

PI: Hotopp, Julie	Title: Extent and Significance of Bacterial DNA Integrations in the Human Cancer Genome	
Received: 10/09/2014	FOA: RM14-003	Council: 05/2015
Competition ID: FORMS-C	FOA Title: NIH TRANSFORMATIVE RESEARCH AWARDS (R01)	
1 R01 CA206188-01	Dual: OD,RM	Accession Number: 3745895
IPF: 820104	Organization: UNIVERSITY OF MARYLAND BALTIMORE	
Former Number: 1R01OD020504-01	Department: Institute for Genome Sciences	
IRG/SRG: ZRG1 BCMB-A (51)	AIDS: N	Expedited: N
<u>Subtotal Direct Costs</u> <u>(excludes consortium F&A)</u> Year 1: 439,960 Year 2: 546,783 Year 3: 480,617 Year 4: 497,324 Year 5: 459,975	Animals: N Humans: Y Clinical Trial: N Current HS Code: 30 HESC: N HFT: N	New Investigator: N Early Stage Investigator: N
<i>Senior/Key Personnel:</i>	<i>Organization:</i>	<i>Role Category:</i>
Julie Hotopp	University of Maryland, Baltimore	PD/PI
Maria Baer	University of Maryland, Baltimore	Other (Specify)-Collaborator

APPLICATION FOR FEDERAL ASSISTANCE

SF 424 (R&R)

3. DATE RECEIVED BY STATE		State Application Identifier MD
1. TYPE OF SUBMISSION*		4.a. Federal Identifier
☐ Pre-application ☒ Application ☐ Changed/Corrected Application		b. Agency Routing Number
2. DATE SUBMITTED 2014-10-09	Application Identifier 23255	c. Previous Grants.gov Tracking Number
5. APPLICANT INFORMATION		
Legal Name*: University of Maryland, Baltimore		Organizational DUNS*: [REDACTED]
Department: Microbiology		
Division: [REDACTED]		
Street1*: [REDACTED]		
Street2: [REDACTED]		
City*: [REDACTED]		
County: [REDACTED]		
State*: [REDACTED]		
Province:		
Country*: [REDACTED]		
ZIP / Postal Code*: [REDACTED]		
Person to be contacted on matters involving this application		
Prefix:	First Name*: [REDACTED]	Middle Name: [REDACTED] Last Name*: [REDACTED] Suffix:
Position/Title:	AVP, Sponsored Programs Administration	
Street1*: [REDACTED]		
Street2: [REDACTED]		
City*: [REDACTED]		
County:		
State*: [REDACTED]		
Province:		
Country*: [REDACTED]		
ZIP / Postal Code*: [REDACTED]		
Phone Number*: [REDACTED]	Fax Number: [REDACTED]	Email: [REDACTED]
6. EMPLOYER IDENTIFICATION NUMBER (EIN) or (TIN)* [REDACTED]		
7. TYPE OF APPLICANT*		H: Public/State Controlled Institution of Higher Education
Other (Specify):		
Small Business Organization Type		☐ Women Owned ☐ Socially and Economically Disadvantaged
8. TYPE OF APPLICATION*		If Revision, mark appropriate box(es).
☒ New ☐ Resubmission		☐ A. Increase Award ☐ B. Decrease Award ☐ C. Increase Duration
☐ Renewal ☐ Continuation ☐ Revision		☐ 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: NIH Transformative Research Awards (R01)
11. DESCRIPTIVE TITLE OF APPLICANT'S PROJECT* Extent and Significance of Bacterial DNA Integrations in the Human Cancer Genome		
12. PROPOSED PROJECT Start Date* Ending Date* 08/01/2015 07/31/2020		13. CONGRESSIONAL DISTRICTS OF APPLICANT MD-007

14. PROJECT DIRECTOR/PRINCIPAL INVESTIGATOR CONTACT INFORMATION

Prefix: First Name*: Julie Middle Name: Christine Last Name*: Hotopp Suffix:

Position/Title: Assistant Professor

Organization Name*: University of Maryland, Baltimore

Department: [REDACTED]

Division: [REDACTED]

Street1*: [REDACTED]

Street2: [REDACTED]

City*: [REDACTED]

County: [REDACTED]

State*: [REDACTED]

Province: [REDACTED]

Country*: [REDACTED]

ZIP / Postal Code*: [REDACTED]

Phone Number* [REDACTED] Fax Number: [REDACTED] Email*: [REDACTED]

15. ESTIMATED PROJECT FUNDING

- a. Total Federal Funds Requested* [REDACTED]
- b. Total Non-Federal Funds* [REDACTED]
- c. Total Federal & Non-Federal Funds* [REDACTED]
- d. Estimated Program Income* [REDACTED]

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: [REDACTED]
- 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. SFLL or OTHER EXPLANATORY DOCUMENTATION

File Name: [REDACTED]

19. AUTHORIZED REPRESENTATIVE

Prefix: First Name*: [REDACTED] Middle Name: [REDACTED] Last Name*: [REDACTED] Suffix:

Position/Title*: Asst Director, Sponsored Program Admin

Organization Name*: University of Maryland, Baltimore

Department: [REDACTED]

Division: [REDACTED]

Street1*: [REDACTED]

Street2: [REDACTED]

City*: [REDACTED]

County: [REDACTED]

State*: [REDACTED]

Province: [REDACTED]

Country*: [REDACTED]

ZIP / Postal Code*: [REDACTED]

Phone Number*: [REDACTED] Fax Number: [REDACTED] Email*: [REDACTED]

Signature of Authorized Representative*

Date Signed*

10/09/2014

20. PRE-APPLICATION File Name: [REDACTED]**21. COVER LETTER ATTACHMENT** File Name: [REDACTED]

424 R&R and PHS-398 Specific Table Of Contents

Page Numbers

SF 424 R&R Cover Page-----	1
Table of Contents-----	3
Performance Sites-----	4
Research & Related Other Project Information-----	5
Project Summary/Abstract(Description)-----	6
Project Narrative-----	7
Facilities & Other Resources-----	8
Equipment-----	11
Research & Related Senior/Key Person-----	14
Research & Related Budget Year - 1-----	22
Research & Related Budget Year - 2-----	25
Research & Related Budget Year - 3-----	28
Research & Related Budget Year - 4-----	31
Research & Related Budget Year - 5-----	34
Budget Justification-----	37
Research & Related Cumulative Budget-----	42
PHS398 Cover Page Supplement-----	43
PHS 398 Research Plan-----	45
Specific Aims-----	46
Research Strategy-----	47
Human Subjects Section-----	59
Protection of Human Subjects-----	59
Women & Minorities-----	60
Planned Enrollment Report-----	61
Children-----	62
Bibliography & References Cited-----	63
Letters Of Support-----	68
Resource Sharing Plans-----	69

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: University of Maryland, Baltimore
Duns Number: [REDACTED]
Street1*: [REDACTED]
Street2: [REDACTED]
City*: [REDACTED]
County: [REDACTED]
State*: [REDACTED]
Province:
Country*: [REDACTED]
Zip / Postal Code*: [REDACTED]
Project/Performance Site Congressional District*: MD-007

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: IGS
DUNS Number:
Street1*: [REDACTED]
Street2: [REDACTED]
City*: [REDACTED]
County:
State*: [REDACTED]
Province:
Country*: [REDACTED]
Zip / Postal Code*: [REDACTED]
Project/Performance Site Congressional District*: MD-007

File Name

Additional Location(s)

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	00007145
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:	Mexico
6.b. Optional Explanation:	
Filename	
7. Project Summary/Abstract*	2014_NCI_TR01_Summary.pdf
8. Project Narrative*	2014_NCI_TR01_Narrative.pdf
9. Bibliography & References Cited	2014_NCI_TR01_References.pdf
10. Facilities & Other Resources	FacilitiesOct2014.pdf
11. Equipment	EquipmentOct2014.pdf

SUMMARY

The integration of exogenous DNA into the human genome can cause somatic mutations associated with oncogenesis. For example, the insertion of HPV DNA into human chromosomes is the single most important event leading to tumorigenesis in cervical cancer. It is also now preventable with vaccines against HPV. In contrast to viral DNA integrations, the instances and repercussions of bacterial DNA integration into the somatic human genome are less clear. Yet bacterial DNA integrations found to be associated with tumorigenesis could also be prevented using therapeutics like vaccines that limit exposure to the bacteria. This proposal has three objectives aimed at addressing our gap in knowledge about bacterial DNA integrations. First, virtual machines will be developed for LGTSeek and LGTview, our bioinformatics tools that we have used previously to detect bacterial DNA integrations in human genome sequencing projects. These virtual machines would enable these tools to be run by a wide variety of users, from the bioinformatically savvy to the naïve. This provides a resource to the community to enable detection of such integrations by a wider variety of scientists. In addition, LGTSeek and LGTView will be used to further interrogate publicly available cancer genome data where such integrations are likely to occur because the tissues are exposed to the microbiome (e.g. colon). Second, genome and transcriptome sequencing will be undertaken of new stomach adenocarcinoma samples and acute myeloid leukemia samples. This objective is aimed at reproducing previous results that suggest the presence of bacterial DNA integrations. These sequencing efforts would include control samples with exogenous bacterial nucleic acids added to the sample in order to quantify the formation of chimeras in modern sequencing techniques. Third, the effect that previously detected bacterial DNA integrations have on transcription will be interrogated using luciferase reporter constructs. Integrations that lead to up-regulation of the gene will be further interrogated by reconstructing the integrations in cells using the CRISPR/Cas9 system. Collectively, this research is expected to improve our understanding of the extent and significance of bacterial DNA integrations in the somatic human genome.

NARRATIVE

The integration of bacterial DNA in the human somatic genome could result in diseases due to insertional mutagenesis. The experiments proposed here aim at developing tools for the research community to detect such integrations, assessing the prevalence of such integrations in cohorts of gastric cancer and acute myeloid leukemia patients, and testing the consequence of known integrations on transcript abundance and phenotype.



The Institute for Genome Sciences (IGS), established within the University of Maryland, School of Medicine (UMSOM) on May 1, 2007, houses an inter-disciplinary, multi-departmental team of collaborative investigators with a broad spectrum of research programs related to the genomics of infectious disease agents, human microbial metagenomics, and bioinformatics. The Institute is led by Claire M. Fraser, Ph.D., one of the pioneering genome scientists and previously the Director and President of The Institute for Genomic Research (TIGR).

IGS is currently located at the UMB BioPark, a biomedical research park located adjacent to the University of Maryland Medical Research Complex. Within the BioPark, the Institute occupies part of the fifth floor and the entire sixth floor encompassing ~38,000 sq ft of total laboratory, office, conference, and interactive gathering space. The newly constructed facility represents a 40 million dollar commitment to genomic research for the School of Medicine.

IGS has two resource centers that support investigators at IGS and UMSOM: the **Genomics Resource Center** (GRC) and the **Informatics Resource Center** (IRC).

The Genomics Resource Center (GRC) is a high-throughput core laboratory and data analysis group led by Dr. Sadzewicz, Administrative Director, and Mr. Luke Tallon, Scientific Director, who together have more than 35 years experience in managing high-throughput sequencing and analysis operations. The GRC includes bioinformatics software engineers, bioinformatics analysts, project managers, and research specialists who have extensive experience in planning and managing projects ranging in scope from small-scale amplicon and plasmid sequencing to large-scale comparative genomic and transcriptome sequencing. The laboratory services offered by the GRC include genomic and transcriptomic library construction, whole genome sequencing, metagenomic sequencing, exome and custom targeted capture sequencing, cDNA/EST sequencing, transcriptome sequencing, epigenomic sequencing, amplicon sequencing, and customized sequencing services.

The GRC occupies 7000 sq. ft. of space within the BioPark II facility on the UMSOM campus. This BSL-2 research space consists of wet laboratory space, a sequencer facility, a cryostorage facility, a cold room, a dark room, a reagent storage facility, conference rooms, and office space. The GRC sequencer fleet currently includes two AB 3730xl DNA Analyzers, one AB 3130xl Genetic Analyzer, three Illumina HiSeq2500/2000 sequencers, three Illumina MiSeq sequencers, and a Pacific Biosciences RS II sequencer. Our combined annual sequencing capacity is more than 90 trillion bases of high-quality, passed-filter data. We continue to scale our operations to meet growing demand and our laboratory and office space could support more than twice the current capacity. The Illumina sequencers represent our highest-throughput platform, with the ability to deeply sequence multiple human genomes per week. The MiSeq platform combines the efficient and trusted Illumina chemistry with a lower-throughput but faster runtime instrument that is capable of generating data sets in less than one day. The PacBio RS II is the first commercially viable single-molecule sequencer and has the potential to change the landscape of high-throughput sequencing. It generates significantly longer reads than any other current platform, with average read lengths exceeding 8 kbp. It has a run time of two to three hours, compared to the many days it takes to run an Illumina sequencer.

Consistency is essential to the efficient operation of a high-throughput genomics laboratory. To ensure this consistency, all activities within the GRC are based on tested and approved Standard Operating Procedures. These SOPs are written in a standardized format and include detailed instructions for the performance of each laboratory or bioinformatics protocol. A limited number of approved employees have the authority to create or modify SOPs. Each SOP has a primary author and at least one independent reviewer. New or modified SOPs undergo review and scientific approval by the assigned reviewer and are approved for publication and use by the GRC by the GRC Directors. Approved and published SOPs are available in read-only PDF format on the

internal IGS web site. All GRC staff members are trained in the use of these SOPs and are notified upon release of a new or modified SOP. At minimum, each SOP is reviewed annually. The author of each SOP is responsible for the regular update and improvement of the SOP on a more frequent basis as needed.

In order to efficiently manage project flow and accounting for a large number of parallel projects the GRC has implemented a custom-designed Laboratory Information Management System (LIMS). The LIMS ensures accurate and robust tracking of all samples and data and features role-based user access control and barcoding of all LIMS objects. The LIMS also tracks storage conditions and expiration dates for reagents to ensure that reagents not stored properly or beyond their expiration dates are eliminated from use in the laboratory. Reagent inventories are carefully managed to ensure appropriate redundancy in stocks. It is important to maintain stocks sufficient to absorb the loss of a reagent lot that fails quality control testing, but also to keep inventories as fresh as possible. A workflow module within the LIMS provides process control and validation to ensure each sample is processed according to established and reviewed protocols. Workflows are an ordered set of structured tasks with required sample types and properties, required data entry points, and task dependencies to prevent downstream tasks from being processed out of order. A customized reports module is used to generate automated and user-initiated reports on all data objects in the LIMS. These reports facilitate project tracking as well as laboratory quality control measures. A project management and scheduling module tracks pipeline capacity utilization and facilitates scheduling of large scale, multi-platform projects.

In addition to the high throughput laboratory resources, the GRC provides new technology development, protocol optimization, sequence data analysis and storage, sequence assembly, genome finishing, and data submission services using state of the art laboratory and software tools. Sequence assembly services are performed using a large suite of validated assemblers, including Celera Assembler, ALLPATHS-LG, Velvet, SOAPdenovo, Abyss, HGAP3, MaSuRCA, Mira, and the AMOS package of tools. Each assembler is optimized to the assembly of particular data sets and we have developed a pipeline to combine assemblers to perform hybrid assembly with any combination of data from our three sequencing platforms. Similarly, large-scale sequence alignment is performed using a series of alignment methods, including MUMmer, BLAST, MOSAIK, BWA, and Bowtie2. Following alignment of genomic or transcriptomic data, we use SAMtools, GATK, TopHat/Cufflinks, and other customized software for variant discovery and expression profiling. We have established data submission pipelines to deliver sequence and assembly data to NCBI's Trace Archive, Short Read Archive, GEO, dbSNP, dbGaP, and Genbank. We also provide data delivery services via FTP hosting and hard media.

The **Informatics Resource Center**, under the direction of Owen White, Ph.D., provides genome annotation and analysis services as well as IT infrastructure, web, and database services to IGS investigators. The IRC includes a staff of over forty scientists, engineers, analysts, and systems engineers that work together to conduct research in bioinformatics and provide genome analyses. The IRC staff is organized along scientific platforms and functional areas of expertise. Scientists, typically biologists, lead the scientific platforms and coordinate with the engineering staff to provide the analysis required for projects. The major scientific platforms supported include prokaryotic, viral, eukaryotic, and mammalian genomics, metagenomics, informatics research, and systems biology. The engineering group has built expertise in a number of functional areas that include genome assembly and annotation, genome visualization, NextGen sequence analysis including transcriptomics, epigenomics, structural variant detection, comparative genomics, and statistical modeling. In addition to the work in support of the scientists at IGS/UMSOM, the IRC has been awarded a number of grants to develop genomic resources that include the NIH Human Microbiome Project Data Analysis and Coordination Center, the IGS Annotation Engine Service, CloVR (Cloud Virtual Resource), and GEMINA (Genomic Metadata for Infectious Agents). To enable our research activities IGS has a state-of-the-art computational infrastructure that includes a computational grid, a 10 Gbps network infrastructure, high performance database servers, and a tiered storage management system.

IGS is connected by a high-performance switched 10 Gbps network, powered by Cisco equipment, to the rest of the campus. All UMB buildings are connected to the LAN backbone and core switches via fiber cabling. IGS, as part of UMB maintains a 10 Gbps connection via the National Lambda Rail (NLR) to other NLR sites, a 1 Gbps connection to Internet2, the high-speed network designed to facilitate collaboration and communication among research institutions, as well as the aggregated bandwidth of 20 Mbps to the regular internet network.

The computational grid is built around ten high-performance high-memory multi-processor machines (64-512 GB RAM, 4-8 CPU multi-core processors) for memory and compute intensive applications such as genome and transcriptome assembly, multiple genome alignment, and over ninety high throughput computational

nodes (16-48 GB RAM 2 CPU multi-core Intel Xeon processor machines) for running distributed applications such as BLAST, HMMsearch, etc. The grid scheduling is managed by Sun N1 Grid Engine (SGE) distributed computing system.

To address the ever expanding data sets generated by next generation genome sequencing technologies at a reasonable cost we have deployed a tiered storage infrastructure consisting of 4 tiers of random access storage and a fourth tier of serial access tape media storage for archival and data backup. Tier 1 storage, with a capacity of 100 TB, is a high-performance replicated, grid-attached storage where all the mission critical data including primary sequence data is stored. Tier 2 storage is built around a scalable high-performance global file system from EMC/Isilon with a capacity of >500 TB that currently support over 2 GB/s throughput. All current project data is stored on this tier and most computational activities occur here. Tier 3 storage, also EMC/Isilon, has a capacity of >600 TB that currently support 600 GB/s throughput and is used for inactive longer-term storage. We have also deployed 250 TB of disk based long-term archival storage. In addition we provide offline storage that is used for data backups as well as archival data such as the raw data generated by sequencers. This tier is built around a tape library and is integrated with Tiers 1-3 to provide daily, weekly, monthly, and annual backups.

IGS has built a high availability (HA) web infrastructure centered on clustered Apache web servers and load balancers. This ensures minimal downtime and uninterrupted access to our site and data. The IGS bioinformatics group supports MySQL and PostgreSQL databases. All database servers are attached to high performance SAN attached disk systems to ensure speed as well as expandability and are accessible from all desktops and servers at IGS. All storage is accessible from the desktops as well as computational servers via network file sharing (NFS) or similar protocol hosted by high-performance redundant file servers from EMC/Isilon.

At IGS we have over 100 commonly used bioinformatics tools installed centrally that are kept up to date. All IGS employees and collaborators have secure external access to this infrastructure via a virtual private network (VPN) powered by Cisco equipment.

In addition to the IT infrastructure dedicated for IGS use, we were awarded an NSF MRI-R2 grant to setup a shared computational infrastructure, the Data Intensive Academic Grid (DIAG). The DIAG meets the bioinformatics needs of over 300 researchers at over 150 institutions worldwide including the University of Maryland School of Medicine. The DIAG includes a computational infrastructure, a high-performance storage network, and optimized data sets generated by mining the data from public data repositories like Camera and NCBI to enable researchers to perform analyses. The DIAG computational infrastructure includes 3,000 cores (125 nodes) for high-throughput computational analysis and 160 cores (5 nodes) for high-performance computational analysis. To handle the large amounts of data generated by sequencing experiments it includes ~600 TB of clustered high-performance parallel file system storage. The bioinformatics community accesses DIAG using Ergatis, a web based pipeline creation and management tool, bioinformatics oriented VMs including IGS' Cloud Virtual Resource (CloVR), as well as interactive and programmatic access using technologies such as Nimbus, and the Virtual Data Toolkit from the Open Science Grid (NSF supported projects).



Major Equipment

Allowing for general molecular biology, large-scale DNA sequencing and functional genomics, technology development, and follow-up characterization of specific gene and protein families, a large equipment inventory supports IGS research activity. Temperature-critical equipment and facilities (freezers, cold boxes, cold rooms, cryovessels, incubators, freezer rooms, sequencer rooms and data rooms) will be attached to the AmegaView Centralized Monitoring System. This environmental monitoring system provides temperature and power surveillance data logged directly to a SQL network database remotely accessible through the network or via the telephone. This system is also equipped with an automated telephone/email/pager alarm. To protect IGS operations from loss of electrical power, the facility is outfitted with back up generators. In the event of a power failure, the emergency back-up generators will resume electrical power within 60 s. An uninterrupted power supply (UPS) unit facility will provide continuous power to servers, sequencers, and arrays that require continuous power.

#	Equipment (<i>partial list</i>)
1	Affymetrix GCS3000 TG System
1	Agilent 2100 Bioanalyzer
1	Agilent 2200 Tape Station
1	AmegaView Centralized Environmental Monitoring System
1	Applied Biosystems 3130xl Genetic Analyzer with GeneMapper Software
2	Applied Biosystems 3730xl Genetic Analyzers
1	Applied Biosystems 7900HT Fast Real-Time PCR System
10	Applied Biosystems 9700 384-well Thermal Cyclers, Dual
6	Applied Biosystems 9700 96-well Thermal Cyclers, Dual
5	Applied Biosystems 9700 96-well Thermal Cyclers, Single
1	Aushon 2470 Arrayer Microarray Printing Platform
1	Axon Instruments GenePix [®] 4000B Microarray Scanners
1	Baker SterilGARD [®] III Biological Safety Cabinet, Vertical Laminar Flow Bench, Class II, Type A2 with Canopy Exhaust Connection
1	Beckman Coulter Multimek TM 96 Automated 66-Channel Pipettor
6	Beckman Coulter Allegra [®] 6R Benchtop Centrifuge, Refrigerated
1	Beckman Coulter Avanti [®] J-26 XP High-Performance Centrifuge
1	Beckman Coulter Avanti [®] J-E Centrifuge
1	Beckman Coulter FXP Dual Multichannel Span 8
1	Beckman Coulter Biomek [®] NXP Liquid Handling Work Station with Cytomat Plate Hotel
1	Beckman Coulter Biomek [®] NXP Liquid Handling Work Station with Hudson Control Group Precipitation Work Cell
1	Beckman Coulter BioMek [®] 3000 Laboratory Automation Workstation
1	Beckman Coulter CEQ [™] 8000 Genetic Analysis System
1	Beckman Coulter DU [®] 800 UV/Vis Spectrophotometer
1	Beckman Coulter Z1 Single Threshold Analyzer
1	Beckman-Coulter LS6000 Liquid Scintillation Counters
1	Bellco Hybridization Ovens
1	Biocane 20 Cryogenic Vessel
1	Bio-Rad CHEF DRIII Chiller Mapper System
2	Bio-Rad DNA Engine 96-well Thermal Cycler

- 2 Bio-Rad Gel Doc™ XR Systems
- 2 Bio-Rad Gene Pulser Xcell™ Total System
- 1 Bio-Rad iCycler 96-well Thermal Cycler
- 1 Caliper LabChip GX System
- 1 Caliper LabChip XT System
- 2 Consolidated Sterilizer Model 2SSR-3A-MC
- 1 Covaris E-210
- 1 Coy Laboratories Type B Anaerobic Chamber with Auto Airlock, Low Temp Range Incubator (Model 2000) and Oxygen/Hydrogen Gas Analyzer, (Model 10x)
- 1 CRS DNA Isolation and Dispensing robotic system
- 1 Diagenode Bioruptor™ UCD-200
- 1 Elga Centra High Capacity Water System
- 12 Elga Purelab Option-Q 7/15 RODI Water System
- 16 Eppendorf Microcentrifuges
- 1 FastPrep®-24 System
- 1 Fisher Scientific Model 100 Sonic Dismembrator
- 1 Fluidigm BioMark Real Time PCR System
- 1 Fotodyne Apprentice WS UV 21 Imaging System
- 1 Freezerworks Unlimited 3.1 with Zebra TLP2844 Thermal Transfer Printer and Brady Code Reader 3 Batch Reader
- 1 Gene Machines® HydroShear® 1 DNA Shearing Device
- 2 General Purpose Refrigerator/Freezer Units
- 1 Hudson Control Robotic Work Cells for Cell Lysis with Compressor
- 1 Illumina HiSeq 2000 Sequencing System & cbot
- 2 Illumina HiSeq 2500 Sequencing System & cbot
- 2 Illumina MiSeq Sequencing System
- 1 Illumina MiSeqDx Sequencing System
- 1 Labconco Class II Type A2 Biosafety Cabinet
- 1 Labconco PCR Enclosure
- 3 LabLine Orbital Shakers
- 1 Laboratory Information Management System (LIMS) Sapio
- 1 Lancer 1300 LX Washer-Disinfector
- 1 Locator 6 Plus Cryovessel with Level Monitor
- 1 Millipore CytoFluor™ 2350 Fluorescence Measurement System
- 2 Milli-Q® Synthesis Water Purification Systems
- 2 MJ Research PTC-200 Peltier Thermal Cycler
- 1 Molecular Devices SpectraMax® M5 Multimode Plate Reader
- 1 MP Biomedical LLC Fast Prep-24
- 1 Multitron Infors Shaker-Double Multitron II
- 3 NanoDrop® ND-1000 and Software
- 1 Nebulization Work Station with Enclosure and Exhaust System
- 1 New Brunswick Scientific Temperature Controlled Excella™ E25 Series Incubator/Shaker
- 1 Nikon SMZ-1 Stereo Microscope with Focusing Base and Fiber Optic Light Source
- 1 Nikon Eclipse TE2000-E2 Imaging System with 10 x Flat Field Phase Objective NA 0.25 WD 6.1 mm; Plan FLELWD DMPhase 20 x Objective NA 0.45 WD 7-8.1 mm CC; Plan Fluor 40 x Objective NA 0.75 WD 72 mm SPG; CFI Plan Fluor 60 x Oil Immersion Objective NA 0.5 – 1.25; Plan Fluor 100 x Oil Immersion Objective NA 1.3 WD 0.20 mm; DAPI, FITC, TRITC and Cyan GFP Fluorescence Filter Cubes; Elements, Advanced Research Software; DS-5M Digital Sight Camera; and Cool Snap ES2 20 MHz Camera.
- 1 Norgren Systems Colony Picker CP7200
- 1 NuAire Biological Safety Cabinet, Vertical Laminar Flow Bench, Class II, Type A2 with carbon filter Model 437-300
- 1 Olympus CK II Binocular Inverted Phase Contrast Microscope with 4 x Objective, 10 x A10PL 0.25 160/0.17 Objective, 20 x LWD C A20PL 0.4 160/1.2 Objective, and 40 x LWD CPPlan40FPL 0.55 160/0.17 Objective

1	Olympus BH-2 Binocular Microscope with 10 x WHK eye piece containing a KR-406B - 21mm 10x10, A-J, 1-10 Reticule as well as 4 x DPlan 0.10 160/0.17 Objective; 10 x DPlan 0.25 160/0.17 Objective, 40 x DPlan 0.65 160/0.17 Objective and 100 x DPlan 1.25 160/0.17 Oil Immersion Objective.
1	Pacific Biosciences RS II Sequencer
1	Qiagen TissueLyser
1	Qiagen QIAgility
1	Qiagen Symphony
2	Roche Applied Sciences 454 Life Sciences™ Genome Sequencer FLX
1	Sage BluePippin Electrophoresis Platform
1	Sanyo Cooled 254 Liter Incubator
1	Sanyo Double Stack Direct Heat Air Jacketed Carbon Dioxide Incubator
1	Sanyo Double Stack Direct Heat Air Jacketed Carbon Dioxide Incubator with UV Light
20	Sanyo VIP Ultra-Low Temperature Freezers (25.7 cu ft)
4	Savant DNA 120 Speed Vacuum
1	SPEX Sample Prep Freezer Mill (Model 6870) and Autoextractor
1	Sorvall Legend 21 Centrifuge with rotor
1	Stratagene Stratalinker
1	Tecan F200 Pro Plate Reader
1	Techne Hybridiser HB-1D hybridization oven
2	Thermo 1300 Series Class II, Type A2 Biosafety Cabinets (1 with Canopy Exhaust Connection)
1	VWR Diurnal Growth Chamber
1	Zebra 110xiPlus Printer
1	-30 °C Sanyo Biomedical Freezer
19	-20 °C Freezers
3	37 °C Incubators
1	4 °C General Purpose Refrigerator (27 cu ft)
7	4 °C Chromatography/Glass Storage Refrigerators

RESEARCH & RELATED Senior/Key Person Profile (Expanded)

PROFILE - Project Director/Principal Investigator				
Prefix:	First Name*: Julie	Middle Name Christine	Last Name*: Hotopp	Suffix:
Position/Title*:	Assistant Professor			
Organization Name*:	University of Maryland, Baltimore			
Department:	[REDACTED]			
Division:	[REDACTED]			
Street1*:	[REDACTED]			
Street2:	[REDACTED]			
City*:	[REDACTED]			
County:				
State*:	[REDACTED]			
Province:				
Country*:	[REDACTED]			
Zip / Postal Code*:	[REDACTED]			
Phone Number*:	[REDACTED]	Fax Number:	E-Mail*: [REDACTED]	
Credential, e.g., agency login: [REDACTED]				
Project Role*: PD/PI			Other Project Role Category:	
Degree Type: Doctor of Philosophy			Degree Year: 2002	
Attach Biographical Sketch*: Attach Current & Pending Support:			File Name 2014Dunning_Hotopp_Biosketch_TR01_format.pdf	

PROFILE - Senior/Key Person				
Prefix:	First Name*: Maria	Middle Name R.	Last Name*: Baer	Suffix:
Position/Title*:	Visiting Assistant Professor			
Organization Name*:	University of Maryland, Baltimore			
Department:	[REDACTED]			
Division:	[REDACTED]			
Street1*:	[REDACTED]			
Street2:	[REDACTED]			
City*:	[REDACTED]			
County:				
State*:	[REDACTED]			
Province:				
Country*:	[REDACTED]			
Zip / Postal Code*:	[REDACTED]			
Phone Number*:	[REDACTED]	Fax Number:	E-Mail*: [REDACTED]	
Credential, e.g., agency login: [REDACTED]				
Project Role*: Other (Specify)			Other Project Role Category: Collaborator	
Degree Type: Bachelor of Arts			Degree Year: 1973	
Attach Biographical Sketch*: Attach Current & Pending Support:			File Name 2014_NCI_TR01_Baer_Biosketch_New_Form.pdf	

BIOGRAPHICAL SKETCH

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

NAME Julie Hotopp		POSITION TITLE Associate Professor, Institute for Genome Sciences, Dept. of Microbiology and Immunology, University of Maryland Baltimore School of Medicine	
eRA COMMONS USER NAME (credential, e.g., agency login) [REDACTED]			
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)	MM/YY	FIELD OF STUDY
University of Rochester Rochester, NY 14627, USA	B.S.	09/93-05/97	Microbiology & Immunology
Michigan State University East Lansing, MI 48824, USA	Ph.D.	05/97-05/02	Microbiology & Molecular Genetics
The Institute for Genomic Research Rockville, MD 20850, USA	Postdoctoral Training	05/02-08/05	Microbial Genomics

A. PERSONAL STATEMENT

I have 12 years of experience in genomics and informatics that includes two paradigm-shifting manuscripts on bacteria-animal lateral gene transfer. This is the greatest indication of my innovativeness as well as my ability and willingness to question paradigms. I began my post-doctoral fellowship shortly after the publication of a series of papers that led some to assume that lateral gene transfer could not occur in Metazoans. While many scientists believed this was a fact, my growing understanding of genomic techniques, a careful reading of the literature, and access to one specific dataset (the *Drosophila ananassae* genome sequencing reads) led me to question this paradigm. The result was our description of widespread lateral gene transfer from *Wolbachia* endosymbionts to their animal hosts. This manuscript also illustrates my ability to foster collaborations between individuals from diverse disciplines at multiple institutions to produce a seminal study, since the impact and interest was clearly amplified by the diverse set of evidence we put forth. Yet it is still held as fact by many that while invertebrates can acquire DNA from their microbiome, humans and other vertebrate animals cannot. Last year, challenging that paradigm, we described computational evidence for bacterial DNA integrations into the somatic human genome via lateral gene transfer. This result illustrates my innovating thinking and accomplishments as an independent researcher and my ability to coordinate a large project. Combined my research projects have involved extensive genome sequencing and analysis including the handling of big data and the development of novel genomic methods to study bacteria-animal interactions. Collectively, this demonstrates my ability to use state-of-the-art methods to challenge paradigms.

B. POSITIONS AND HONORS**Professional Experience**

2014-current Associate Professor (tenure track), Institute for Genomic Sciences and Department of Microbiology and Immunology, School of Medicine, University of Maryland Baltimore
2011-2014 Assistant Professor (tenure track), Institute for Genomic Sciences and Department of Microbiology and Immunology, School of Medicine, University of Maryland Baltimore
2007-2011 Assistant Professor (non-tenure track), Institute for Genomic Sciences and Department of Microbiology and Immunology, School of Medicine, University of Maryland Baltimore
2005-2007 Staff Scientist, Microbial Genomics, The Institute for Genomic Research/J. Craig Venter Institute

Other Experience and Professional Memberships

2002-current Member, American Society for Microbiology
2009 USDA Peer Review Panel Member
2009,2014 NSF Peer Review Panel Member

- 2009-current Ad Hoc reviewer: Swiss National Science Foundation, United Kingdom Medical Research Council, Portuguese Fundacao para a Ciencia e a Technologia, Ministerio da Educacao e Ciencia, Austrian Science Fund, French National Research Agency
- 2005-current Ad Hoc reviewer: Bioinformatics, BMC Evolutionary Biology, Central European Journal of Biology, Current Biology, Current Microbiology, Expert Reviews in Proteomics, ISME Journal, Insect Molecular Biology, Genome Biology and Evolution, Marine Biotechnology, Microbial Ecology, Mobile Genetic Elements, Molecular Biology and Evolution, Molecular Ecology, Molecular Genetics and Genomics, Nature Communications, Plant Physiology, PLoS Genetics, PLoS ONE, Proceedings of the National Academy of Science USA, Proceedings of the Royal Society B, Science, Trends in Parasitology

Honors & Awards

- 2010 NIH Director's New Innovator Award
- 2010 Daily Record's Leading Women of 2010 in Maryland
- 2010 Genome Technology Young Investigator, Genome Technology
- 2001 Rudolph Hugh Fellowship, Department of Microbiology and Molecular Genetics, Michigan State University
- 2001 Dr. Marvin Hensley Endowed Fellowship, College of Natural Science, Michigan State University
- 2000-2001 Distinguished Fellowship, Michigan State University
- 1997-1998 College of Natural Science Recruiting Fellowship, Michigan State University
- 1996, 1995 Howard Hughes Summer Fellow, University of Rochester
- 1994 McNair Freshman/Sophomore Scholars Program, University of Rochester
- 1993-1997 Xerox Undergraduate Scholarship, University of Rochester
- 1993-1997 Bausch and Lomb Scholarship, University of Rochester

C. SELECTED PEER-REVIEWED PUBLICATIONS (of a total of 37 publications; listed in order of relevance as it relates to those that illustrate my exceptional innovativeness and the significance of my past accomplishments)

1. Riley, D. R., K. B. Sieber, K. M. Robinson, J. R. White, A. Ganesan, S. Nourbakhsh, and **J. C. Dunning Hotopp**. 2013. Bacteria-human somatic cell lateral gene transfer is enriched in cancer samples. PLoS Comp Bio 9(6):e1003107. PMCID: PMC3688693
Summary: Paradigm-shifting paper presenting computational evidence for bacterial DNA integrations in the human somatic genome.
2. **Dunning Hotopp, J. C.**, M. E. Clark, D. C. S. G. Oliveira, J. M. Foster, P. Fischer, M. C. Muñoz Torres, J. D. Giebel, N. Kumar, N. Ishmael, S. Wang, J. Ingram, R. V. Nene, J. Shepard, J. Tomkins, S. Richards, D. J. Spiro, E. Ghedin, B. E. Slatko, H. Tettelin, and J. H. Werren. (2007) "Widespread lateral gene transfer from intracellular bacteria to multicellular eukaryotes." Science 317(5845):1753-6.
Summary: Paradigm-shifting paper presenting experimental and computational evidence for inherited bacterial *Wolbachia* DNA integrations into their Metazoan hosts. It is now an accepted fact that invertebrate genomes have bacterial DNA integrations that are inherited, of which some subset are functional and adaptive.
3. Robinson K. M., **J. C. Dunning Hotopp**. "Mobile elements and viral integrations prompt considerations for bacterial DNA integration as a novel carcinogen." Cancer Lett 2014, 352(2):137-144. PMCID: PMC4134975
Summary: In this review, we outline our rationale for hypothesizing that bacterial DNA integrations may occur, based on what is known about viral DNA integrations and mobile element insertions.
4. Robinson, K. M., K. B. Sieber, **J. C. Dunning Hotopp**. (2013) "A review of bacteria-animal lateral gene transfer may inform our understanding of disease like cancer." PLoS Genet 9(10):e1003877. PMCID: PMC3798261
Summary: In this invited review, we outline our rationale for hypothesizing that bacterial DNA integrations may occur, based on what is known about bacterial DNA integrations in invertebrate animals.
5. **Dunning Hotopp, J. C.** (2011) "Horizontal gene transfer between bacteria and animals." Trends in Genetics, 27(4):157-63. PMCID: PMC3068243.

Summary: Following our 2007 manuscript in Science, there was an explosion of instances of bacterial DNA integrations described in Metazoan genomes. In this invited review, we summarize these transfers and put forth hypotheses to be tested about lateral gene transfer from bacteria to animals.

6. Tallon L. J., X. Liu, S. Bennuru, M. C. Chibucos, A. Godinez, S. Ott, X. Zhao, L. Sadzewicz, C. M. Fraser, T. B. Nutman, **J. C. Dunning Hotopp** (2014) "Single molecule sequencing and genome assembly of a clinical specimen of *Loa loa*, the causative agent of loiasis." BMC Genomics 15:788. PMCID: PMC4175631

Summary: Demonstrating our ability to use state-of-the-art techniques, we were the first group to publish a Metazoan genome using the PacBio platform. We demonstrated that use of this platform yields a more complete genome sequence for far fewer dollars than any of the prior efforts to sequence filarial nematodes. This was done from a single clinical specimen, demonstrating the utility in the clinic.

7. Ioannidis P., Y. Lu, N. Kumar, T. Creasy, S. Daugherty, M. C. Chibucos, J. Orvis, A. Shetty, S. Ott, M. Flowers, N. Sengamalay, L. J. Tallon, L. Pick, **J. C. Dunning Hotopp**. (2014) "Rapid transcriptome sequencing of an invasive pest, the brown marmorated stink bug *Halyomorpha halys*." BMC Genomics 15:738. PMCID: PMC4174608

Summary: We applied novel experimental methodologies to shotgun transcriptome sequencing and novel bioinformatics methodologies to assembly and analysis to yield a comprehensive set of sequences from this invasive pest that is the USDA's number one priority for control today. This analysis included the detection of novel bacterial DNA integrations that may explain why this pest ravages crops.

8. Ioannidis, P, K. L. Johnston, D. R. Riley, N. Kumar, J. R. White, K. T. Olarte, S. Ott, L. J. Tallon, J. M. Foster, M. J. Taylor, and **J. C. Dunning Hotopp**. (2013) "Extensively duplicated and transcriptionally active recent lateral gene transfer from a bacterial *Wolbachia* endosymbiont to its host filarial nematode *Brugia malayi*." BMC Genomics 14:639. PMCID: PMC3849323
9. Budroni, S., E. Siena, **J. C. Dunning Hotopp**, K. Sieb, D. Serruto, C. Nofroni, M. Comanducci, D. Riley, S. Daugherty, S. Angiuoli, A. Covacci, M. G. Pizza, R. Rappuoli, R. Moxon, H. Tettelin, and D. Medini. (2011) "*Neisseria meningitidis* population structure in phylogenetic clades, associated with restriction-modification systems that modulate the effect of homologous recombination." Proc Natl Acad Sci U S A 108(11):4494-9. PMCID: PMC3060241
10. **Dunning Hotopp, J. C.**, M. Lin, R. Madupu, J. Crabtree, S. V. Angiuoli, J. Eisen, R. Seshadri, Q. Ren, M. Wu, T. R. Utterback, S. Smith, M. Lewis, H. Khouri, C. Zhang, N. Hua, Q. Lin, N. Ohashi, N. Zhi, W. Nelson, L. M. Brinkac, R. J. Dodson, M. J. Rosovitz, J. Sundaram, S. C. Daugherty, T. Davidsen, A. Durkin, M. Gwinn, D. H. Haft, J. D. Selengut, S. A. Sullivan, N. Zafar, L. Zhou, F. Benahmed, H. Forberger, R. Halpin, S. Mulligan, J. Robinson, Y. Rikihisa, and H. Tettelin. (2006) "Comparative genomics of emerging human ehrlichiosis agents." PLoS Genet. 2(2):e21. PMCID: PMC1366493

D. RESEARCH SUPPORT

Current support

(Dunning Hotopp) PI NIH New Innovator Award DP2OD007372 09/01/10-07/31/15

"Impact of Bacterial-Animal Lateral Gene Transfer on Human Health"

The major goals of this project involve examining the human health implications that the transfer of bacterial DNA to filarial nematode genomes and the human somatic genome.

(Fraser/Rasko/White) Core Director NIH NIAID U19 AI110820-01 04/15/14-03/31/19

"Host, Pathogen, and the Microbiome: Determinants of Infectious Disease Outcome"

Infectious diseases cause significant morbidity and mortality around the world. This project will use current genomics-based approaches to reveal new information about disease-causing agents and the hosts that they infect to produce results that will enable the design of better diagnostics, anti-microbial compounds and vaccines. Dr. Dunning Hotopp is the Technology Core Director for this U19.

(Fraser/Rasko/White) Co-Project Leader NIH NIAID U19 AI110820-01 04/15/14-03/31/19

"Host, Pathogen, and the Microbiome: Determinants of Infectious Disease Outcome"

Infectious diseases cause significant morbidity and mortality around the world. Dr. Dunning Hotopp will be the Co-Project Leader on a project aimed at using genomics to study filarial nematodes important to human health. The goals of the project include generating a more comprehensive strand-specific transcriptome profile of the life cycle, comparing the genomes of laboratory isolates and wild isolates of *Brugia* spp., and examining the function of filarial genes using RNAi.

Completed support (past three years)

(Fraser) Subproject PI NIH-NIAID HHSN272200900009C 07/01/11-06/30/13

"Whole Genome Sequencing of Emergent Meningococcal Clones of Public Health Significance"

This project focuses on examining the genetic factors involved in the emergence of 142 *N. meningitidis* isolates of public health importance from globally distributed geographic areas. The whole genome sequence will be obtained for multiple pairs of isolates fitting this pattern and will be examined for common attributes that may explain the emergence of disease.

(Fraser) Subproject PI NIH-NIAID HHSN272200900009C 10/01/12-09/30/13

"Sequencing and Comparison of Genomes and Transcriptome Profiles of Human Ehrlichiosis Agents"

This project is aimed at identifying the strain specific pathogenicity factors in a collection of *Ehrlichia chaffeensis* isolates using genome and transcriptome sequencing. Insights gained from this study will be instrumental in increasing our understanding of human ehrlichiosis.

(Fraser) Subproject PI NIH-NIAID HHSN272200900009C 04/01/13-03/31/14

"Sequencing Rickettsiales Genomes"

With this research we will improve our understanding of vector borne diseases caused by bacteria in the order Rickettsiales by obtaining further diverse genomes from the *Anaplasma*, *Ehrlichia*, *Orientia*, *Rickettsia*, and *Neoehrlichia* genera. The analysis is expected to elucidate the molecular mechanisms of host tropism, immune response, and pathogenicity/virulence as reflected in the isolates for sequenced.

(Fraser) Subproject PI NIH-NIAID HHSN272200900009C 06/01/13-03/31/14

"Sequencing and Analysis of *Mansonella perstans* and Its Supergroup F *Wolbachia* Endosymbiont"

Mansonella perstans is a filarial nematode that infects ~114 million people and is widespread in parts of Africa, Latin America, and the Caribbean with an average prevalence of 20% in endemic areas. This project is expected to generate a draft annotated genome sequence of *M. perstans* and its *Wolbachia* endosymbiont genome along with transcriptomics data from the available sample to aid in annotation.

(Dunning Hotopp) PI NSF EF-0826732 09/01/08-08/31/11

"Genome Sequence of *Wolbachia*/Drosophila Lateral Gene Transfer."

The goal of this project is to examine the genomic changes in the fruit fly *Drosophila* due to LGT from the bacterium *Wolbachia*. The DNA of both the fruit fly insert and the resident bacterial genome will be sequenced using extremely high-throughput, next generation sequencing technologies. State-of-the-art bioinformatic techniques will be used to decipher the role and impact of the transferred DNA in flies.

BIOGRAPHICAL SKETCH

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

NAME Maria R. Baer, M.D.		POSITION TITLE Professor of Medicine and Molecular Medicine Director, Hematologic Malignancies	
eRA COMMONS USER NAME (credential, e.g., agency login) [REDACTED]			
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)	MM/YY	FIELD OF STUDY
Harvard University, Cambridge MA	B.A.	1973	Visual/Environmental Studies
Harvard University, Cambridge MA	None	1975	Premedical Curriculum
Johns Hopkins University, Baltimore MD	M.D.	1979	Medicine

A. PERSONAL STATEMENT

I am Director of Hematologic Malignancies, University of Maryland Greenebaum Cancer Center, Professor of Medicine, University of Maryland School of Medicine and Professor of Molecular Medicine, University of Maryland Graduate School. I oversee Hematologic Malignancies patient care and clinical and translational research at the Greenebaum Cancer Center and the Baltimore Veterans Administration Medical Center. I have a long track record of clinical and translational research in acute leukemia, having conducted clinical trials and studies of acute leukemia patient samples for over twenty-five years. I have conducted clinical and translational research in acute leukemia both in single-institution studies and within Cancer and Leukemia Group B, having served as principal investigator of a large frontline study of multidrug resistance modulation in older acute myeloid leukemia patients (CALGB 9720), and of immunophenotyping (CALGB 8361) and drug resistance (CALGB 9760) correlative sciences protocols. I have also served as a Member of the CALGB/Alliance for Clinical Trials in Oncology Leukemia, Leukemia Correlative Sciences, Pharmacology and Experimental Therapeutics Core Committees, Executive Committees and Board of Directors, and serve as the University of Maryland Alliance for Clinical Trials in Oncology principal investigator. The focus of my laboratory research is mechanisms of drug resistance in acute leukemia and novel approaches to overcoming drug resistance. Current work focuses on signal transduction pathways regulating drug resistance in acute myeloid leukemia.

B. POSITIONS AND HONORS

1979-1982	Intern and Resident, Department of Medicine, Vanderbilt University, Nashville, TN.
1982-1986	Clinical Fellow and National Research Service Award Trainee, Division of Hematology, Vanderbilt University
1986-2007	Clinical Staff, Leukemia Service, Roswell Park Cancer Institute, Buffalo, NY
1986-1993	Assistant Professor, Department of Medicine, State University of New York at Buffalo
1993-	Fellow, American College of Physicians
1993-1998	Assistant Research Professor, Department of Physiology, Roswell Park Graduate Division, State University of New York at Buffalo
1993-1999	Research Associate Professor, Department of Medicine, State University of New York at Buffalo
1994-1998	Associate Professor, Department of Medicine, Roswell Park Cancer Institute, Buffalo, NY
1998-2004	Chief, Leukemia Section, Division of Medicine, Roswell Park Cancer Institute, Buffalo, NY
1999-2003	Assistant Research Professor, Department of Experimental Pathology, Roswell Park Graduate Division, State University of New York at Buffalo
1999-2002	Associate Professor (tenured), Department of Medicine, State University of New York at Buffalo
2002-2007	Professor (tenured), Department of Medicine, State University of New York at Buffalo
2003-2007	Associate Professor, Department of Molecular Pharmacology and Cancer Therapeutics, Roswell Park Graduate Division, State University of New York at Buffalo
2006-2007	Professor of Oncology, Roswell Park Cancer Institute, Buffalo, NY
2007-	Professor of Medicine (tenured), University of Maryland School of Medicine, Baltimore, MD
2007-	Director, Hematologic Malignancies, University of Maryland Greenebaum Cancer Center

2009- Professor of Molecular Medicine, University of Maryland Graduate School, Baltimore
 2011-2015 Chartered member, NCI Clinical Oncology (CONC) Study Section
 2014- Co-leader, Experimental Therapeutics Program, University of Maryland Greenebaum Cancer Center

C. SELECTED PEER-REVIEWED PUBLICATIONS (of a total of 200 publications and 294 abstracts)

1. Gourdin TS, Zou Y, Ning Y, Emadi A, Duong VH, Tidwell ML, Chen C, Rassool FV, **Baer MR**. High frequency of rare structural chromosome abnormalities at relapse of cytogenetically normal acute myeloid leukemia with FLT3 internal tandem duplication. In press, *Cancer Genetics*. *Epub 2014 Sep 13*.
2. Bhatnagar B, Duong VH, Gourdin TS, B, Tidwell ML, Cen C, Ning Y, Emadi A, Sausville EA, **Baer MR**. Ten-day decitabine as initial therapy for newly diagnosed acute myeloid leukemia patients unfit for intensive chemotherapy. *Leukemia and Lymphoma* 55:1533-7, 2014. PMID: 24144313
3. Natarajan K, Xie Y, Burcu M, Linn DE, Qiu Y, **Baer MR**. Pim-1 kinase phosphorylates and stabilizes 130 kDa FLT3 and promotes aberrant STAT5 signaling in acute myeloid leukemia with FLT3 internal tandem duplication. *PLoS One* 8:e74653, 2013. PMID: PMC3764066
4. Bhullar J, Natarajan K, Shukla S, Mathias T, Sadowska M, Ambudkar SV, **Baer MR**. The FLT3 inhibitor quizartinib inhibits ABCG2 at pharmacologically relevant concentrations, with implications for both chemosensitization and adverse drug interactions. *PLoS One* 8:e71266, 2013. PMID: PMC3743865
5. Natarajan K, Bhullar J, Shukla S, Burcu M, Chen Z-S, Ambudkar SV, **Baer MR**. The Pim kinase inhibitor SGI-1776 decreases cell surface expression of ABCB1 and ABCG2 and drug transport by Pim-1-dependent and -independent mechanisms. *Biochemical Pharmacology* 85:514-24, 2013. PMID: PMC3821043
6. Scheibner KA, Teaboldt B, Hauer MC, Chen X, Cherukuri S, Guo Y, Kelley SM, Liu Z, **Baer MR**, Heimfeld S, Civin CI. MiR-27a functions as a tumor suppressor in acute leukemia by regulating 14-3-3 theta. *PLoS ONE* 2012;7:e50895. PMID: PMC3517579
7. Sen R, Natarajan K, Bhullar J, Shukla S, Fang H, Cai L, Chen Z-S, Ambudkar SV, **Baer MR**. The novel BCR-ABL and FLT3 inhibitor ponatinib is a potent inhibitor of the multidrug resistance-associated ATP-binding cassette transporter ABCG2. *Molecular Cancer Therapeutics* 11:2033-44; 2012. PMID: PMC3683995
8. Tobin LA, Robert C, Rapoport AP, Gojo I, **Baer MR**, Tomkinson AE, Rassool FV. Targeting abnormal DNA double strand break repair in tyrosine kinase inhibitor-resistant chronic myeloid leukemias. *Oncogene* 32:1784-93, 2013. PMID: PMC3752989
9. **Baer MR**, George SL, Sanford BL, Mrózek K, Kolitz JE, Moore JO, Stone RM, Powell BL, Anastasi J, Caligiuri MA, Bloomfield CD, Larson RA. Escalation of daunorubicin and addition of etoposide in the ADE regimen in acute myeloid leukemia patients 60 years and older: Cancer and Leukemia Group B study 9720. *Leukemia* 25:800-7, 2011. PMID: PMC3821040
10. Xie Y, Burcu M, Linn DE, Qiu Y, **Baer MR**. Pim-1 kinase protects P-glycoprotein from degradation and enables its glycosylation and cell surface expression. *Molecular Pharmacology* 78:310-318, 2010.
11. Lancet J, Gojo I, Burton M, Quinn M, Tighe SM, Kersey K, Zhong Z, Albitar MX, Bhalla K, Hannah AL, **Baer MR**. Phase I, pharmacokinetic and pharmacodynamic study of the heat shock protein 90 inhibitor alvespimycin (KOS-1022, 17-DMAG) administered intravenously twice weekly to patients with acute myeloid leukemia. *Leukemia* 24:699-705, 2010.
12. **Baer MR**, George SL, Caligiuri MA, Sanford BL, Bothun SM, Mrózek K, Kolitz JE, Powell BL, Moore JO, Stone RM, Anastasi J, Bloomfield CD, Larson RA. Low-dose interleukin-2 immunotherapy does not improve outcome of patients 60 years and older with acute myeloid leukemia in first complete remission: Cancer and Leukemia Group B study 9720. *Journal of Clinical Oncology* 30:4934-4939, 2008. PMID: PMC2652081
13. Qadir M, O'Loughlin KL, Fricke SM, Williamson NA, Greco WR, Minderman H, **Baer MR**. Cyclosporin A is a broad-spectrum multidrug resistance modulator. *Clinical Cancer Research* 11:2320-6, 2005.
14. Suvannasankha A, Minderman H, O'Loughlin KL, Nakanishi T, Greco WR, Ross DD, **Baer MR**. Breast cancer resistance protein (BCRP/MXR/ABCG2) in acute myeloid leukemia: Discordance between expression and function. *Leukemia* 18:1252-1257, 2004.
15. **Baer MR**, George SL, Dodge RK, O'Loughlin KO, Minderman H, Caligiuri MA, Anastasi J, Powell BL, Kolitz JE, Schiffer CA, Bloomfield CD, Larson RA. Phase III study of the multidrug resistance modulator PSC-833 in previously untreated patients 60 years of age and older with acute myeloid leukemia: Cancer and Leukemia Group B study 9720. *Blood* 100:1224-1232, 2002.

D. RESEARCH SUPPORT

Current support

[REDACTED]

[REDACTED]

(Perrotti) NIH/NCI 7 R01 CA163800-03 9/1/13-1/31/17

“Role of microRNAs in the regulation of stem cell survival and renewal”

Specific aims are to identify miRs dysregulated in TKI-resistant quiescent CML HSCs that interfere with the BCR-ABL1-Jak2-SET- β -catenin pathway through their canonical and/or hnRNP A1 RNA decoy activities; to assess whether modulation of miR expression impairs *in vitro* and *in vivo* survival and self-renewal of quiescent CML HSCs, and to determine the therapeutic role of modulation of miR expression in eradication of CML by using 2-OMethylphosphorothioate miRs and antagomiRs.

[REDACTED]

(Hoffman) 5P01CA108671 NIH/NCI 9/28/08-6/30/16

Myeloproliferative Disorders Research Consortium

Funding for clinical trials in myeloproliferative disorders

(Bertagnolli) U10 CA180821 NIH/NCI 4/11/13-4/10/15

CALGB/Alliance for Clinical Trials in Oncology

Funding for CALGB/Alliance for Clinical Trials in Oncology clinical trials at the University of Maryland.

(Baer) Veterans Administration Merit Review Award 10/1/14-9/30/18

“Inhibition of Pim kinases in acute myeloid leukemia”

Specific Aims are to determine the mechanisms by which Pim kinase inhibition sensitizes FLT-ITD AML cells to induction of apoptosis by chemotherapy drugs and by FLT3 inhibitors, to optimize scheduling of administration of Pim kinase inhibitors, chemotherapy drugs and by FLT3 inhibitors in FLT3-ITD AML, and to test the effects of Pim kinase inhibitors administered *in vivo* with chemotherapy drugs and with FLT3 inhibitors on FLT3-ITD AML cells and AML stem cells and on normal hematopoietic cells.

BUDGET JUSTIFICATION

University of Maryland, Baltimore, School of Medicine, Institute for Genome Sciences (UMSOM IGS)

A 2% inflation rate is applied to all the costs estimated for this proposal (salary, fringe, travel, supplies and other costs), with the exception of the sequencing cost.

A. SENIOR/KEY PERSONNEL:

A.1. Principal Investigator

Julie C. Dunning Hotopp, Ph.D., Associate Professor, UMSOM IGS, Department of Microbiology and Immunology and Member of Greenebaum Cancer Center:

Dr. Dunning Hotopp has developed an expertise in genomics focused on bacterial DNA integration in animal genomes. She has overseen several ongoing genome sequencing projects including sequencing and annotation of hundreds of bacterial genomes, numerous eukaryotic whole genome and shotgun transcriptome sequencing projects, and RNA-Seq based differential expression analysis of both bacteria and eukaryotes. She directs several projects investigating lateral gene transfer between bacteria and their eukaryotic hosts including research with her NIH New Innovator award focused on analysis of bacterial DNA integration in the human genome. Dr. Dunning Hotopp will oversee the entire project; interact with internal and external collaborators Drs. Maria Baer and Javier Torres, respectively; oversee management of Mr. Kumar; and mentor the graduate students and postdoctoral fellow. Funds are requested for 40% of Dr. Dunning Hotopp's salary (4.8 cal months) in all years of the project.

A.2. Collaborator

Maria Baer, M.D., Professor, UMSOM, Department of Medicine and Director of Hematologic Malignancies, University of Maryland Greenebaum Cancer Center:

Dr. Baer has a long track record of clinical and translational research in acute leukemia. She has conducted clinical trials and studies of acute leukemia patient samples for over twenty-five years both in single-institution studies and within Cancer and Leukemia Group B. Her group sees ~70 patients/year with acute myeloid leukemia and she will provide access to specimens from a subset of consented patients. In addition, she will provide expertise on AML and assist in interpreting the results and preparing publications. Funds are requested for 5% of Dr. Baer's salary (0.6 cal months) in years 2-5 of the project.

B. OTHER PERSONNEL

Nikhil Kumar, M.S., Laboratory Manager, UMSOM IGS

Mr. Kumar has extensive experience investigating the genomics of complex host-pathogen relationships including integration of bacterial DNA in animal genomes. He specializes in experimental identification and verification of genomic and transcriptional changes associated with host-microbe relationships including the development of novel experimental techniques. This has included generating much of the preliminary data for Aim 3 of this project. He will also be responsible for managing the lab, including ordering supplies. Funds for 50% effort (6.0 calendar months) in all years are requested for Mr. Kumar.

Sonia Agrawal, M.S., Software Engineer, UMSOM IGS

Ms. Agrawal is a Software Engineer at the Institute for Genomic Sciences at UMSOM. She has an undergraduate degree in Computer Science from India, a graduate degree in Bioinformatics from Georgia Institute of Technology, and 6+ years of sequencing-related bioinformatics experience. Her expertise includes bioinformatics analysis and tool development, the application of scalable analytical methods, using a high-throughput computing architecture, and cloud computing. Ms. Agrawal will be responsible for developing, testing, and improving the VMs constructed in Aim 1. Funds are requested for 100% effort in year 1 (12 calendar months), 15% effort in year 2 (1.8 calendar months), 10% effort in year 3 (1.2 calendar months) and 5% effort in year 4 (0.6 calendar months).

Postdoctoral Fellow, UMSOM IGS

A postdoctoral fellow will be supported who will be responsible for identifying putative bacterial DNA integrations using publicly available cancer genome sequencing data. He/she will also test the CloVR VM ensuring that the documentation and training videos adequately describe the approach, ensuring that others will be able to use it in a broad research community. He/she will also be responsible for experimental validation

of the results that she/she discovers. He/She will be responsible for analyzing data, preparing manuscripts, and presenting his/her research at national and/or international meetings. Salary support is requested for 100% effort in years 2-5 (12 calendar months).

Graduate Students, UMSOM Graduate Program in Life Sciences (GPILS)

Two graduate students will be recruited from the GPILS programs for which Dr. Dunning Hotopp and/or Dr. Baer are members. These include the programs in Microbiology & Immunology; Biochemistry; and Molecular Medicine, which includes both Cancer Biology and Genome Biology tracks. Collectively these two students will be trained in, and responsible for, computation- and laboratory-based studies in all aims. Funds are requested for 100% (12.0 calendar months) for one student in all years of the project and 100% (12.0 calendar months) for a second student in years 2-5.

C. FRINGE BENEFIT RATE

UMB has a stated fringe benefits rate. Effective July 1, 2012, rates vary dependent upon the specific employee category. Fringe benefit rates for faculty are [REDACTED], staff are [REDACTED], and postdoctoral fellows are [REDACTED] throughout this grant.

D. MATERIALS AND SUPPLIES

D.1. General Lab Supplies

Funds of [REDACTED] in each year [REDACTED] are requested for general laboratory consumables for DNA and RNA isolation, PCR, sequence verification, and other consumables to be used by the graduate students, postdoctoral fellow, and laboratory manager. Consumables include disposable gloves, plastic tips, and disposable tubes as well as kits for DNA purification, oligonucleotides, and molecular biology enzymes, including those needed for CRISPR/Cas9 constructs.

D.2. Acute myeloid leukemia specimens

Bone marrow aspirates will be acquired from the University of Maryland Medical System Pathology Biorepository and Research Shared Service. For Greenebaum Cancer Center members, the first vial costs [REDACTED], with each additionally costing [REDACTED]. We will acquire two vials for the RNA sequencing, DNA sequencing, and culturing at [REDACTED]. We will acquire vials from 20, 24, and 28 patients for [REDACTED] year 2, [REDACTED] in year 3, and [REDACTED] in year 4, respectively [REDACTED].

D.3 Promoter/UTR constructs

Funds of [REDACTED] are requested for construction and experimental interrogation of each promoter/UTR luciferase reporter construct with 5 different bacterial DNA integrations with the described controls. We will test one promoter/UTR region for [REDACTED] in year 1 and two promoter/UTR regions for [REDACTED] in year 2 ([REDACTED]). This includes synthesizing the initial plasmid containing the promoter/UTR; all enzymes for cloning, sub-cloning, and site-directed mutagenesis; kits for plasmid and amplicon purification; acquiring the necessary cell lines; transfection reagents; reporter assay reagents for the single and the dual assay; cell culture media; and cell culture consumables.

E. TRAVEL

Domestic Travel:

Funds of [REDACTED] in each year [REDACTED] are requested for two domestic trips per year for the PI, postdoctoral fellow, and/or graduate students to present results at national conferences, or international conferences hosted within the US.

Foreign Travel:

Funds of [REDACTED] in each year [REDACTED] are requested for one international trip per year for the PI or graduate students to present results at international conferences.

F. OTHER DIRECT COSTS

F.1. Sequencing Costs

Total funds of [REDACTED] over 5 years are requested for DNA and RNA sequencing with Illumina paired end libraries for whole genome sequencing [REDACTED]; Illumina paired end libraries for RNA-Seq [REDACTED]; whole exome capture and library construction ([REDACTED]); and Illumina HiSeq sequencing lanes

██████████.

Year 1 (Total: ██████████)

Library preparation: ██████████

Control samples (Aim 2): 8 libraries for paired end whole genome sequencing, 8 libraries for RNA-seq, 8 libraries for whole exome sequencing using the Agilent v5 capture system

Stomach cancer (Aim 2): 2 libraries for paired end whole genome sequencing, 20 libraries for RNA-seq, 20 libraries for whole exome sequencing using the Agilent v5 capture system

Sequencing: ██████████

Funds are requested for 33 multiplexed Illumina HiSeq runs with 100-bp reads for sequencing the above libraries, such that each RNA-seq sample will be sequenced on one lane; 6 whole exome samples will be sequenced on one lane; and each whole genome sample will be sequenced on 3.25 lanes.

Year 2 (Total: ██████████)

Library preparation: ██████████

AML (Aim 2): 4 libraries for paired end whole genome sequencing, 15 libraries for RNA-seq, 15 libraries for whole exome sequencing using the Agilent v5 capture system

Other samples for validation of results: 4 libraries for paired end whole genome sequencing, 20 libraries for RNA-seq, 20 libraries for whole exome sequencing using the Agilent v5 capture system

Sequencing: ██████████

Funds are requested for 34 multiplexed Illumina HiSeq runs with 100-bp reads for sequencing the above libraries, such that each RNA-seq sample will be sequenced on one lane; 6 whole exome samples will be sequenced on one lane; and each whole genome sample will be sequenced on 3.25 lanes.

Year 3 (Total: ██████████)

Library preparation: ██████████

AML (Aim 2): 4 libraries for paired end whole genome sequencing, 15 libraries for RNA-seq, 15 libraries for whole exome sequencing using the Agilent v5 capture system

Other samples for validation of results: 6 libraries for paired end whole genome sequencing, 25 libraries for RNA-seq, 25 libraries for whole exome sequencing using the Agilent v5 capture system

Sequencing: ██████████

Funds are requested for 40 multiplexed Illumina HiSeq runs with 100-bp reads for sequencing the above libraries, such that each RNA-seq sample will be sequenced on one lane; 6 whole exome samples will be sequenced on one lane; and each whole genome sample will be sequenced on 3.25 lanes.

Year 4 (Total: ██████████)

Library preparation: ██████████

AML (Aim 2): 4 libraries for paired end whole genome sequencing, 15 libraries for RNA-seq, 15 libraries for whole exome sequencing using the Agilent v5 capture system

Other samples for validation of results: 6 libraries for paired end whole genome sequencing, 25 libraries for RNA-seq, 25 libraries for whole exome sequencing using the Agilent v5 capture system

Sequencing: ██████████

Funds are requested for 40 multiplexed Illumina HiSeq runs with 100-bp reads for sequencing the above libraries, such that each RNA-seq sample will be sequenced on one lane; 6 whole exome samples will be sequenced on one lane; and each whole genome sample will be sequenced on 3.25 lanes.

Year 5 (Total: ██████████)

Library preparation: ██████████

Other samples for validation of results: 10 libraries for paired end whole genome sequencing, 40 libraries

for RNA-seq, 40 libraries for whole exome sequencing using the Agilent v5 capture system

Sequencing: [REDACTED]

Funds are requested for 40 multiplexed Illumina HiSeq runs with 100-bp reads for sequencing the above libraries, such that each RNA-seq sample will be sequenced on one lane; 6 whole exome samples will be sequenced on one lane; and each whole genome sample will be sequenced on 3.25 lanes.

F.2. Publication costs

Funds to cover the costs of one publication in a peer-reviewed journal ([REDACTED]) are requested in each year ([REDACTED]).

F.3. ADP/Computer (FTE-based):

IT and bioinformatics support is based on FTE effort. This FTE-based fee allows the institute to maintain a state-of-the-art IT/data analysis infrastructure as described below. Funds in the amount of [REDACTED] in year one, [REDACTED] in year two, [REDACTED] in year three, [REDACTED] in year four, and [REDACTED] in year five are requested for the project for the following services:

IT Charges: The IT charge will support the minimal computational infrastructure required to analyze the data generated by the project. This will cover basic data storage and access to the computational grid, database server, and web server for the duration of the project. This charge will also cover routine data backup and archival to insure against catastrophic data loss.

Informatics Charges: The informatics charge covers the minimal shared informatics services needed for the successful execution of the project. These services include, but are not limited to, access to computational pipelines, maintenance of these pipelines, customization of the pipelines, database administration, web development to create static project pages, internal and external project portal, and custom analysis support. This does not cover the specialized software or analysis required for the project.

Project-specific Charges: In addition to the support above, this project requires an additional 40 TB of storage/year with monthly snapshot back-up and full yearly back-ups to enable the analysis of publicly available genome sequence data described in Aim1 years 2-5. We chose this level of back-up as opposed to more frequent back-ups since many of the pipelines are automated and can be re-launched. As such, the additional costs of more frequent back-ups is not justified. The cost of storage with this level of back-up is [REDACTED]. This storage will be segregated, facilitating FISMA compliance.

F.4. Laboratory Equipment Maintenance (FTE-based):

IGS faculty share a large laboratory facility as described in the equipment and resources page. Funds are requested to cover fees charged to each IGS-PI for each full time laboratory employee (FTLE) in the amount of [REDACTED] to cover the maintenance fees and replacement of equipment in this shared laboratory space [REDACTED]

F.5. Graduate Student Fees and Health Insurance

The graduate student is exempt of fringe benefits. Funds of [REDACTED] are requested to cover graduate student health insurance costs and student fees at the University of Maryland for one graduate student in years 1-5 and another graduate student in years 2-5.

F.6. Equipment

Funds for 10 dual processor grid nodes ([REDACTED]) are requested in year 2 ([REDACTED]). While the ADP/Computer fees listed above provide us with some access to grid nodes, it is not sufficient for the LGTSeek analysis proposed in Aim 1 for years 2-5. These nodes will also be segregated, facilitating FISMA compliance.

G. INDIRECT COSTS

MTDC, On-Campus Rate: The negotiated indirect cost rate for the period 07/01/11-6/30/15 for UMB is [REDACTED] per DHHS agreement dated June 21, 2011. The allocation base is modified total direct costs, consisting of all salaries and wages, fringe benefits, materials, supplies, services, travel and sub-grants and

sub-contracts up to the first [REDACTED] of each sub-grant or sub-contract (regardless of the period covered by the sub-grant or sub-contract). Modified total direct costs shall exclude equipment, capital expenditures, charges for patient care, student tuition remission, rental costs of off-site facilities, scholarships, and fellowships as well as the portion of each sub-grant and subcontract in excess of [REDACTED].

PHS 398 Cover Page Supplement

OMB Number: 0925-0001

1. Project Director / Principal Investigator (PD/PI)

Prefix:

First Name*: Julie

Middle Name: Christine

Last Name*: Hotopp

Suffix:

2. Human SubjectsClinical Trial? ☒ No ☐ YesAgency-Defined Phase III Clinical Trial?* ☐ No ☐ Yes**3. Permission Statement***

If this application does not result in an award, is the Government permitted to disclose the title of your proposed project, and the name, address, telephone number and e-mail address of the official signing for the applicant organization, to organizations that may be interested in contacting you for further information (e.g., possible collaborations, investment)?

☒ Yes ☐ No**4. 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

5. Human Embryonic Stem Cells

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://grants.nih.gov/stem_cells/registry/current.htm. 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.

6. Inventions and Patents (For renewal applications only)

Inventions and Patents*: ☐ Yes ☐ No

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

Previously Reported*: ☐ Yes ☐ No

7. Change of Investigator / Change of Institution Questions

☐ 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

Please attach applicable sections of the research plan, below.

OMB Number: 0925-0001

1. Introduction to Application

(for RESUBMISSION or REVISION only)

2. Specific Aims

2014_NCI_TR01_Specific_Aims.pdf

3. Research Strategy*

2014_NCI_TR01_Research_Plan.pdf

4. Progress Report Publication List**Human Subjects Sections****5. Protection of Human Subjects**

2014_NCI_TR01_Protection_of_Human_Subjects.pdf

6. Inclusion of Women and Minorities

2014_NCI_TR01_Inclusion_of_Women_and_Minorities.pdf

7. Inclusion of Children

2014_NCI_TR01_Inclusion_of_Children.pdf

Other Research Plan Sections**8. Vertebrate Animals****9. Select Agent Research****10. Multiple PD/PI Leadership Plan****11. Consortium/Contractual Arrangements****12. Letters of Support**

2014_NCI_TR01_Torres_Letter_of_Support.pdf

13. Resource Sharing Plan(s)

2014ResourceSharing.pdf

Appendix (if applicable)**14. Appendix**

CHALLENGE, INNOVATION, AND IMPACT STATEMENT

Fundamental paradigm being addressed. We recently presented evidence of bacterial DNA integrating into the somatic human genome in tumor samples [63]. This finding goes against the current dogma that endogenous DNA in the human genome arises only from viral and mitochondrial DNA integrations. Yet this dogma is not grounded in evidence, but instead stems from a lack of evidence, which is in turn due to a lack of examination of vertebrate animal genome sequence data for such integrations. Bacterial sequences are frequently removed *a priori*, without justification, from eukaryotic genome sequencing projects, removing evidence of bacterial DNA integrations (BDIs). For instance, the experimentally well-characterized BDIs in the aphid genome [56,57] were recently removed when the aphid genome was re-curated at NCBI. Exemplifying these barriers, in our own research on BDIs in invertebrate genomes, we often have to fight to include validated BDIs in our NCBI submissions. This is despite a change in dogma in invertebrate genomics over the past decade that now recognizes that such integrations occur, can be functional, and can even be adaptive. This change in dogma is a direct result of our seminal paper that described widespread BDIs from *Wolbachia* endosymbionts to their invertebrate host genomes [23]. This current proposal aims to further characterize BDIs in the human somatic genome in order to rectify this circular reasoning in which BDIs cannot be detected, because they are removed, because they “do not exist,” because they are not detected. At the conclusion of this study, the extent to which BDIs occur in the human somatic genome will be established.

RATIONALE

Using data from the Cancer Genome Atlas, we identified putative BDIs enriched in the 5'-UTRs of known proto-oncogenes in stomach adenocarcinoma (STAD) samples and in mitochondrial genes in acute myeloid leukemia samples (AML) [63]. While they are present in tumors, they may or may not be oncogenic since tumors may merely be permissive to such BDI, or clonal expansion of the tumor may have facilitated detection of such BDIs. Here, we propose further studies aimed at determining the frequency of BDIs, experimentally validating the BDIs, and assessing the functional consequences of BDIs in the human somatic genome.

AIM 1: We will generate bioinformatic resources and use these resources to identify further BDIs in Illumina-based human cancer genome data. We have already developed LGTSeek to detect heterologous DNA integration in a reference genome and LGTView to visualize/interact with the LGTSeek results and available metadata. At the completion of this aim, we will have made these tools available in virtual machines for use by the research community. We will also use LGTSeek and LGTView to analyze further datasets, with an emphasis on those datasets where it will be possible to validate the results.

AIM 2: We will generate genome data for detection of BDIs and experimentally validate the results. Our previously completed analysis supports *Pseudomonas* DNA integrations into the 5'-UTR of known proto-oncogenes in a subset of STAD samples and *Acinetobacter* DNA integrations into the mitochondrial genes in a subset of AML samples. We propose to replicate these results in independent samples and undertake the necessary experimental validation of these findings that was not previously possible. Using human genomic DNA samples with and without exogenous bacterial DNA added, we will assess the prevalence of artifactual chimeras in the sequencing that mimic BDIs. We will also use genome and transcriptome sequencing to assess the prevalence of BDIs in a cohort of *Pseudomonas*-containing gastric samples and a cohort of AML samples. We propose generating these new datasets, where we have access to patient specimens, so we can validate the BDIs.

AIM 3: We will determine the effect of BDIs on transcription *in vitro*. BDIs were detected in the 5'-UTRs of *CEACAM5*, *CEACAM6*, and *CD74* from patients with stomach adenocarcinoma. Bacterial integrations were detected in five patients, with some patients having integrations in multiple genes. Considering that BDIs are integrated into the 5'-UTRs, we hypothesize that the BDIs alter the gene expression in all three cases. Therefore, we seek to recreate the BDIs to test their effects on transcription. Using the raw TCGA sequence data from stomach adenocarcinoma samples, we will develop models of the BDIs in the 5'-UTRs of *CEACAM5*, *CEACAM6*, and *CD74*. Using the models, we will reconstruct the BDIs in the *CEACAM5*, *CEACAM6*, and *CD74* 5'-UTRs and place it downstream of the native promoter to measure the effect on transcription with a luciferase reporter construct. For those cases where the BDI leads to up-regulation of transcription, we will reconstruct the integrations in cell lines using the CRISPR/Cas9 system to examine the effect on cell phenotype.

In summary, our results will provide a strong rationale for the potential use of vaccines targeting these specific bacteria as a way to limit such cancer-inducing integrations. Given that 15-20% of cancers worldwide are linked to infections, more work is needed to understand the role of infectious agents in oncogenesis and the mechanisms underlying that role.

APPROACH

A. SIGNIFICANCE

A.1 How will we attempt to test a novel paradigm?

Recently, we presented evidence of bacterial DNA integrating into the somatic human genome in tumor samples [63]. We hypothesized that such mutations can be oncogenic through mutagenesis of proto-oncogenes and tumor suppressors [63,65]. **Here, we propose a series of experiments aimed at better understanding the extent of bacterial DNA integrations (BDI) in the human genome and their significance.**

The integration of exogenous DNA into the human genome can cause somatic mutations associated with oncogenesis. For example, the insertion of HPV DNA into human chromosomes is the single most important event leading to tumorigenesis in cervical cancer. Clonal integration of HPV-16 or HPV-18, two of the “high-risk” strains [77] occurs in 80-100% of cervical carcinoma tumors [17,21,52]. HPV can integrate into the host nuclear genome in a manner that leads to disruption of the HPV E2 protein [78], which in turn leads to deregulation of the HPV E6 and E7 proteins [67]. Uncontrolled expression of E6 leads to down-regulation of the TP53 pathway, and thus reduction of apoptosis, contributing to cellular transformation to a malignant phenotype [69]. Deregulation of E7 leads to increased cell proliferation through many mechanisms, including degradation of the retinoblastoma protein [10], which controls the duration of the G1 phase of the cell cycle [32]. Fortunately, through childhood HPV vaccination exposure to HPV can be limited, which will also prevent oncogenic integrations of HPV in the somatic genome.

In contrast to viral DNA integrations, the instances and repercussions of BDI into the somatic human genome are less clear. We recently identified putative BDIs enriched in tumors and in the promoters of known proto-oncogenes in stomach adenocarcinoma (STAD) samples [63]. Should these BDIs be validated and found to be associated with tumorigenesis, vaccines targeting specific bacteria could offer great therapeutic potential in limiting such integrations. Therefore, further studies are needed, which we outline below, that are aimed at examining the frequency of BDIs, validating BDIs, and assessing the consequences of these BDIs in the human somatic genome.

A.2 How does our approach significantly differ from the current state of the art in the field?

Currently the state of the art in cancer genomics is to examine primary tumor tissue with some combination of normal adjacent tissue and/or matched germline samples from 200-300 patients [6-8,11-15,35,50]. Samples are subjected to a nearly exhaustive battery of nucleic acid-based techniques including whole genome sequencing, whole exome sequencing, mRNA sequencing, miRNA sequencing, and DNA methylation profiling. These techniques are employed in an effort aimed at assessment of single nucleotide polymorphisms, structural rearrangements, copy number alterations, methylation status, transcriptional differences, and the presence of viral DNA in these samples [6-8,11-15,35,50]. These projects are typically discovery-based initiatives and techniques like unsupervised learning are used to uncover trends that are reported. Such an analysis of STAD revealed a molecular classification that defines four major genomic subtypes of gastric cancer: EBV-infected tumors; microsatellite instable tumors; genomically stable tumors; and chromosomally unstable tumors [8]. This is an extremely powerful technique, and we laud these efforts and the contributions they are likely to make in cancer treatment. However, while viral DNA integrations are examined, other large genomic integrations and rearrangements are not routinely addressed in these marker publications. Nuclear mitochondrial transfers [44] and retrotransposition [45] have been addressed separately. However, BDI has been overlooked, as has other exogenous DNA like DNA from food and parasites. The experiments we propose here aim to address this by focusing on BDI.

A.3 How will our rationale and/or approach overcome existing challenges or barriers in the field?

The biggest barriers in detecting BDI in the human genome arise from: (a) a disbelief that such transfers occur and (b) a disbelief that if such transfer happens that they would ever alter biology. These experiments are designed to enable detection of such transfers, validation of such transfers, and examination of the consequences of such integrations as described below in more detail.

B. INNOVATION

My group has a history of innovation through challenging dogma. When we began our work on BDIs in animal genomes, it was a “fact” that bacterial DNA was not found in animal genomes. Yet, we were the first group to demonstrate that extensive amounts of bacterial DNA can integrate into many animal genomes through lateral gene transfer (LGT), specifically between *Wolbachia* endosymbionts and their arthropod and nematode hosts

[23]. Now LGT is routinely described between many different bacteria and their Metazoan hosts (e.g. [1-3,19,20,23,33,36,38-41,48,49,51,53,55-59,73,79,80,82-84]). Challenging dogma again, we examined human traces for evidence of BDIs in the human somatic genome [63]. Our idea was that if integrations occur in somatic tissues, they will be easiest to detect in the clonally expanding population of cells in the tumor, whether they are oncogenic or not. This appears to be the case, since evidence for BDI is higher in tumor samples than normal samples [63]. The experiments proposed here would continue our research in this area.

C. BACKGROUND

An Open Question: Does Bacterial DNA Integrate in the Human Genome? We hypothesize that BDIs in the genome of cells in somatic tissues (the somatic human genome) may contribute to bacteria-associated diseases like cancer, chronic inflammatory diseases, and autoimmune disease [64,65]. Most research on the integration of bacterial DNA in the human genome has focused not on the somatic human genome, but the inherited human genome, where convincing recent BDI has not been detected [68,72]. One barrier to inheritance of BDIs is the segregation of gametes. Unlike in insects, nematodes, and plants, where bacterial DNA integrates with some frequency, the human germ line is both physically and immunologically well-protected from bacteria, which likely limits integrations that are inherited. However, the somatic human genome is less well protected, yet incredibly important to human health. Some human cells can be bathed in bacteria that outnumber them >10:1, providing an opportunity for mutagenesis by bacterial DNA through BDIs. Such integrations will not become inherited by offspring of the human, but may be propagated through the individual's lifetime, if the cell is not terminally differentiated, is capable of undergoing clonal expansion, and is not destroyed by the immune system.

Overview of Bacteria-Animal Lateral Gene Transfer. In order to understand the potential for BDI in human chromosomes in somatic tissues, as well as its significance, it is important to consider what is known about such integrations in other animals [65]. In invertebrates, such transfers occur as lateral gene transfer (LGT) and are vertically inherited. When we started working on LGT of bacterial DNA to animal genomes a decade ago, the prevailing paradigm was that it was non-existent. Subsequently, instances of bacteria-animal LGT have been observed in multiple invertebrates [1-3,19,20,23,33,36,38-41,48,49,51,53,55-59,73,79,80,82-84], including many such integrations of genes that are functional [1,3,19,20,33,38,39,49,53,55,59,73,79,82,83]. *Hypothenemus hampei*, the coffee berry borer, acquired a bacterial mannanase gene that allows it to exploit coffee berries as a new ecological niche relative to its sister taxa [1]. The invasive brown marmorated stink bug that ravages crops in the mid-Atlantic region is thought to also have several LGTs from bacteria, including a mannanase gene [37]. Several plant parasitic nematodes have acquired cellulases, pectate lyases, and expansin-like proteins from bacteria that allow them to degrade plant material [20,49]. In mealybugs, LGTs from at least three different bacterial lineages have resulted in hybrid biosynthetic pathways, like riboflavin biosynthesis that requires three insect genes that arose via LGT from diverse bacteria [33].

An examination of LGT from bacterial *Wolbachia* endosymbionts reveals that this process is ongoing and occurring frequently. In 2007, most of the genome sequencing projects (8/11) containing *Wolbachia* endosymbiont sequences showed evidence of having recent LGT between the endosymbiont genome and the host chromosome [23]. Since we successfully characterized LGT in all 5 of the hosts that we examined, we estimated that ~70% of sequenced *Wolbachia*-infected hosts may have at least one *Wolbachia*-host LGT [23]. In contrast to the other examples from insects described above, *Wolbachia*-insect LGTs span many hundreds of kilobases of DNA, have no evidence of being functional or beneficial, and have minimal divergence from the likely donor suggesting that they have happened in modern times [2,23,36,58].

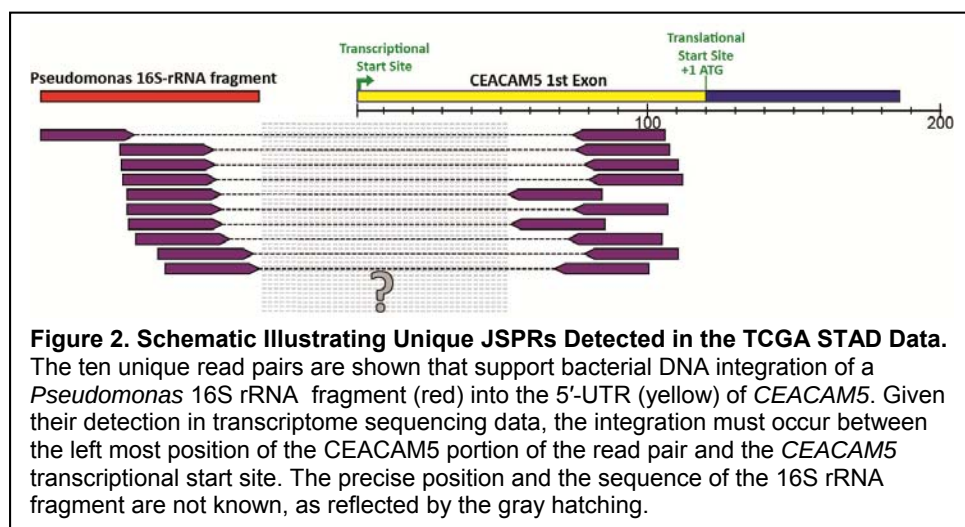
The high prevalence of *Wolbachia*-host LGT is unique and not reflective of all microbe-host relationships. One of the major factors limiting the occurrence of LGT is the proximity of the two organisms, which helps explain the increased prevalence of LGT between obligate intracellular endosymbionts and their hosts. However, other factors are also likely at work, as evidenced in aphids and mealybugs where functional LGTs are detected from facultative reproductive endosymbionts and not the primary mutualistic endosymbionts [33,56]. One influencing factor may be proximity to stem cells. For example, facultative *Wolbachia* endosymbionts can sometimes colonize their host's germ stem cells [26] where a LGT will be inherited by a larger number of progeny when compared to a transfer in a single gamete [65]. In humans, bacterial DNA integration in a stem cell, for example a colon or stomach stem cell, could be propagated over the course of his/her life.

D. PRELIMINARY RESULTS

BDI into Somatic Cells Could Induce Oncogenic Mutations. Our hypothesis is that BDI occurs in human somatic cells and can be detected in tumors due to a clonally expanding population (**Figure 1**) [64,65]. Such integrations that disrupt and mutagenize proto-oncogenes or tumor suppressor genes could provide an

additional mechanism associated with bacteria-associated oncogenesis, in addition to other known mechanisms of bacteria-induced oncogenesis that include chronic inflammation and toxins [64,65]. These integrations could arise through a directed mechanism, as has been observed with *Agrobacterium*-induced crown gall disease in plants. Alternatively, these integrations may merely result from the release of nucleic acids following lysis of bacteria. Integrations could occur specifically at particular integration sites, as is observed with mobile elements, or randomly in a manner analogous to mutations induced by exposure to known carcinogens, like UV radiation and cigarette smoke [64,65].

The availability of large cancer genome datasets like The Cancer Genome Atlas (TCGA) have facilitated testing this hypothesis across a wide variety of cancer types. Through an analysis of this data, we presented evidence supporting BDIs of *Acinetobacter*-like DNA in acute myeloid leukemia (AML) samples and of *Pseudomonas*-like DNA in stomach adenocarcinoma (STAD) samples [63]. We focused on integrations supported by >4 pairs of reads, where a sequencing read matching the human genome was linked to a sequencing read matching only bacterial DNA sequences (e.g. **Figure 2**). The reads had to be unique (not PCR or optical duplicates) such that each read is an independent observation. Integrations were found only from specific members of the microbiome, suggesting that BDIs are limited to a subset of bacteria. There was

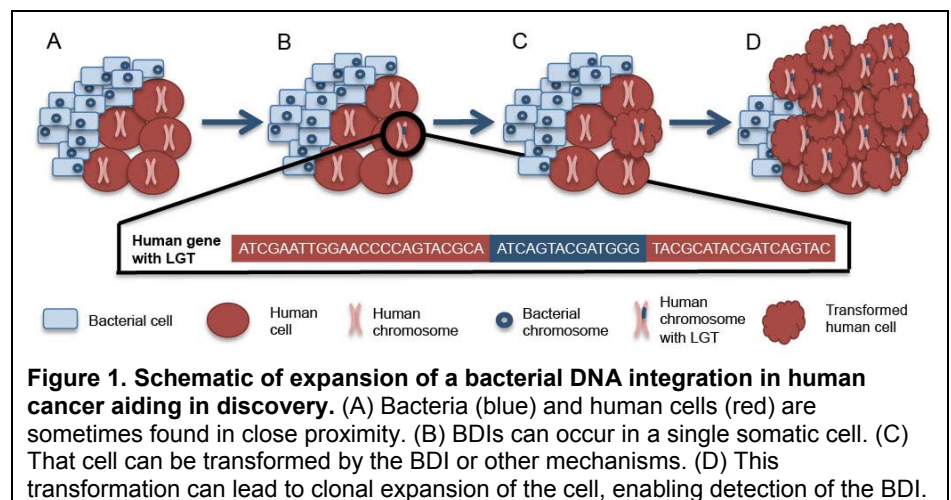


alternatively, occurred as passenger mutations during cancer progression. For instance, cancer cells may become more permissive to mutations during carcinogenesis and thus receptive to integration of bacterial DNA. Regardless, the clonal expansion of tumor cells containing BDIs likely facilitated this discovery.

Here, we propose a three pronged approach aimed at (1) developing bioinformatic resources to detect and present putative BDIs in cancer genomes, (2) sequencing efforts with specific cohorts of gastric cancer and AML samples to identify and validate BDIs, and (3) *in vitro* experimentation to examine the effects of specific BDIs detected previously in STAD.

E. AIM 1 We will generate bioinformatic resources and use these resources to identify further BDIs in Illumina-based human cancer genome data.

We have developed LGTSeek to detect heterologous DNA integrations in a reference genome and LGTView to visualize/interact with the LGTSeek results and available metadata. At the completion of this aim, we will harden both of these tools and make them available in virtual machines (VMs) for use by the research



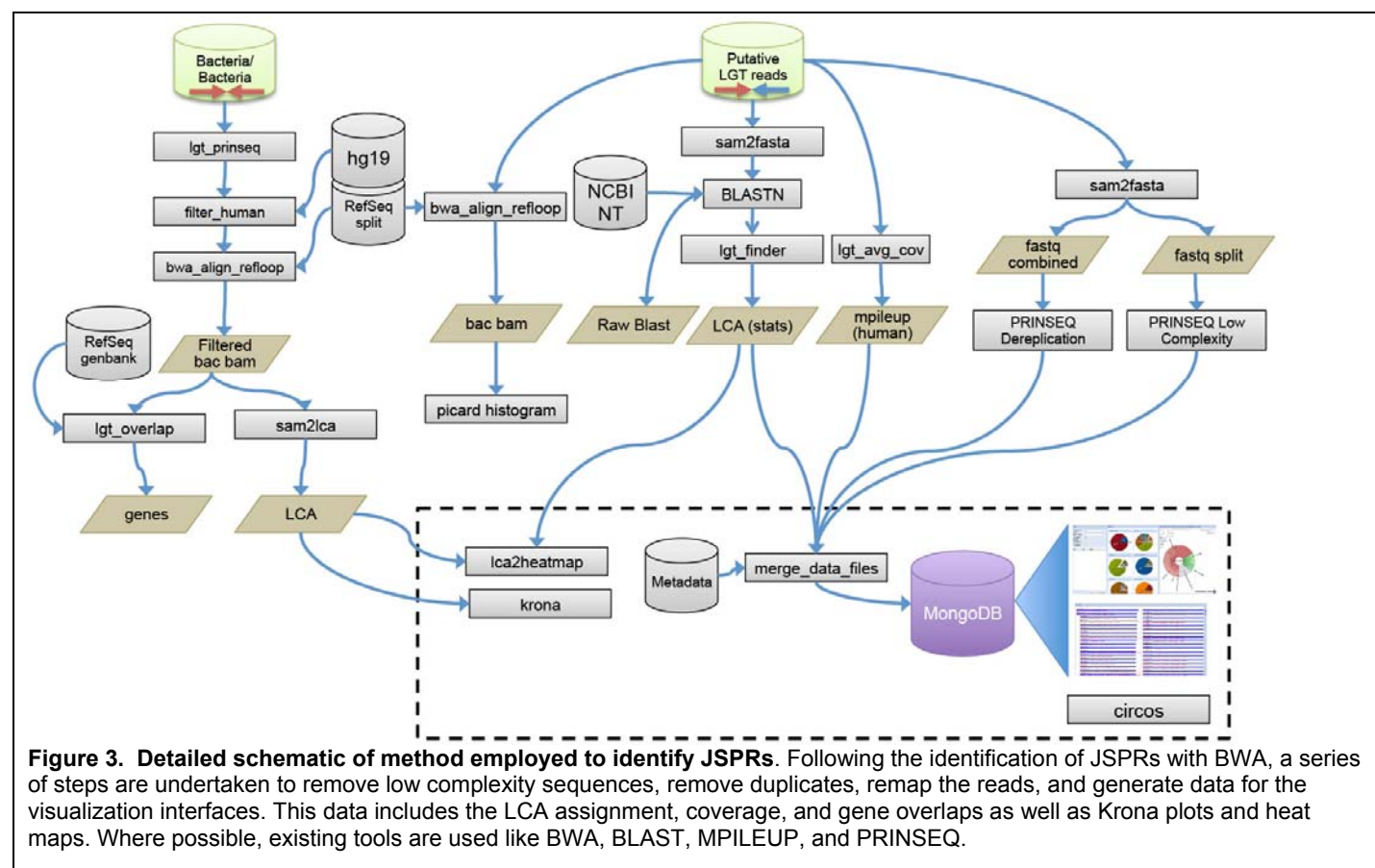
a higher frequency of BDIs in the tumor samples when compared to the available normal samples [63]. The integrations found in STAD samples were in known oncogenes and tumor suppressor genes, while the integrations in AML samples were in mitochondrial genes [63]. In STAD, the same integrations were found in multiple patients, while in LAML the large number of integrations spanned nearly every site in the mitochondrial genome. It was not possible with this analysis to determine if BDIs contributed to carcinogenesis or,

community. We will also apply them to further publicly available cancer genome sequencing datasets and create a public LGTView instance for interrogating these results.

LGTSeek. The LGTSeek pipeline (**Figure 3**) is described in a published analysis of BDI in the human somatic genome that used Illumina sequences that were downloaded from the NCBI Short Read Archive for the 1000 Genomes Project and the TCGA [65]. Junction-spanning paired reads (JSPRs, pronounced “jaspers”) are identified when one read maps only to the donor genome and its mate maps only to the recipient genome. For example, when examining BDIs in tumors, the donor “genome” is a composite genome of all complete bacterial genomes deposited in RefSeq and the recipient genome is the latest release of the human genome. Therefore, we are identifying pairs of reads where one read maps to the human genome and the other read maps exclusively to bacterial sequences. While we have focused our efforts almost entirely on identifying BDIs, donor genomes could be other entities with the potential to induce insertional mutagenesis like the genomes from mitochondria, viruses, and parasites.

This pipeline will be made available to the research community in a CloVR virtual machine (VM), which is a piece of software that encapsulates an entire operating system and can be bundled with pre-installed and pre-configured applications. A single CloVR VM with a graphical user interface (GUI) ensures that users will not have to download and install a complex pipeline and all of the necessary dependencies or know how to use a UNIX-based environment. For this reason, in recent years, a number of VMs have been created specifically for bioinformatics analyses including CloVR [5] at IGS, Qiime [16,42], and Galaxy [9,28]. We will create a CloVR LGTSeek VM that will bundle the LGTSeek pipeline [63] and all of its dependencies including Ergatis/Workflow [62], Grid Engine, BWA [46], NCBI-BLAST [4], SAMTOOLS [47], custom scripts, and libraries. We will use the Open Source configuration management tool Chef (<http://www.opscode.com/chef/>) for efficient installation and maintenance of libraries, tools, and applications. Chef “recipes” are small pieces of code that are used to configure the VM and install necessary applications. These recipes are typically portable in nature and work across most operating systems. By building recipes for all the needed applications and tools, we can rapidly create VMs that are compatible with the major hypervisors KVM, Xen, VirtualBox, and VMPlayer.

With the sequencing of thousands of cancer genomes, the growing appreciation for the prevalence of exogenous DNA integrations in eukaryotic genomes, and the potential relevance of such integrations on human health, the interest in identifying such transfers has grown and will continue to grow substantially over



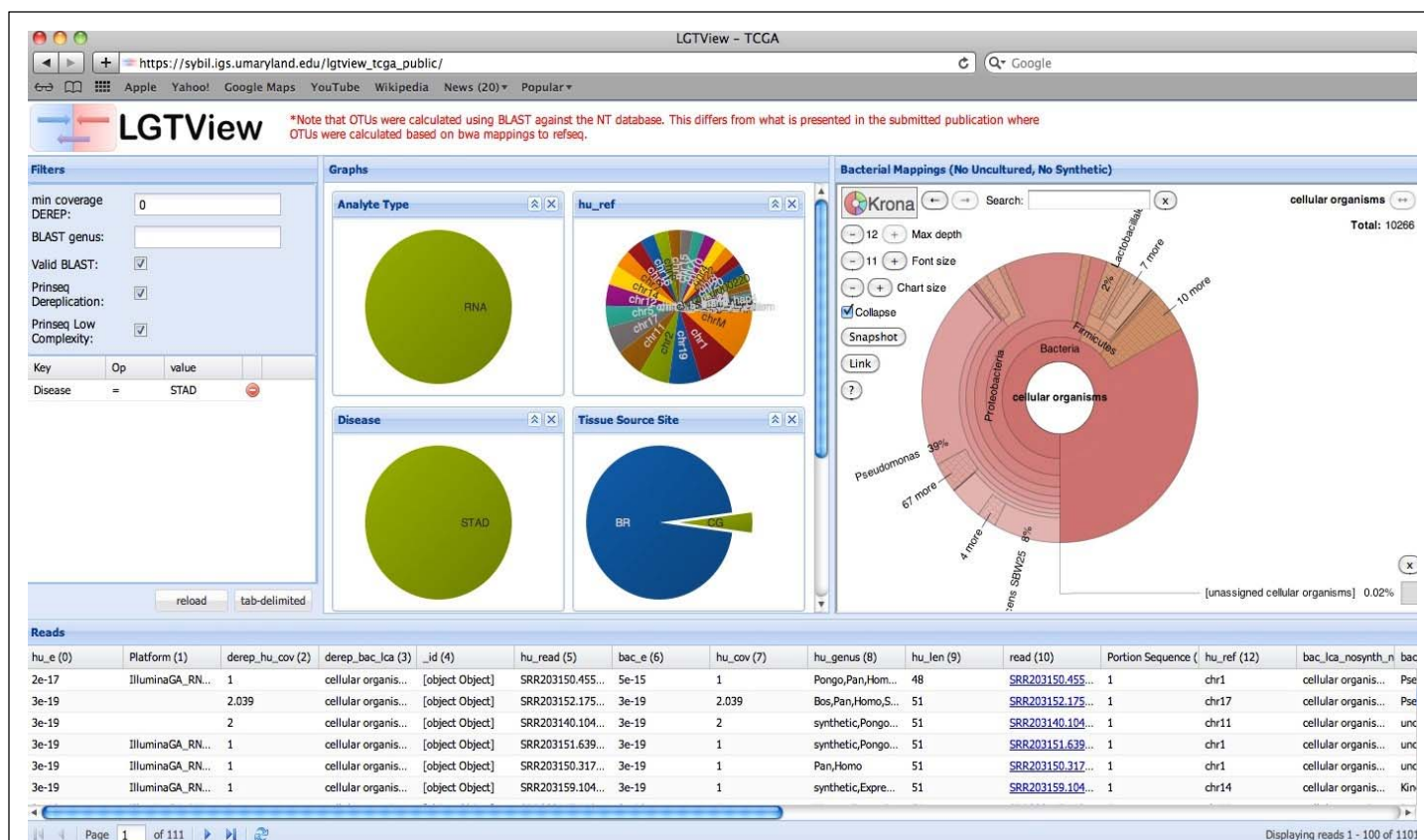


Figure 4. LGTView Interface. The LGTView interface aggregates the results from the pipelines in Objective 1 with metadata. The pipeline results are shown in the table at the bottom with links to TwinBlast while the metadata are summarized in the pie charts. The pie charts are interactive such that clicking on a slice of the pie will limit the data. This is shown here by limiting to STAD as seen in the left panel. After limiting the data, all of the panes of the interface refresh with the new limited dataset.

the coming years. With such growth, we anticipate that we will not be able to collaborate effectively with every group that contacts us to conduct these analyses. In fact, we already are unable to keep up with the requests that we receive. In addition, we believe some standardization and distribution of the pipeline is necessary for results to be easily compared. As such we need to provide a resource that both bioinformatically savvy and naïve individuals can adopt to their particular project and infrastructure. To do this, the VM needs to run on local machines, institutional grids/clusters, and cloud infrastructures. To be sustainable and ensure research integrity, particularly as it relates to human subjects research, it will be FISMA compliant and we will not host any raw or processed data. Instead, the user will maintain complete control over, and responsibility for, their data.

LGTView. Once JSPRs are identified, they need to be further examined, but LGTSeek results can still be a bit obtuse for those that lack the skills to parse large tab-delimited text files. The LGTview tool (**Figure 4**) allows the researcher to query the JSPRs identified by LGTSeek. We have two public LGTview instances that have been available for >1 year (<http://lgt.igs.umaryland.edu/tcg/> & <http://lgt.igs.umaryland.edu/1000genomes/>) that allow the user to search the JSPRs detected in the TCGA and 1000 genomes data, respectively [63]. The results are connected with available metadata, which can include data from the sequencing center as well as the research subjects, and which can be used to limit the data by clicking on the pie charts or using the left most panel. Hyperlinks are used to connect to public information about the study, sample, and sequence as well as to connect to TwinBLAST, a resource we developed that allows for simultaneous review of the blast results for both reads (e.g. <http://tinyurl.com/o3wn35g>).

Using CloVR, we will make the LGTview interface available to researchers to examine their LGTSeek-predicted JSPRs along with their metadata. We will also add further functionality that we have found essential in our many analyses. LGTView will be able to dynamically generate circle figures with Circleator [18] illustrating the linkages between the two genomes for which there is evidence of BDI. It will also generate heat maps using the R package, in much the same way that heat maps are generated in Qiime [16,42]. Both heat maps and circles would be generated upon request following selection of a dataset, for example after limiting the data by clicking on a pie piece. Both of these analyses would be additional analyses viewed in the panel

Table 1. Summary of available colon cancer sequencing projects and the preliminary results from LGTSeek.

Name	Accession	Library	Patients	Sample	Reads Supporting BDI	25 bp regions >2x coverage
1. German	n/a	WGS	1	Tumor	84	97
2. Fusobacterium RNA-Seq	SRP007854	RNA	12	Normal	350	248
				Tumor	922	362
3. Fusobacterium	SRP009542	WGS	9	Normal	2,430	7
				Tumor	3,462	0
4. TCGA	Phs000178	WGS	112	Normal	8,296	174
				Tumor	30,842	374
		WXS	447	Normal	TBD	TBD
				Tumor	TBD	TBD

currently containing the Krona plots [60]. A dropdown menu would allow the user to select between the different analyses, as well as download publication quality SVG files.

Applying LGTView and LGTSeek to further cancer genome datasets. Using these two resources, we propose to examine TCGA and other publicly available cancer genome datasets for evidence of BDI. While we have demonstrated BDIs in STAD and AML, the datasets available at that time were incomplete. Therefore, we first propose to examine the remaining samples and any controls (e.g. adjacent normal tissue and blood samples) in both of those datasets including all exome, whole genome, and transcriptome paired end reads not previously examined. We anticipate that examining all three sequencing platforms, when available, will add an additional level of validation to the results, in the absence of actual experimental validation in these samples. Experimental validation is not possible as samples are not available [63]. Both of these studies are published [8,14] and not under embargo, enabling an analysis of the entire dataset. We also propose a similar analysis to examine the other 7 TCGA publicly available datasets that have a published marker paper [6,7,11-13,15,35,50] and are also not under embargo. An LGTView interface will be created that enables a pan-TCGA analysis of BDI in cancer similar to other pan-TCGA analyses (e.g. [31,44,45]).

E.1 Preliminary results for Aim 1

The LGTSeek pipeline has its roots in the many peer-reviewed analyses we have undertaken to identify bacteria-animal lateral gene transfer using sequencing read pairs from shotgun-based sequencing projects. All of these predictions have been experimentally validated, with the exception of the BDIs in the human genome. Our use of this pipeline and its validation began with an analysis of paired reads in capillary-based shotgun sequencing projects [23], continued with Roche 454-based sequencing [22,51], and now has an emphasis on paired end reads and mate pairs in Illumina-based sequencing projects [36,37,63]. We are even considering how best to do these analyses with the latest platform (PacBio) and recently published on the formation of chimeras in PacBio sequencing and assembly [75] that may impede such analyses in genetically heterogeneous samples.

Demonstrating the utility of applying LGTSeek to further datasets, we have analyzed four colon cancer genome sequencing datasets for evidence of BDI (**Table 1**). The presence of numerous >25-bp regions with >2X coverage across multiple datasets suggests the presence of BDIs warranting a more comprehensive evaluation. This is not surprising since colon cancer is one of the most likely locations for BDIs considering the prolific microbiome in the human colon.

E.2 Expected outcomes, potential pitfalls, and alternative approaches for Aim 1

At the completion of Aim 1, we expect to have the LGTSeek and LGTView pipelines available in CloVR that allow for the detection of BDI with Illumina sequence reads from cancer genomes and enable inspection and analysis of the results. We will test these pipelines using existing data, recreating published analyses, as well as undertaking new analyses of data. All of these interfaces will be made available via the LGTView web page and manuscripts describing the results.

F. AIM 2 We will generate genome data for detection of BDIs and experimentally validate the results

Our previously completed analysis of RNA-Seq data from gastric tumors and AML samples from TCGA resulted in the identification of novel BDIs associated with specific microbiome signatures [63]. Evidence

supports *Pseudomonas* DNA integrations into the 5'-UTR of known proto-oncogenes in a subset of STAD samples and *Acinetobacter* DNA integrations into the mitochondrial genes in a subset of acute myeloid leukemia samples [63]. Unfortunately, for this release of data, which was a subset of data from ten different tumor types and was released to the NCBI SRA, there were no AML or STAD normal samples for comparison. Even more unfortunately, the results could not be experimentally validated due to an inability to acquire the necessary samples. Therefore, it is extremely important to replicate this sequencing in independent samples and undertake the necessary validation of these findings.

Assess the prevalence of chimeras in mock samples that mimic BDIs. One concern about detecting BDIs arises from artifact chimeras that occur in Illumina sequencing during library construction and in clustering on the flow cell. In the TCGA data, we found identical results in two sequencing channels made from a single library suggesting that the results do not stem from artifacts related to flow cell clustering. But all of the sequencing was done from a single library such that artifacts could occur during this step. For that reason we discarded all reads with characteristics of PCR duplicates, requiring reads to start and end in different locations, ensuring that they are independent observations. However, further experiments are needed to understand the formation of chimeras in Illumina sequencing, at least as it relates to the detection of BDI. Therefore, we will collect DNA and RNA samples from two healthy volunteers (n=2). Aliquots will be created of both where one is maintained as isolated and the other has the corresponding nucleic acids added from either *Pseudomonas* or *Acinetobacter* isolates (n=2). From each sample, two libraries will be constructed and sequenced on separate Illumina channels (n=2; n=2). This will be done for transcriptome, whole exome (WXS), and whole genome sequencing (WGS) (n=3). The result will be a comprehensive set of sequencing data (n=64) that allows us to compare the formation of bacteria-human chimeras with respect to the human source, bacterial source, library construction, and flow cell construction for three different common types of libraries constructed on the most widely used sequencing platform today.

Assess the prevalence of putative BDIs in *Pseudomonas*-containing gastric samples. While specific BDIs were identified in a subset of STAD samples, they were only identified in samples with *Pseudomonas* reads, of which >50% had evidence of BDI. This suggests that studies aimed at examining BDI in STAD should focus on sequencing samples containing *Pseudomonas* DNA. Putative BDIs were only identified in TCGA samples from collection centers in two specific geographical regions, suggesting that such patients may not be identified everywhere. A recent analysis of the microbiome of stomach cancer in a cohort from Mexico

Table 2. *Pseudomonas* spp. score of gastric cancer samples that each have 2.5 µg of DNA and RNA available for sequencing

	ID	<i>Pseudomonas</i> score
Tumor Tissue	1CG-002	233
	3CG-046	288
	3CG-051	172
Adjacent Tissue	1CG-002	N/A
	3CG-046	N/A
	3CG-051	N/A
Lesion Tissue	4GB011 (IM)*	86
	4GB014(NAG)*	93

* IM=intestinal metaplasia; NAG=non-atrophic gastritis

revealed that *Pseudomonas* DNA was more abundant in gastric cancer samples than non-atrophic gastritis samples [6]. Therefore, this Mexican cohort provides an opportunity to independently validate our observations made with the TCGA data. Therefore, we have established a collaboration with Dr. Javier Torres (**Letter of Support**), who led this study, in order to sequence DNA and RNA from samples with a higher abundance of *Pseudomonas* DNA recovered from the microbiome (**Table 2**). We propose to sequence eight samples including five samples that have been shown to contain *Pseudomonas* DNA and three control samples. Based on our prior results analyzing the TCGA data where >50% of *Pseudomonas*-containing samples had BDI, we anticipate identifying BDIs in three samples. For all eight samples (**Table 2**), the transcriptome will be sequenced following polyA-selection and the exome will be sequenced

following capture with the Agilent SureSelect Human All Exon version 5 kit. Following analysis of the transcriptome and exome with LGTSeek, two samples will be selected for whole genome sequencing and analysis. In all cases, two libraries will be constructed at separate times and loaded on separate channels to allow for a comparison of two independent datasets as described above. We will have confidence in putative BDIs identified in both libraries relative to those found in only a single library. We will examine the properties of these JSRs and the corresponding putative BDIs using numerous tests and analyses including those listed in **Table 3**. We will further validate the DNA integrations using PCR amplification with sequence validation.

Assess the prevalence of putative BDIs in *Acinetobacter*-containing acute myeloid leukemia samples.

All of the AML samples that were sequenced were from a single cohort centered on St. Louis, MO, USA. The presence of putative BDIs were not associated with any particular type of sample relative to the available metadata. However, 58% of AML samples sequenced had evidence for BDI, with 17% having regions with >4X

Table 3. Analysis proposed to evaluate bacterial integrations of gastric cancer and AML samples.

1. Calculate the number of unique: human read pairs microbiome read pairs bacteria-human JSPRs	4. Compare coverage of JSPRs to: consistency of JSPRs within patients, tissue type, sequencing libraries and runs previous integrations we identified in STAD and AML [63]
2. Determine coverage across JSPRs human read bacterial read	5. Compare JSPRs between samples by: clinical features library sequencing lane
3. Compare the number of JSPRs to the number of genomic positions for human: genes protein coding regions intergenic regions exons (coding) exons (non-coding; e.g. UTRs) introns promoters mobile elements fragile sites	6. Compare the number of JSPRs to the number of genomic positions for representative bacteria: genes protein-coding regions intergenic regions rRNA
	7. Compare number and position of bacteria-human JSPRs to the number of other JSPRs: EBV for STAD numt for AML

coverage of bacteria-human JSPRs on the human reference genome. These included samples with >100,000 JSPRs supporting BDIs. We are seeking to reproduce these results in a cohort of 45 patients centered in Baltimore, MD, USA, through a collaboration with Dr. Maria Baer, a clinician scientist and fellow member of the Greenebaum Cancer Center at the University of Maryland School of Medicine. If the Baltimore cohort mirrors the St. Louis cohort, we should identify BDIs in 26 samples and highly supported BDIs in 9 samples. For all 45 bone marrow aspirate samples, the transcriptome will be sequenced following polyA-selection and the exome will be sequenced following capture with the Agilent SureSelect Human All Exon version 5 kit. We will also conduct 16S rRNA assays to measure the microbiome composition from the isolated genomic DNA and use a small aliquot of the bone marrow aspirate to culture for *Acinetobacter* on several non-selective media including Leeds agar and sheep's blood agar plates. Following analysis of the transcriptome and exome with LGTSeek, two samples will be selected for whole genome sequencing and analysis. All subsequent experiments will be the same as those for STAD including the use of two independently constructed and sequenced libraries.

F.1 Preliminary results for Aim 2

We have already obtained the necessary STAD specimens from Dr. Torres. For AML, ~70 untreated AML patients are seen per year at the University of Maryland Medical Center, such that collection of 15 bone marrow aspirates per year is feasible. Samples are frozen in vials containing 10^7 cells, and 50% recovery is expected. This is consistent with the number of cells and recovery levels for samples used in the TCGA AML genome study [14]. As a prior NIAID-funded Genome Sequencing Center for Infectious Disease and now supporting a technology core led by PI Dunning Hotopp for an NIAID Genome Center for Infectious Disease, our genome sequencing center has sequenced thousands of genomes, including many projects that include whole genome sequencing, whole exome sequencing, and/or RNA-seq with human samples. PI Dunning Hotopp's research group has analyzed more than a thousand samples with the LGTSeek pipeline. Therefore, we do not anticipate any challenges in accomplishing this aim.

F.2 Expected outcomes, potential pitfalls, and alternative approaches for Aim 2

While we are proposing to sequence two datasets that reasonably should be able to reproduce our prior results, until they are sequenced we cannot ensure that there will be BDIs to examine. While this is a risk, we believe it is a reasonable risk given our prior findings and the potential implications if the results are duplicated in a system where they can be verified. To further mitigate risk, we will use these sequences to look at other characteristics of these tumors. For example, recent studies on STAD have suggested that classification systems used for gastric cancers have little to no predictive ability for treatments and success rates. As such, new classifications based on genomic analyses have been put forth [76]. We wish to see if the already characterized bacterial composition of the microbiome [6] correlates with these molecular classifications.

G. AIM 3 We will determine the effect of BDIs on transcription *in vitro*.

Our previous analysis of the STAD genomes revealed BDI into the 5'-UTR of *CEACAM5*, *CEACAM6*, and *CD74*. BDIs were detected in five patients, with some patients having integrations in two genes and some integrations found in two patients. The greatest number of these JSPRs were in *CEACAM5* (or *CEA*). Other JSPRs were identified in *CD74* (or *HLA-G*) and *CEACAM6* (or *NCA*). The *CEACAM* family of proteins has

collectively been shown to have a role in cell adhesion, differentiation, and apoptosis [24,25,34,61,66,74,81] as well as bacterial pathogen adhesion and colonization [50,54]. *CD74* is a glycoprotein with several immunological functions and may have a role in cancers, where it acts as a signaling molecule for cell proliferation (for review, [8]). It has also been found to facilitate adhesion of *Helicobacter pylori* in gastric epithelial cells [7]. Considering that BDIs are integrated into the 5'-UTR of all three of these genes, we hypothesize that the BDIs altered gene expression in all three cases. Consistent with this, in all cases, the patients with putative BDIs had up-regulation of the corresponding gene relative to the entire set of tumor samples studied [63]. Intriguingly, increased expression of these genes would promote bacterial adhesion and colonization as well as have the potential to be oncogenic.

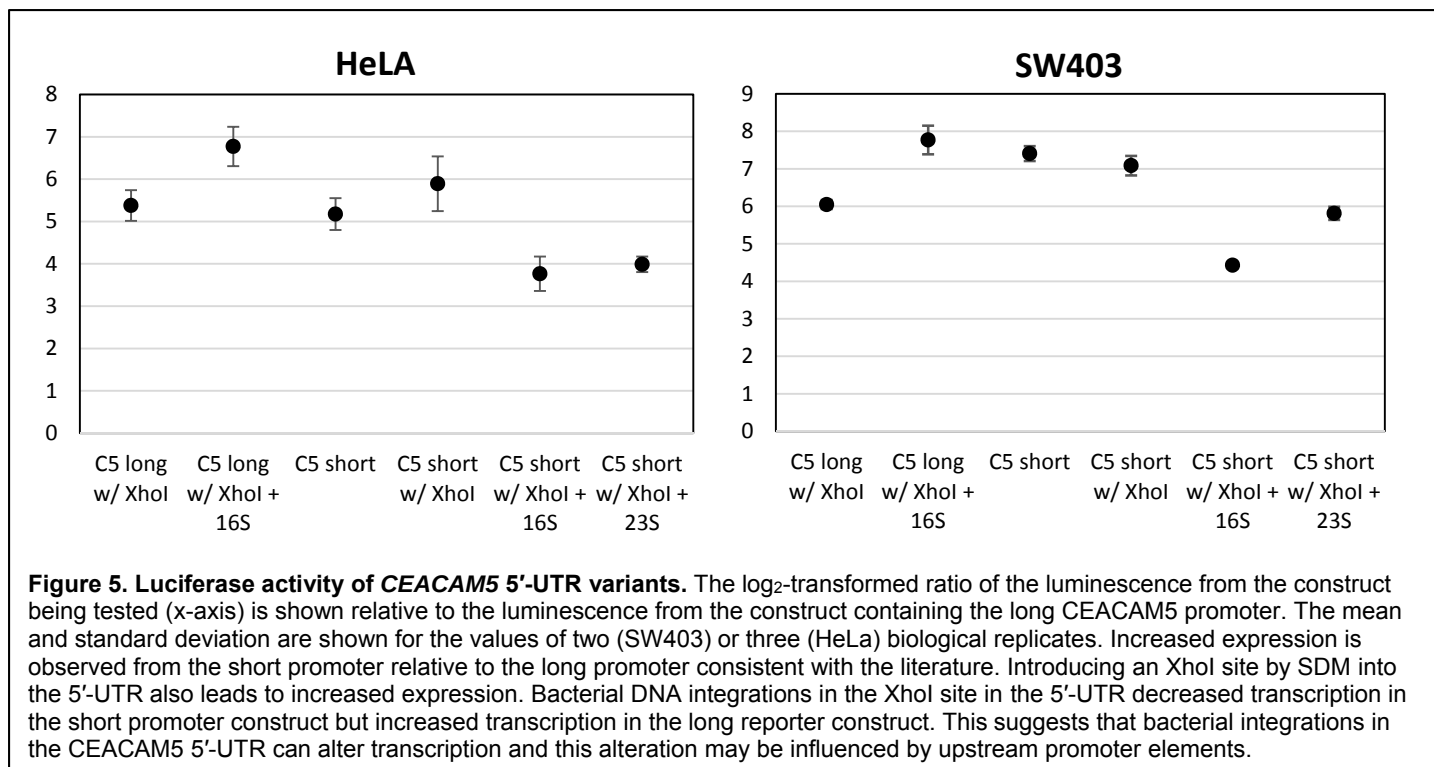
The role of the 5'-UTR in all three of these genes is poorly understood. Based on analysis of mutations near the transcriptional start site of other genes, we hypothesize that the BDIs may disrupt a transcriptional pausing site [43] or repressor-binding site [27,30,71]. In this aim, we seek to recreate the BDIs to test their effect on transcription.

Modelling. Using the identified bacteria-human JSPRs, we will develop models of the BDIs in *CEACAM5*, *CEACAM6*, and *CD74*. Thus far, we have been unable to assemble these BDIs likely owing to the tumor heterogeneity, proximity of the integration to the poorly sequenced 5'-ends of transcripts, and the short read length in these studies. Instead, we will use the JSPRs detected to model the BDI location and sequence using two methods, namely the minimum average difference (AD) and the Jensen-Shannon distance (JSD). For both methods, the consensus sequence is determined for the bacterial portion of the JSPRs and the human portion of the JSPRs. The model estimates the distance between these two consensus sequences, or X . For the AD model, the average difference was calculated between the median insert size of the library and the distance between each JSPR for each potential value of X , such that $0 \leq X \leq 100$. The best AD model(s) is(are) selected as the model(s) with the minimum average distance between reads. The Jensen Shannon distance is calculated between the frequency distribution of the insert size of the human read pairs and the frequency distribution of the JSPRs for each value of X using the Kullback-Leibler divergence, such that $0 \leq X \leq 100$ with 1000 bootstraps. The local minimum was selected for X , assuming it was supported by the bootstrapping. The value of 100 was chosen based on the insert size distribution of these specific libraries; longer insert size libraries might require a higher value to be chosen.

Effect on transcription. Using the models above, we will reconstruct the integrations of bacterial DNA in the *CEACAM5*, *CEACAM6*, and *CD74* UTRs in reporter plasmids to measure the effect on transcription. More specifically, we will reconstruct the BDI into the cloned promoter/UTR using a restriction site introduced by site-directed mutagenesis (SDM). We will test the full promoter with and without the restriction site and with all of the BDIs identified, in case characteristics of the bacterial DNA (e.g. secondary structure) influence transcription. One reason to expect that this might be the case is that the integrations arise from positions 345-455 of the 16S rRNA and positions 1300-1415 and 1626-1701 in the 23S rRNA that contain numerous stem-loop structures that have the potential to alter transcription. For this reason, the effect of these BDIs in the promoter/UTR constructs will also be compared to (a) a construct containing a sequence of the same size that is not expected to be able to form a stem-loop structure, (b) a construct with point mutations that destabilize the stem-loop structure, and (c) a construct with compensatory mutations that restore the stem-loop structure.

While there is a great deal of literature on the promoters for these genes, little is known about the function of the 5'-UTR. However, when there is existing knowledge about the promoters, we will include constructs analogous to those tested previously as positive controls. For *CEACAM5*, we will include a shorter promoter that removes a repressor binding site and was previously shown to have higher expression than the full-length promoter [29,70]. We will use SW403 and HeLa human cell lines, which have been proposed to have high and low levels of native *CEACAM5* expression [29,70]. The reporter plasmids will be transfected into SW403 and HeLa cells using Lipofectamine and Polyplus, respectively. *CEACAM6* will be examined in LR-73 and SW403 cells, which have previously been shown to have a 10-fold difference in expression of this promoter. While *CEACAM5* and *CEACAM6* are paralogs, *CEACAM6* shares only three of five promoter elements found in *CEACAM5* and lacks the repressor that is removed when generating the short *CEACAM5* promoter. However, given our preliminary results below, we will include long and short versions of the *CEACAM6* promoter in case a cryptic silencer can be revealed with these experiments.

To measure transcript abundance, we will use both single and dual luciferase reporter assays. With the single plasmid assay there is no control for the transfection efficiency or number of viable cells; as such a difference in luminescence can occur if a promoter construct effects transfection or viability. The dual luciferase assay uses transfection of a second control plasmid containing the Renilla luciferase gene driven by the SV40 viral



promoter plasmid to measure the transfection efficiency and cell viability. However, that SV40 promoter construct can have trans-effects that require the optimization of the amount of control plasmid transfected. Despite optimization, we have already detected an effect in preliminary experiments. Therefore, and because BDI into the human genome is so controversial, we believe that performing both assays is needed to ensure confidence in the accuracy of the result. Congruence between the two will lend support for the results obtained.

CRISPR/Cas9. When the BDI is found to up-regulate transcription, we propose to make similar constructs in nuclear copies of the genes in cell lines using the CRISPR/Cas9 system. We will characterize transcriptional and phenotypic changes in these cells.

G.1 Preliminary Results for Aim 3

As a proof-of-principle, the predicted BDIs in CEACAM5 were tested in the transcription assay. After optimizing the maintenance, seeding, and transfection protocol of the HeLa and SW403 cell lines, single transformant luciferase assays were completed with HeLa cells transfected with the reporter constructs containing (a) the long 1.1 kbp CEACAM5 promoter with 5'-UTR, (b) the long promoter with an XhoI site introduced by SDM in the 5'-UTR, (c) the long promoter with a 16S rRNA fragment introduced into the XhoI site, (d) a ~400 bp truncated CEACAM5 ("short") promoter that removes an upstream repressor site, (e) the short promoter with an XhoI site introduced in the 5'-UTR by SDM, (f) the short promoter with the 16S rRNA integration in the XhoI site, and (g) the short promoter with a 23S rRNA fragment integration into the XhoI site (**Figure 5**). The mean and standard deviation was calculated for each of two (SW403) or three (HeLa) biological replicates using the log₂-transformed value of the ratio of the luminescence of each sample relative to the luminescence of the native long promoter and 5'-UTR. Consistent with prior published results, the longer promoter is transcribed more poorly than the short promoter. Transcript abundance is also increased upon introduction of the XhoI site by SDM. Transcript abundance was decreased after the introduction of the BDIs at the XhoI site in the short promoter/UTR, but is increased after introduction of the BDI in the long promoter/UTR. This suggests that the BDI alters transcription, and that the direction of that alterations is related to differences between the long promoter and short promoter.

G.2 Expected outcomes, potential pitfalls, and alternative approaches

At the completion of this aim, we expect to demonstrate that the UTRs with the BDIs have altered gene expression relative to their native promoters. It is unlikely that DNA can integrate into the UTR without disrupting it, so down-regulation of transcription would not be surprising. However, with the preliminary results we already observed up-regulation of transcription. We are particularly interested in mutations that lead to up-regulation of transcription as it suggests that the BDIs have converted these proto-oncogenes into oncogenes. This can be examined in cell cultures by introducing the mutations using the CRISPR/CAS9 system. A

phenotype may be difficult to assess in many cell lines, which are derived from cancer samples. Therefore, it may be necessary to test them in primary cell lines or stem cells. Given that both of these proteins also have a role in bacterial colonization, we would like to test these cell lines for bacterial cell adhesion and invasion. We anticipate a change in bacterial cell adhesion based on the literature that suggests that these proteins are involved not only in carcinogenesis, but also bacterial adhesion. Should bacterial adhesion be increased, this may suggest that bacterial DNA integrations provide a benefit to the bacterial donor, which may indicate that these transfers are not random, which should be examined further in the future.

H. Summary

Over the past decade, we have discovered and validated bacterial DNA integrations in multiple animals. Thus far in humans, we have only been able to computationally identify bacterial DNA integrations in the somatic genome. The research supported by this proposal emphasizes identifying such integrations in samples where they can be validated and interrogated experimentally in order to understand their function and relevance.

APPROPRIATENESS FOR THE TRANSFORMATIVE RESEARCH INITIATIVE

Why is the proposed research uniquely suited to the goals of the Transformative Research Awards initiative, rather than a conventional research grant application?

The biggest barriers in obtaining funding to detect BDI in the human genome arise from: (a) a disbelief that such transfers occur and (b) a disbelief that if such transfers occur that they would ever alter biology. In the current funding climate, it is understandable that study sections are risk-averse, investing in programs that will surely yield an outcome, even if that outcome may be incremental and not transformative. However, the NIH HRHR programs are designed as a counter balance, providing funding for transformative research. Therefore, we look at the transformative R01 program, review criteria, and reviewers as providing an opportunity to test these unconventional hypotheses and are applying to this program instead of submitting an R01 to a standing study section. It saddens me to say that the **current paradigm on bacterial DNA integration in the human genome is based largely on our ignorance**. Essentially, the argument is that since we have never validated a BDI in the human genome, they do not exist. Yet, besides the research in my group, scientists do not even look for such integrations, and when my group found such integrations, we could not be given access to samples for validation. Therefore, it does not seem unreasonable to thoroughly test this hypothesis. Should the scientific process reveal that such transfers cannot be identified and validated within a reasonable number of datasets, this is an important finding that will be added to the scientific literature. If, instead, the scientific process reveals that there are such transfers that can be validated and are oncogenic, therapeutics can be developed that limit the interactions that lead to these BDIs, like vaccines. Therefore, we propose three efforts here aimed at closing this knowledge gap by: (a) enabling us and others to detect BDIs in existing genome sequencing data, (b) facilitating further sequencing of samples from cohorts deemed as more likely to have BDIs, and (c) testing the functional consequences of integrations that are detected. Such a comprehensive analysis will allow us to better understand the extent and significance of BDIs.

How does the proposed research significantly differ from mainstream science being done in my laboratory and in other laboratories?

Currently, similar research is supported in my laboratory by the NIH New Innovator Award that ends next summer. At that point, as it currently stands, none of the funded research in my group will be on bacteria-animal lateral gene transfer. Such transfers in invertebrates are now commonly reported by the corresponding genome sequence project. We have submitted numerous proposals aimed at understanding the mechanism and frequency of such transfers from *Wolbachia* endosymbionts, but since it appears that they are not functional, there has not been the enthusiasm for these proposals needed in this funding climate. The other research in my lab is focused on more traditional genomics-based studies on infectious disease. We were funded through an NIAID genome sequencing contract to sequence 140+ *Neisseria meningitidis* genomes, 10 *Ehrlichia chaffeensis* genomes, and 40+ genomes from the order Rickettsiales. In addition, this contract funded the sequencing of 30+ *Ehrlichia chaffeensis* and host transcriptomes from multiple bacterial strains in vertebrate and invertebrate hosts. We have been included on numerous proposals as subcontracts for sequencing and analysis for further bacterial or invertebrate genomics projects. In addition, I currently lead the Technology Core and co-lead a parasite genomics project funded under an NIAID U19 on Genome Sequencing for Infectious Disease. The parasite genomics focuses on filarial nematodes with an emphasis of identifying novel drug targets using genome sequencing, transcriptome sequencing, and RNAi.

TIMELINE

PI Dunning Hotopp will oversee all aspects of the project in all years.

Year 1

Aim 1: Put LGTSeek and LGTView in CloVR (Agrawal)

Aim 2: Experiments assessing chimera formation on the Illumina platform (Graduate Student #1)
Stomach adenocarcinoma sequencing and analysis (Graduate Student #1)

Aim 3: Testing modelled bacterial DNA integrations in reporter constructs (Graduate Student #1 & Kumar)

Year 2

Aim 1: Continued improvement of LGTSeek and LGTView in CloVR (Agrawal)
CloVR testing and putative BDI discovery (Postdoctoral fellow)

Aim 2: AML sequencing and analysis (Graduate Student #2 & Baer)

Aim 3: Testing bacterial DNA integrations in reporter constructs (Graduate Student #1 & Kumar)

Year 3

Aim 1: Continued improvement of LGTSeek and LGTView in CloVR (Agrawal)
CloVR testing and putative BDI discovery (Postdoctoral fellow)

Aim 2: AML sequencing and analysis (Graduate Student #2 & Baer)

Aim 3: Testing bacterial DNA integrations in cells lines (Graduate Student #1 & Kumar)

Year 4

Aim 1: Continued improvement of LGTSeek and LGTView in CloVR (Agrawal)
CloVR testing and putative BDI discovery (Postdoctoral fellow)

Aim 2: AML sequencing and analysis (Graduate Student #2 & Baer)

Aim 3: Testing bacterial DNA integrations in cells lines (Graduate Student #1 & Kumar)

Year 5

Aim 1: CloVR testing and putative BDI discovery (Postdoctoral fellow)

Aim 2: AML sequencing and analysis (Graduate Student #2 & Baer)

Aim 3: Testing bacterial DNA integrations in cells lines (Graduate Student #1 & Kumar)

Continual reassessment

Our primary emphasis at this time is on experimental validation of our computational results. We will continue to pursue collaborations to gain access to specimens for validation when possible. However, thus far we have encountered insurmountable barriers for obtaining the samples containing the BDIs that we have detected. Therefore, most of the resources we are requesting are for new sequencing efforts aimed at specimens from cohorts where validation is possible. We have also included extra sequencing in the event that samples or cohorts become available in the future that are worth examining for BDI. For instance, upon examination of the TCGA colon cancer sequencing, we may find it worthwhile to sequence colon cancers from the University of Maryland Medical Center Biorepository in order to identify samples for validation. Because the primary roadblock to obtaining funding from more traditional mechanisms is the current lack of validation of our computational results, the only samples that will be sequenced are those where the results can be validated and experimentally interrogated. While the proposal is written to follow up on our current discoveries, given the emphasis on experimentally validated BDIs, we are able and willing to switch priorities and strategies to investigate and interrogate validated BDIs where the impact of the research is likely to be highest.

Alternative approaches

The biggest barriers in obtaining funding to detect BDI in the human genome arise from: (a) a disbelief that such transfers occur and (b) a disbelief that if such transfers occur that they would ever alter biology. We think that after the experiments proposed here, we will be able to convincingly state whether or not BDIs occur in the human somatic genome. Currently, the evidence points toward the presence of such integrations. However, upon further analysis and experimentation, we may find that they do not occur. This will still be an important finding, particularly as it relates to understanding the limits of nucleic acid movement and integration in genomes, aspects of bacterial pathogenesis and immunity, and even the limits of gene therapy. We do not anticipate that this will be clear before the end of this proposal. However, if it becomes clear that bacterial DNA integrations do not occur, we will focus our efforts on analysis of genome sequencing data collected using traditional sequence analysis.

PROTECTION OF HUMAN SUBJECTS

All personnel on this project will receive training on the proper handling of human subjects through HIPAA training and the CITI course.

Aim 1 - Identification of bacterial DNA integrations in publicly available cancer genome data

We will analyze available human genomic sequences available through dbGap. On August 4, 2010, the UMB IRB determined that such analysis does not require IRB review. There is no identifiable private information about the individual and the data are publicly available, preexisting specimens and data. Inclusion of women, minorities, and children is not applicable as we are using pre-existing samples and sequences.

Aim 2 – Genome sequencing of stomach adenocarcinoma samples

Our proposed experiments sequencing stomach adenocarcinomas samples involves the use of pre-existing nucleic acids samples collected in Mexico under FWA (federal wide assurance): 00004956; IRB (Institutional Review Board-CNIC) : 00003566; IORG (Institutional Organization Number-IMSS): 0002957; and DUNS (FIS): 815520374. There is no identifiable private information about the individuals.

Aim 2 – Genome sequencing of acute myeloid leukemia samples

Our proposed experiments sequencing acute myeloid leukemia samples involves the use of existing cryopreserved human leukemia cells, which have been collected previously, with IRB approval. In addition, new bone marrow samples from leukemia patients will also be aspirated for our research, under existing IRB-approved protocols. For all patient samples studied, the information is recorded in such a way that subjects cannot be linked or identified by laboratory researchers. No samples are associated with identifiable information about the donors.

1) Recruitment and Informed Consent. No recruitment of subjects is planned. Proper consent procedures will be used for all subjects studied and written informed IRB-approved consent documents will be signed and witnessed by the physician obtaining the samples.

2) Protection against risk. To minimize risks, marrow aspiration will be performed only by fully qualified individuals. Subjects will be excluded if they give a history of adverse reactions to local anesthetics, contact sensitivity to iodine, or fainting during phlebotomy; if phlebotomy or marrow aspiration is technically difficult (e.g. due to obesity); or if there are underlying medical illnesses which might increase risks to the subject. These safeguards have been found effective in minimizing risks to patients. Patients on clinical trials must meet the eligibility criteria specified for the protocol. The investigators and the institution will provide physician's services to any human subject who is physically injured through marrow aspiration. All subjects are given a card with the name and phone number of an investigator and instructed to call immediately if persistent tenderness, swelling, heat and/or redness appear in marrow aspirate sites. Any subject with such complaints will be seen promptly and treated as indicated. Obtaining marrow or peripheral blood leukemia specimens for laboratory studies includes the following potential risks: (a) allergic reactions to lidocaine (local anesthesia) or iodine tincture (skin preparation); (b) local hematoma or infection at site of skin puncture; and (c) osteomyelitis (marrow). Some discomfort at the site of the marrow aspirations is expected. However, since bone marrow from leukemia patients will be obtained at times when it is being obtained for other purposes, there will be no added risks to the patients for providing the marrow.

INCLUSION OF WOMEN AND MINORITIES

Samples from men and women and from all racial and ethnic groups will be studied in the laboratory, and no patient will be excluded on the basis of gender or racial/ethnic group. We expect the racial/ethnic composition of our studies to be similar to that of our patient population demographics.

Planned Enrollment Report

Study Title: Extent and Significance of Bacterial DNA Integrations in the Human Cancer Genome.

Domestic/Foreign: Domestic

Comments: The laboratory studies described in this project will require bone marrow specimens from patients with AML. Adequate numbers of samples are already available to us through the Hematologic Malignancies Program at the University of Maryland's Greenebaum Cancer Center. Confidentiality of donor samples will be maintained at all times. Donor name, identification number, etc. are removed from samples, which are assigned a Tissue Biorepository unique identification number.

Racial Categories	Ethnic Categories				Total
	Not Hispanic or Latino		Hispanic or Latino		
	Female	Male	Female	Male	
American Indian/Alaska Native	0	0	0	0	0
Asian	2	2	0	0	4
Native Hawaiian or Other Pacific Islander	0	0	0	0	0
Black or African American	8	8	0	0	16
White	9	12	2	2	25
More than One Race	0	0	0	0	0
Total	19	22	2	2	45

Study 1 of 1

INCLUSION OF CHILDREN

No children are included in our studies because acute myeloid leukemia occurs predominantly in adults, and the acute myeloid leukemia population at the University of Maryland Greenebaum Cancer Center is almost exclusively an adult population.

REFERENCES

1. **Acuna, R., B. E. Padilla, C. P. Florez-Ramos, J. D. Rubio, J. C. Herrera, P. Benavides, S. J. Lee, T. H. Yeats, A. N. Egan, J. J. Doyle, et al.** 2012. Adaptive horizontal transfer of a bacterial gene to an invasive insect pest of coffee. *Proc Natl Acad Sci U S A* **109**:4197-202. PubMed PMID: 22371593; PubMed Central PMCID: PMC3306691.
2. **Aikawa, T., H. Anbutsu, N. Nikoh, T. Kikuchi, F. Shibata, and T. Fukatsu.** 2009. Longicorn beetle that vectors pinewood nematode carries many *Wolbachia* genes on an autosome. *Proc Biol Sci* **276**:3791-8. PubMed PMID: 19692404; PubMed Central PMCID: PMC2817283.
3. **Altincicek, B., J. L. Kovacs, and N. M. Gerardo.** 2012. Horizontally transferred fungal carotenoid genes in the two-spotted spider mite *Tetranychus urticae*. *Biol Lett* **8**:253-7. PubMed PMID: 21920958; PubMed Central PMCID: PMC3297373.
4. **Altschul, S. F., T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman.** 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**:3389-402.
5. **Angiuoli, S. V., M. Matalka, A. Gussman, K. Galens, M. Vangala, D. R. Riley, C. Arze, J. R. White, O. White, and W. F. Fricke.** 2011. CloVR: a virtual machine for automated and portable sequence analysis from the desktop using cloud computing. *BMC Bioinformatics* **12**:356. PubMed PMID: 21878105; PubMed Central PMCID: PMC3228541.
6. **Aviles-Jimenez, F., F. Vazquez-Jimenez, R. Medrano-Guzman, A. Mantilla, and J. Torres.** 2014. Stomach microbiota composition varies between patients with non-atrophic gastritis and patients with intestinal type of gastric cancer. *Sci Rep* **4**:4202. PubMed PMID: 24569566; PubMed Central PMCID: PMC3935187.
7. **Beswick, E. J., D. A. Bland, G. Suarez, C. A. Barrera, X. Fan, and V. E. Reyes.** 2005. *Helicobacter pylori* binds to CD74 on gastric epithelial cells and stimulates interleukin-8 production. *Infect Immun* **73**:2736-43. PubMed PMID: 15845476; PubMed Central PMCID: PMC1087363.
8. **Beswick, E. J., and V. E. Reyes.** 2009. CD74 in antigen presentation, inflammation, and cancers of the gastrointestinal tract. *World J Gastroenterol* **15**:2855-61. PubMed PMID: 19533806; PubMed Central PMCID: PMC2699002.
9. **Blankenberg, D., G. Von Kuster, N. Coraor, G. Ananda, R. Lazarus, M. Mangan, A. Nekrutenko, and J. Taylor.** 2010. Galaxy: a web-based genome analysis tool for experimentalists. *Curr Protoc Mol Biol* **Chapter 19**:Unit 19 10 1-21. PubMed PMID: 20069535.
10. **Boyer, S. N., D. E. Wazer, and V. Band.** 1996. E7 protein of human papilloma virus-16 induces degradation of retinoblastoma protein through the ubiquitin-proteasome pathway. *Cancer Res* **56**:4620-4. PubMed PMID: 8840974.
11. **Cancer Genome Atlas, N.** 2012. Comprehensive molecular characterization of human colon and rectal cancer. *Nature* **487**:330-7. PubMed PMID: 22810696; PubMed Central PMCID: PMC3401966.
12. **Cancer Genome Atlas, N.** 2012. Comprehensive molecular portraits of human breast tumours. *Nature* **490**:61-70. PubMed PMID: 23000897; PubMed Central PMCID: PMC3465532.
13. **Cancer Genome Atlas Research, N.** 2012. Comprehensive genomic characterization of squamous cell lung cancers. *Nature* **489**:519-25. PubMed PMID: 22960745; PubMed Central PMCID: PMC3466113.
14. **Cancer Genome Atlas Research, N.** 2013. Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. *N Engl J Med* **368**:2059-74. PubMed PMID: 23634996; PubMed Central PMCID: PMC3767041.
15. **Cancer Genome Atlas Research, N.** 2011. Integrated genomic analyses of ovarian carcinoma. *Nature* **474**:609-15. PubMed PMID: 21720365; PubMed Central PMCID: PMC3163504.
16. **Caporaso, J. G., J. Kuczynski, J. Stombaugh, K. Bittinger, F. D. Bushman, E. K. Costello, N. Fierer, A. G. Pena, J. K. Goodrich, J. I. Gordon, et al.** 2010. QIIME allows analysis of high-throughput community sequencing data. *Nat Methods* **7**:335-6. PubMed PMID: 20383131; PubMed Central PMCID: PMC3156573.
17. **Corden, S. A., L. J. Sant-Cassia, A. J. Easton, and A. G. Morris.** 1999. The integration of HPV-18 DNA in cervical carcinoma. *Mol Pathol* **52**:275-82. PubMed PMID: 10748877; PubMed Central PMCID: PMC395710.
18. **Crabtree, J., S. Agrawal, A. Mahurkar, G. S. Myers, D. A. Rasko, and O. White.** 2014. Circleator: Flexible Circular Visualization of Genome-Associated Data with BioPerl and SVG. *Bioinformatics*. PubMed PMID: 25075113.

19. **Craig, J. P., S. Bekal, M. Hudson, L. Domier, T. Niblack, and K. N. Lambert.** 2008. Analysis of a horizontally transferred pathway involved in vitamin B-6 biosynthesis from the soybean cyst nematode *Heterodera glycines*. *Molecular Biology and Evolution* **25**:2085-2098. PubMed PMID: ISI:000259327900002.
20. **Danchin, E. G. J., M.-N. Rossoa, P. Vieiraa, J. d. Almeida-Englera, P. M. Coutinhob, B. Henrissatb, and P. Abada.** 2010. Multiple lateral gene transfers and duplications have promoted plant parasitism ability in nematodes. *Proc Natl Acad Sci U S A* **107**:17651–17656.
21. **Das, P., A. Thomas, U. Mahantshetty, S. K. Shrivastava, K. Deodhar, and R. Mulherkar.** 2012. HPV genotyping and site of viral integration in cervical cancers in Indian women. *PLoS One* **7**:e41012. PubMed PMID: 22815898; PubMed Central PMCID: PMC3397968.
22. **Desjardins, C. A., G. C. Cerqueira, J. M. Goldberg, J. C. Dunning Hotopp, B. J. Haas, J. Zucker, J. M. Ribeiro, S. Saif, J. Z. Levin, L. Fan, et al.** 2013. Genomics of *Loa loa*, a Wolbachia-free filarial parasite of humans. *Nat Genet* **45**:495-500. PubMed PMID: 23525074.
23. **Dunning Hotopp, J. C., M. E. Clark, D. C. Oliveira, J. M. Foster, P. Fischer, M. C. Torres, J. D. Giebel, N. Kumar, N. Ishmael, S. Wang, et al.** 2007. Widespread lateral gene transfer from intracellular bacteria to multicellular eukaryotes. *Science* **317**:1753-6. PubMed PMID: 17761848.
24. **Duxbury, M. S., H. Ito, M. J. Zinner, S. W. Ashley, and E. E. Whang.** 2004. CEACAM6 gene silencing impairs anoikis resistance and in vivo metastatic ability of pancreatic adenocarcinoma cells. *Oncogene* **23**:465-73. PubMed PMID: 14724575.
25. **Eidelman, F. J., A. Fuks, L. DeMarte, M. Taheri, and C. P. Stanners.** 1993. Human carcinoembryonic antigen, an intercellular adhesion molecule, blocks fusion and differentiation of rat myoblasts. *Journal of Cell Biology* **123**:467-475.
26. **Fast, E. M., M. E. Toomey, K. Panaram, D. Desjardins, E. D. Kolaczyk, and H. M. Frydman.** 2011. *Wolbachia* enhance *Drosophila* stem cell proliferation and target the germline stem cell niche. *Science* **334**:990-2. PubMed PMID: 22021671.
27. **Gan, L., S. J. Hahn, and L. K. Kaczmarek.** 1999. Cell type-specific expression of the Kv3.1 gene is mediated by a negative element in the 5' untranslated region of the Kv3.1 promoter. *J Neurochem* **73**:1350-62. PubMed PMID: 10501178.
28. **Goecks, J., A. Nekrutenko, J. Taylor, and T. Galaxy.** 2010. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol* **11**:R86. PubMed PMID: 20738864; PubMed Central PMCID: PMC2945788.
29. **Hauck, W., and C. P. Stanners.** 1995. Transcriptional regulation of the carcinoembryonic antigen gene. Identification of regulatory elements and multiple nuclear factors. *J Biol Chem* **270**:3602-10. PubMed PMID: 7876096.
30. **Hernandez, I., P. de la Torre, J. Rey-Campos, I. Garcia, J. A. Sanchez, R. Munoz, R. A. Rippe, T. Munoz-Yague, and J. A. Solis-Herruzo.** 2000. Collagen alpha1(I) gene contains an element responsive to tumor necrosis factor-alpha located in the 5' untranslated region of its first exon. *DNA Cell Biol* **19**:341-52. PubMed PMID: 10882233.
31. **Hoadley, K. A., C. Yau, D. M. Wolf, A. D. Cherniack, D. Tamborero, S. Ng, M. D. Leiserson, B. Niu, M. D. McLellan, V. Uzunangelov, et al.** 2014. Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. *Cell* **158**:929-44. PubMed PMID: 25109877; PubMed Central PMCID: PMC4152462.
32. **Hollingsworth, R. E., Jr., P. L. Chen, and W. H. Lee.** 1993. Integration of cell cycle control with transcriptional regulation by the retinoblastoma protein. *Curr Opin Cell Biol* **5**:194-200. PubMed PMID: 8507491.
33. **Husnik, F., N. Nikoh, R. Koga, L. Ross, R. P. Duncan, M. Fujie, M. Tanaka, N. Satoh, D. Bachtrog, A. C. Wilson, et al.** 2013. Horizontal gene transfer from diverse bacteria to an insect genome enables a tripartite nested mealybug symbiosis. *Cell* **153**:1567-78. PubMed PMID: 23791183.
34. **Ilantzis, C., L. DeMarte, R. A. Screatton, and C. P. Stanners.** 2002. Deregulated expression of the human tumor marker CEA and CEA family member CEACAM6 disrupts tissue architecture and blocks colonocyte differentiation. *Neoplasia* **4**:151-63. PubMed PMID: 11896570; PubMed Central PMCID: PMC1550325.
35. **Ilver, D., A. Arnqvist, J. Ogren, I. M. Frick, D. Kersulyte, E. T. Incecik, D. E. Berg, A. Covacci, L. Engstrand, and T. Boren.** 1998. *Helicobacter pylori* adhesin binding fucosylated histo-blood group antigens revealed by retagging. *Science* **279**:373-7. PubMed PMID: 9430586.

36. **Ioannidis, P., K. L. Johnston, D. R. Riley, N. Kumar, J. R. White, K. T. Olarte, S. Ott, L. J. Tallon, J. M. Foster, M. J. Taylor, et al.** 2013. Extensively duplicated and transcriptionally active recent lateral gene transfer from a bacterial *Wolbachia* endosymbiont to its host filarial nematode *Brugia malayi*. *BMC Genomics* **14**:639. PubMed PMID: 24053607.
37. **Ioannidis, P., Y. Lu, N. Kumar, T. Creasy, S. Daugherty, M. C. Chibucos, J. Orvis, A. Shetty, S. Ott, M. Flowers, et al.** 2014. Rapid transcriptome sequencing of an invasive pest, the brown marmorated stink bug *Halyomorpha halys*. *BMC Genomics* **15**:738. PubMed PMID: 25168586; PubMed Central PMCID: PMC4174608.
38. **Ioannidis, P., Y. Lu, N. Kumar, T. Creasy, S. Daugherty, M. C. Chibucos, J. Orvis, A. Shetty, S. Ott, M. Flowers, et al.** 2014. Rapid transcriptome sequencing of an invasive pest, the brown marmorated stink bug *Halyomorpha halys*. *BMC Genomics* **15**:738. PubMed PMID: 25168586.
39. **Klasson, L., Z. Kambris, P. E. Cook, T. Walker, and S. P. Sinkins.** 2009. Horizontal gene transfer between *Wolbachia* and the mosquito *Aedes aegypti*. *BMC Genomics* **10**:33. PubMed PMID: 19154594; PubMed Central PMCID: PMC2647948.
40. **Kondo, N., N. Nikoh, N. Ijichi, M. Shimada, and T. Fukatsu.** 2002. Genome fragment of *Wolbachia* endosymbiont transferred to X chromosome of host insect. *Proc Natl Acad Sci U S A* **99**:14280-5. PubMed PMID: 12386340.
41. **Koutsovoulos, G., B. Makepeace, V. N. Tanya, and M. Blaxter.** 2014. Palaeosymbiosis revealed by genomic fossils of *Wolbachia* in a stronglyloidean nematode. *PLoS Genet* **10**:e1004397. PubMed PMID: 24901418; PubMed Central PMCID: PMC4046930.
42. **Kuczynski, J., J. Stombaugh, W. A. Walters, A. Gonzalez, J. G. Caporaso, and R. Knight.** 2011. Using QIIME to analyze 16S rRNA gene sequences from microbial communities. *Curr Protoc Bioinformatics* **Chapter 10**:Unit 10 7. PubMed PMID: 22161565; PubMed Central PMCID: PMC3249058.
43. **Kwak, H., N. J. Fuda, L. J. Core, and J. T. Lis.** 2013. Precise maps of RNA polymerase reveal how promoters direct initiation and pausing. *Science* **339**:950-3. PubMed PMID: 23430654; PubMed Central PMCID: PMC3974810.
44. **Larman, T. C., S. R. DePalma, A. G. Hadjipanayis, N. Cancer Genome Atlas Research, A. Protopopov, J. Zhang, S. B. Gabriel, L. Chin, C. E. Seidman, R. Kucherlapati, et al.** 2012. Spectrum of somatic mitochondrial mutations in five cancers. *Proc Natl Acad Sci U S A* **109**:14087-91. PubMed PMID: 22891333; PubMed Central PMCID: PMC3435197.
45. **Lee, E., R. Iskow, L. Yang, O. Gokcumen, P. Haseley, L. J. Luquette, 3rd, J. G. Lohr, C. C. Harris, L. Ding, R. K. Wilson, et al.** 2012. Landscape of somatic retrotransposition in human cancers. *Science* **337**:967-71. PubMed PMID: 22745252.
46. **Li, H., and R. Durbin.** 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**:1754-60. PubMed PMID: 19451168; PubMed Central PMCID: PMC2705234.
47. **Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, and R. Durbin.** 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**:2078-9. PubMed PMID: 19505943; PubMed Central PMCID: PMC2723002.
48. **Li, Z. W., Y. H. Shen, Z. H. Xiang, and Z. Zhang.** 2011. Pathogen-origin horizontally transferred genes contribute to the evolution of Lepidopteran insects. *BMC Evol Biol* **11**:356. PubMed PMID: 22151541; PubMed Central PMCID: PMC3252269.
49. **Mayer, W. E., L. N. Schuster, G. Bartelmes, C. Dieterich, and R. J. Sommer.** 2011. Horizontal gene transfer of microbial cellulases into nematode genomes is associated with functional assimilation and gene turnover. *BMC Evol Biol* **11**:13. PubMed PMID: 21232122; PubMed Central PMCID: PMC3032686.
50. **McCaw, S. E., E. H. Liao, and S. D. Gray-Owen.** 2004. Engulfment of *Neisseria gonorrhoeae*: revealing distinct processes of bacterial entry by individual carcinoembryonic antigen-related cellular adhesion molecule family receptors. *Infect Immun* **72**:2742-52. PubMed PMID: 15102784; PubMed Central PMCID: PMC387857.
51. **McNulty, S. N., J. M. Foster, M. Mitreva, J. C. Dunning Hotopp, J. Martin, K. Fischer, B. Wu, P. J. Davis, S. Kumar, N. W. Brattig, et al.** 2010. Endosymbiont DNA in endobacteria-free filarial nematodes indicates ancient horizontal genetic transfer. *PLoS One* **5**:e11029. PubMed PMID: 20543958; PubMed Central PMCID: PMC2882956.
52. **Melsheimer, P., S. Vinokurova, N. Wentzensen, G. Bastert, and M. von Knebel Doeberitz.** 2004. DNA aneuploidy and integration of human papillomavirus type 16 e6/e7 oncogenes in intraepithelial

- neoplasia and invasive squamous cell carcinoma of the cervix uteri. *Clin Cancer Res* **10**:3059-63. PubMed PMID: 15131043.
53. **Moran, N. A., and T. Jarvik.** 2010. Lateral transfer of genes from fungi underlies carotenoid production in aphids. *Science* **328**:624-7. PubMed PMID: 20431015.
 54. **Muenzner, P., C. Dehio, T. Fujiwara, M. Achtman, T. F. Meyer, and S. D. Gray-Owen.** 2000. Carcinoembryonic antigen family receptor specificity of *Neisseria meningitidis* Opa variants influences adherence to and invasion of proinflammatory cytokine-activated endothelial cells. *Infect Immun* **68**:3601-7. PubMed PMID: 10816518; PubMed Central PMCID: PMC97649.
 55. **Nakabachi, A., K. Ishida, Y. Hongoh, M. Ohkuma, and S. Y. Miyagishima.** 2014. Aphid gene of bacterial origin encodes a protein transported to an obligate endosymbiont. *Curr Biol* **24**:R640-1. PubMed PMID: 25050957.
 56. **Nikoh, N., J. P. McCutcheon, T. Kudo, S. Y. Miyagishima, N. A. Moran, and A. Nakabachi.** 2010. Bacterial genes in the aphid genome: absence of functional gene transfer from *Buchnera* to its host. *PLoS Genet* **6**:e1000827. PubMed PMID: 20195500; PubMed Central PMCID: PMC2829048.
 57. **Nikoh, N., and A. Nakabachi.** 2009. Aphids acquired symbiotic genes via lateral gene transfer. *BMC Biol* **7**:12. PubMed PMID: 19284544; PubMed Central PMCID: PMC2662799.
 58. **Nikoh, N., K. Tanaka, F. Shibata, N. Kondo, M. Hizume, M. Shimada, and T. Fukatsu.** 2008. *Wolbachia* genome integrated in an insect chromosome: evolution and fate of laterally transferred endosymbiont genes. *Genome Res* **18**:272-80. PubMed PMID: 18073380; PubMed Central PMCID: PMC2203625.
 59. **Novakova, E., and N. A. Moran.** 2012. Diversification of genes for carotenoid biosynthesis in aphids following an ancient transfer from a fungus. *Mol Biol Evol* **29**:313-23. PubMed PMID: 21878683.
 60. **Ondov, B. D., N. H. Bergman, and A. M. Phillippy.** 2011. Interactive metagenomic visualization in a Web browser. *BMC Bioinformatics* **12**:385. PubMed PMID: 21961884; PubMed Central PMCID: PMC3190407.
 61. **Ordóñez, C., R. A. Screaton, C. Ilantzis, and C. P. Stanners.** 2000. Human carcinoembryonic antigen functions as a general inhibitor of anoikis. *Cancer Res* **60**:3419-24. PubMed PMID: 10910050.
 62. **Orvis, J., J. Crabtree, K. Galens, A. Gussman, J. M. Inman, E. Lee, S. Nampally, D. Riley, J. P. Sundaram, V. Felix, et al.** 2010. Ergatis: a web interface and scalable software system for bioinformatics workflows. *Bioinformatics* **26**:1488-92. PubMed PMID: 20413634; PubMed Central PMCID: PMC2881353.
 63. **Riley, D. R., K. B. Sieber, K. M. Robinson, J. R. White, A. Ganesan, S. Nourbakhsh, and J. C. Dunning Hotopp.** 2013. Bacteria-Human Somatic Cell Lateral Gene Transfer Is Enriched in Cancer Samples. *PLoS Comput Biol* **9**:e1003107.
 64. **Robinson, K. M., and J. C. Dunning Hotopp.** 2014. Mobile elements and viral integrations prompt considerations for bacterial DNA integration as a novel carcinogen. *Cancer Lett* **352**:137-44. PubMed PMID: 24956175; PubMed Central PMCID: PMC4134975.
 65. **Robinson, K. M., K. B. Sieber, and J. C. Dunning Hotopp.** 2013. A review of bacteria-animal lateral gene transfer may inform our understanding of diseases like cancer. *PLoS Genet* **9**:e1003877.
 66. **Rojas, M., L. DeMarte, R. A. Screaton, and C. P. Stanners.** 1996. Radical differences in functions of closely related members of the human carcinoembryonic antigen gene family. *Cell Growth Differ* **7**:655-62. PubMed PMID: 8732675.
 67. **Romanczuk, H., and P. M. Howley.** 1992. Disruption of either the E1 or the E2 regulatory gene of human papillomavirus type 16 increases viral immortalization capacity. *Proc Natl Acad Sci U S A* **89**:3159-63. PubMed PMID: 1313584; PubMed Central PMCID: PMCPMC48824.
 68. **Salzberg, S. L., O. White, J. Peterson, and J. A. Eisen.** 2001. Microbial genes in the human genome: lateral transfer or gene loss? *Science* **292**:1903-6. PubMed PMID: 11358996.
 69. **Scheffner, M., J. M. Huibregtse, R. D. Vierstra, and P. M. Howley.** 1993. The HPV-16 E6 and E6-AP complex functions as a ubiquitin-protein ligase in the ubiquitination of p53. *Cell* **75**:495-505. PubMed PMID: 8221889.
 70. **Schrewe, H., J. Thompson, M. Bona, L. J. Hefta, A. Maruya, M. Hassauer, J. E. Shively, S. von Kleist, and W. Zimmermann.** 1990. Cloning of the complete gene for carcinoembryonic antigen: analysis of its promoter indicates a region conveying cell type-specific expression. *Mol Cell Biol* **10**:2738-48. PubMed PMID: 2342461; PubMed Central PMCID: PMC360634.

71. **Singh, I. S., J. R. He, S. Calderwood, and J. D. Hasday.** 2002. A high affinity HSF-1 binding site in the 5'-untranslated region of the murine tumor necrosis factor- α gene is a transcriptional repressor. *J Biol Chem* **277**:4981-8. PubMed PMID: 11734555.
72. **Skaar, E. P., and H. S. Seifert.** 2002. The misidentification of bacterial genes as human cDNAs: was the human D-1 tumor antigen gene acquired from bacteria? *Genomics* **79**:625-7. PubMed PMID: 11991711.
73. **Sloan, D. B., A. Nakabachi, S. Richards, J. Qu, S. C. Murali, R. A. Gibbs, and N. A. Moran.** 2014. Parallel histories of horizontal gene transfer facilitated extreme reduction of endosymbiont genomes in sap-feeding insects. *Mol Biol Evol* **31**:857-71. PubMed PMID: 24398322; PubMed Central PMCID: PMC3969561.
74. **Soeth, E., T. Wirth, H. J. List, S. Kumbhani, A. Petersen, M. Neumaier, F. Czubyko, and H. Juhl.** 2001. Controlled ribozyme targeting demonstrates an antiapoptotic effect of carcinoembryonic antigen in HT29 colon cancer cells. *Clin Cancer Res* **7**:2022-30. PubMed PMID: 11448920.
75. **Tallon, L. J., X. Liu, S. Bennuru, M. C. Chibucos, A. Godinez, S. Ott, X. Zhao, L. Sadzewicz, C. M. Fraser, T. B. Nutman, et al.** 2014. Single molecule sequencing and genome assembly of a clinical specimen of *Loa loa*, the causative agent of loiasis. *BMC Genomics* **15**:788. PubMed PMID: 25217238; PubMed Central PMCID: PMC4175631.
76. **The Cancer Genome Atlas Research, N.** 2014. Comprehensive molecular characterization of gastric adenocarcinoma. *Nature* **513**:202-209.
77. **Van Tine, B. A., J. C. Kappes, N. S. Banerjee, J. Knops, L. Lai, R. D. Steenbergen, C. L. Meijer, P. J. Snijders, P. Chatis, T. R. Broker, et al.** 2004. Clonal selection for transcriptionally active viral oncogenes during progression to cancer. *J Virol* **78**:11172-86. PubMed PMID: 15452237; PubMed Central PMCID: PMC521852.
78. **Wentzensen, N., S. Vinokurova, and M. von Knebel Doeberitz.** 2004. Systematic review of genomic integration sites of human papillomavirus genomes in epithelial dysplasia and invasive cancer of the female lower genital tract. *Cancer Res* **64**:3878-84. PubMed PMID: 15172997.
79. **Werren, J. H., S. Richards, C. A. Desjardins, O. Niehuis, J. Gadau, J. K. Colbourne, L. W. Beukeboom, C. Desplan, C. G. Elsik, C. J. Grimmelikhuijzen, et al.** 2010. Functional and evolutionary insights from the genomes of three parasitoid *Nasonia* species. *Science* **327**:343-8. PubMed PMID: 20075255; PubMed Central PMCID: PMC2849982.
80. **Wheeler, D., A. J. Redding, and J. H. Werren.** 2013. Characterization of an ancient lepidopteran lateral gene transfer. *PLoS One* **8**:e59262. PubMed PMID: 23533610; PubMed Central PMCID: PMC3606386.
81. **Wirth, T., E. Soeth, F. Czubyko, and H. Juhl.** 2002. Inhibition of endogenous carcinoembryonic antigen (CEA) increases the apoptotic rate of colon cancer cells and inhibits metastatic tumor growth. *Clin Exp Metastasis* **19**:155-60. PubMed PMID: 11964079.
82. **Woolfit, M., I. Iturbe-Ormaetxe, E. A. McGraw, and S. L. O'Neill.** 2009. An ancient horizontal gene transfer between mosquito and the endosymbiotic bacterium *Wolbachia pipientis*. *Mol Biol Evol* **26**:367-74. PubMed PMID: 18988686.
83. **Wu, B., J. Novelli, D. Jiang, H. A. Dailey, F. Landmann, L. Ford, M. J. Taylor, C. K. Carlow, S. Kumar, J. M. Foster, et al.** 2013. Interdomain lateral gene transfer of an essential ferrochelatase gene in human parasitic nematodes. *Proc Natl Acad Sci U S A* **110**:7748-53. PubMed PMID: 23610429; PubMed Central PMCID: PMC3651471.
84. **Zhu, B., M. M. Lou, G. L. Xie, G. Q. Zhang, X. P. Zhou, B. Li, and G. L. Jin.** 2011. Horizontal gene transfer in silkworm, *Bombyx mori*. *BMC Genomics* **12**:248. PubMed PMID: 21595916; PubMed Central PMCID: PMC3116507.

PLAN FOR SHARING RESEARCH DATA

Data release plan

Data generated by this project will be made available to the community of researchers and clinicians according to the most recent guidelines in NOT-OD-14-124, “NIH Genomic Data Sharing Policy”. More specifically, we propose the following plan:

Submission of genome sequence data to public databases. Upon generation, raw sequence data will be made available through dbGap in accordance with the IRB for these samples. These data will also include quality values for each sequence.

Release of analysis performed under the project. It is our intention to use scientific publications as the primary means of releasing the analyses that will be performed in the course of these studies.

Presentations at national/international scientific meetings. We anticipate being given the opportunity to present this research at national and international meetings to further disseminate the analyses and results obtained in this study.

Resources sharing plan – Software and Analysis Tools

The IGS informatics group and Dr. Dunning Hotopp believe in open-access to software and bioinformatic pipeline and have a track record of making all the source codes of our software and pipelines available under open-access at sourceforge.org and/or [github](https://github.com).