At the beginning of the project, the PIs will conduct a kickoff joint group meeting at The Jackson Laboratory for Genomic Medicine, to be attended by all key personnel and significant contributors to discuss program objectives, tasks, roles and responsibilities. Thereafter, the PIs will meet by video- or tele-conference on a biweekly basis to discuss research progress, address challenges and find solutions and alternative outcomes as needed. They will share their respective discussions with other key personnel to ensure that everyone is up to date with the latest status of the projects and results.

The PIs will also meet with their respective project staff weekly to evaluate progress and discuss any issues that have arisen in their work. These meetings can often clarify processes, bring to light new data sources, and to identify areas where others need to be consulted or indicate what additional resources are needed to accomplish a particular aspect of the project. Additional one-on-one communications (e.g., between key personnel, laboratory and administrative personnel) will take place on an *ad hoc* basis by email, phone or videoconference. A group meeting with all PIs and key personnel will be scheduled quarterly at a minimum to discuss project progress as a group. These meetings are also important for maintaining open communication between all key personnel and enable all to work together to discuss any changes in direction of the research projects.

The PIs will seek maximal communication with TOPMed Project Scientists and leaders of the data production, data analysis and coordinating centers of the TOPMed Program. As a collaborative NIH-funded program, the PIs also anticipate interactions with NHLBI scientific program officials. The PIs will attend TOPMed steering committee meetings through regular teleconferences and two annual in-person meetings at the NIH.

Organization, Governance and Administration:
All three of the Principal Investigators, Drs. Lee, Ding and Gerstein, participated in the preparation of this proposal, and are in agreement with the scientific and management plans. In particular, each has agreed to accept responsibility for the scientific leadership of the this project. Although Drs. Lee, Gerstein and Ding have synergistic expertise, the scientific direction for each Aim will generally be the domain of the PI with relevant expertise. Any dispute will be addressed first by the PIs. Should there be a scientific conflict, the PI with the most relevant expertise for the question will make the decision. If a resolution cannot be obtained, then the PIs will work with NHLBI program staff responsible for the TOPMed Program to resolve the dispute. All PIs agree to abide by the decisions made in such an instance.

The PIs have an established track record of successful scientific collaboration and will communicate frequently and regularly via email. The research group under the direction of each PI will meet weekly in person, and include, via video or teleconference, other project researchers as appropriate. Through the standing biweekly video- or teleconference meetings, the PIs will collectively oversee and coordinate of the scientific direction of the project and discuss management issues. New and major changes in research direction will be discussed with NHLBI program officials, as appropriate.

The Jackson Laboratory will be the prime grantee, and subaward agreements will be established with the McDonnell Genome Institute at the Washington University in St. Louis and Yale University. All institutions participating in this proposal have signed a Statement of Intent to enter into an inter-institutional agreement to include terms and conditions consistent with the NHLBI TOPMed Program and NIH policies for the U01 Cooperative Agreement mechanism. These include agreeing to accept close coordination and participation of NHLBI Project Scientists, and adhering to NHLBI policies regarding intellectual property, data release and other policies that might be established during the course of the TOPMed Program. The anticipated budget and funding distributions for the project, as delineated in the proposal budget and budget justifications, have been agreed upon by the PIs. The PIs will be responsible for the fiscal management, research compliance and related research administration management at their respective institutions.