**Specific Aims**

Exercise is a powerful physiological stimulus that contributes to disease prevention and is used as a disease treatment. While the protective effects of exercise are well known for many chronic diseases, the molecular mechanisms by which exercise prevents disease and improves health are still poorly understood. The MoTrPAC consortium will address this knowledge gap by developing a ‘molecular map’ of responses to physical activity (PA) and by connecting the molecular responses with health outcomes. Based on our extensive prior work, we are ideally positioned to establish a bioinformatics center for the consortium that will interface with other consortium sites to (A) establish and maintain an infrastructure for tracking the data and biosample resources; (B) develop data processing pipelines; (C) run the pipelines to create a database of molecular profiles; (D) develop tools for other investigators to analyze the profiling data and synthesize it into network modules and pathway knowledge; (E) conduct preliminary analyses of the data; and (F) provide analytical input to facilitate the construction of molecular response models. Specifically, the bioinformatics center will accomplish the following:

**Aim 1 Develop an informatics infrastructure.** The center will develop an infrastructure for data processing, storage and analysis via the following two subaims: **(a)** develop a cloud-based infrastructure for data processing and analysis; **(b)** develop an open interoperable data storage system for cloud-based management of data and metadata based on Linked Open Data standards.

**Aim 2 Create the database of molecular profiles.** The center will create the database of molecular profiles via the following two subaims: **(a)** develop data processing protocols and pipelines for genomic, epigenomic, transcriptomic, metabolomic, and proteomic data and implement associated quality control standards; and **(b)** organize standardization of data and metadata by the consortium and apply this framework to the processing of raw data to populate the MoTrPAc database with molecular profiles and implement access-controlled sharing, analysis, reuse, public distribution and archiving of the molecular profiles.

**Aim 3** **Perform data analysis and construct response models.** The center will construct models of response to physical activity via the following four subaims: **(a)** normalize the processed multi-omics profiles between individuals, deconvolute profiles of complex tissues, and register time points for time-course analyses; **(b)** analyze and cluster gene activity data to identify coordinated acute and durable response modules for both single-perturbation and time-course experiments; **(c)** build integrative models and identify biomarkers that predict response to PA based on multi-omics profiles in the context of other relevant public data; and **(d)** integrate genomic information into exercise response models to identify the genetic variants associated with individual response variation.

**Aim 4 Organize and provide informatics support for cross-consortium projects.** The center will provide organizational and informatics support for cross-consortium projects via the following six subaims: **(a)** deploy a consortium portal for use by the coordinating center and other consortium members; **(b)** develop a system for tracking animal and human specimens across the consortium sites; **(c)** support a data working group to engage the consortium in the development of the MoTrPAc database; **(d)** support an analysis working group to provide expertise in data management and analytics and facilitate communication about data analysis methods and tools; **(e)** organize the development by the consortium of a section within WikiPathways devoted to molecular pathways relevant for PA; and **(f)** support integration of animal study data and develop plans for replication studies.

By accomplishing these aims, the bioinformatics center will enable the MoTrPAc consortium to develop a molecular map of PA responses, to connect them with health outcomes, and thus establish a knowledge base for personalized exercise prescriptions for disease prevention and treatment.

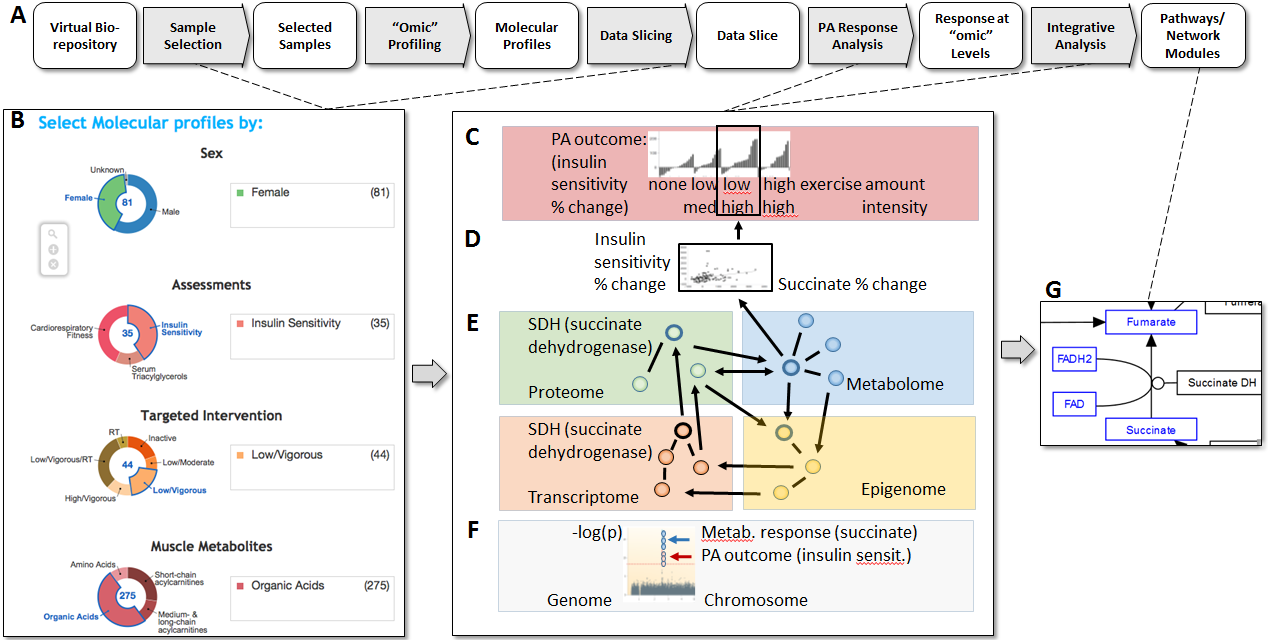
**RESEARCH STRATEGY**

**Significance**

Exercise is a powerful physiological stimulus contributing to disease prevention and may serve as an effective disease treatment. Physical inactivity, prevalent in modern lifestyles, is a contributing factor to many unhealthy conditions and chronic diseases [[11](#_ENREF_11)]. The protective effects of exercise are well known for many chronic diseases including metabolic, neurodegenerative, and cardiovascular diseases, as well as cancer. In addition, it is commonly used as therapeutic intervention for those suffering from diabetes and cardiovascular disease. While the extensive benefits of exercise are widely acknowledged, there is very little understanding about the molecular mechanisms underpinning its associated health benefits [[53](#_ENREF_53), [73](#_ENREF_73)]. The MoTrPAc project will address this knowledge gap by providing a molecular map of physical activity (PA) responses connected to individual health outcomes, taking into account genetic and other modulating factors, thus enabling personalized exercise prescriptions. Working with the consortium, the Bioinformatics Center will process a large volume of diverse molecular profiles, assemble them into an accessible database, and engage the consortium in integrative collaborative interpretation of these profiles using cloud-hosted analysis and visualization tools.

The MoTrPAc consortium will develop conceptual frameworks for constructing the map and general methodological templates with specific ‘use cases’ to inform the design of the bioinformatics systems required for map construction. One such template may be derived from our studies such as STRRIDE[[46](#_ENREF_46)], the first metabolic profiling investigation to examine signatures of PA intervention in human skeletal muscle and to relate them to cardio-metabolic health outcomes and gene expression responses to exercise training. By integrating skeletal muscle metabolomics and transcriptomics we identified the roles of fatty acid oxidation and succinate dehydrogenase as key mediators of the differential effects of exercise intensity on exercise-induced responses in cardiorespiratory fitness, insulin sensitivity, and serum triglyceride levels. Our earlier studies found mediators of insulin sensitivity responses to exercise in plasma [[48](#_ENREF_48), [49](#_ENREF_49)]; integrated peripheral blood and skeletal muscle metabolomics signatures with proteomics to implicate glycine metabolism as a key mediator of improvements in insulin sensitivity after a resistance exercise program[[40](#_ENREF_40)]; and “triangulated” health outcomes and metabolomics, genomics, and epigenomics by mapping metabolite intermediate phenotypes using GWAS onto the same locus[[59](#_ENREF_59)]. As illustrated in **Fig. 1**,a general methodological template for the analysis of “multi-omic” responses to exercise has emerged from these studies. Building upon this and additional templates contributed by other consortium members, we will establish a series specific ‘use cases’, develop bioinformatic systems and construct a 'molecular map' of PA in humans that will enable personalized exercise prescriptions.

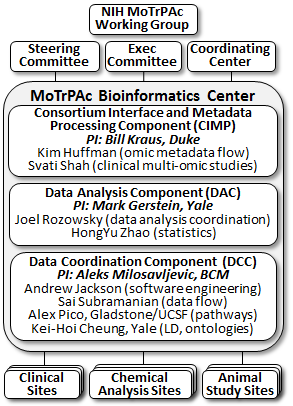
**Figure 1. A methodological template and data flow for constructing the 'molecular map' of PA in humans**. We designed a multi-step process **(A)** for constructing the molecular map of PA that extrapolates our recent study[[46](#_ENREF_46)]. The process starts by sample selection from a virtual biorepository for molecular profiling. **(B)** Upon completion of molecular profiling, selection of “data slices” for PA response analysis is facilitated by rich metadata about subjects, interventions, and ‘omic’ features such as metabolite classes. **(C)** The PA interventions showing the strongest measurable outcomes (e.g. insulin sensitivity improvement) are identified as are **(D)** specific metabolites that mediate PA response and **(E)** interacting network modules. **(F)** Health improvements and associated molecular responses are ‘triangulated’ by mapping them onto overlapping genomic loci. **(G)** The integrative analysis provides pathway knowledge that is collaboratively developed using WikiPathways and used for visually organizing chains of molecular responses.



**Innovation**

By providing innovative approaches we will address the following bioinformatics problems that are essential for the construction of the molecular map of PA responses:

* We will develop a series of novel methodological templates (such as the one in **Fig. 1**) to inform the design of bioinformatic systems and the construction of molecular maps of PA responses.
* To address confounding factors due to variation in cell-type proportions within profiled tissues, and to provide insights into metabolic coupling and other cellular interactions within muscle and other tissues as they respond to PA, we will apply the novel Epigenomic Deconvolution method.
* To dissect the temporal dynamic effects of different regulatory subsystems as tissues respond to PA, we will employ Lomb-Scargle periodograms and DREISS[[4](#_ENREF_4)], our innovative continuous signal modeling method.
* RNA-seq provides much more information than simple quantification of gene expression levels. To detect alternative RNA splicing and RNA transducers of PA, we will apply an array of our novel tools.
* To study PA responses of regulatory modules involving both proximally and distally regulating transcription factors (TFs) and miRNAs we will deploy our novel Loregic method[[102](#_ENREF_102)]. To further integrate human regulatory responses with those in animal studies, we will apply our novel OrthoClust[[105](#_ENREF_105)] tool that maps gene co-expression modules across species.



**Figure 2.** MoTrPAc Bioinformatics Center organization and interactions.

* To disentangle acute molecular PA responses from durable tissue-specific changes in chromatin modifications that epigenetically reprogram regulatory elements, we will link the methylation and ChIP-seq data generated by MoTrPAc to the epigenomic profiles generated by the ENCODE, Epigenome Roadmap, and GTEx consortia.
* Building upon our extensive experience with crowdsourcing to develop WikiPathways content, we will engage the MoTrPAc consortium in an innovative project to synthesize knowledge gained into a “PA Pathways” section of WikiPathways. This structured graphical knowledge representation, along with other WikiPathway content and Cytoscape tools will in turn be used to visualize molecular responses to PA in the context of networks and pathways.
* To enhance web-based data tracking and interoperability we will implement newly adopted W3C Linked Open Data standards. We will enable tracking of biosample data across MoTrPAc sites using the W3C Linked Data Platform 1.0 standard. To enable collaborative and continuous metadata development, we will implement the novel concept of dynamic metadata using the recently adopted W3C PROV-O standard.

**Approach**

As indicated in **Fig. 2**, the Bioinformatics Center will be organized into three integrated components, led by PIs with complementary expertise they will collaborate with NIH scientists and staff, and with other MoTrPAC participants. The Center components will communicate via weekly teleconferences and at least three in-person meetings per year. The Center participants will organize at least two consortium working groups (**Aim 4(c,d)**) that will have weekly teleconferences and semi-annual in-person meetings and will participate in Steering Committee calls. Taken together, the Duke CIMP, Yale DAC, and Baylor DCC staff will co-participate in at least three teleconferences per week, thus ensuring sufficient communication within the Bioinformatics Center and with the MoTrPAC Consortium.

Below we describe our plan for accomplishing Aims 1-4 during Years 2-6 of the project. During *Year 1 (the planning year)*, we will help plan the study lead by the Steering Committee. We will develop a detailed plan for tracking biospecimens; establish standards for data, provenance, and ontology-based metadata; develop a plan for database assembly, curation and storage; help to establish analysis methods and data query approaches; address various bioethical issues and develop de-identification and anonymization protocols, as well as a common data sharing plan and data release policies consistent with achieving the goals of the program. *In Year 1, we will produce detailed timelines with concrete milestones for the entire study, which may lead to adjustments in our approach and the timelines that are detailed next.*

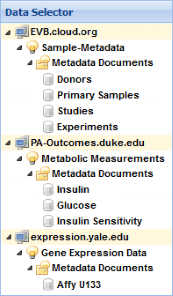
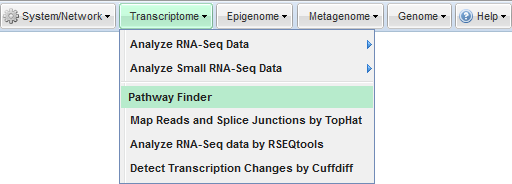
**Aim 1. Develop an informatics infrastructure.**

The center will develop an infrastructure for data processing, storage and analysis. Our approach will be based on our experience in providing data coordination and performing analyses for genome projects at the Baylor Human Genome Sequencing Center [[37](#_ENREF_37), [38](#_ENREF_38), [43](#_ENREF_43), [72](#_ENREF_72), [91](#_ENREF_91)], TCGA[[3](#_ENREF_3)], NIH Roadmap Epigenomics [[8](#_ENREF_8), [68](#_ENREF_68), [83](#_ENREF_83)], Clinical Genome Resource (ClinGen)[[74](#_ENREF_74), [81](#_ENREF_81)] and, most recently, the Extracellular RNA Communication project [[1](#_ENREF_1), [99](#_ENREF_99)]. The extensive data coordination experience gained and best practices developed during our work on these projects are captured in Genboree Workbench and GenboreeKB free open source software and associated web services; these will be further developed in **Aims 1(a)** and **1(b)** respectively.

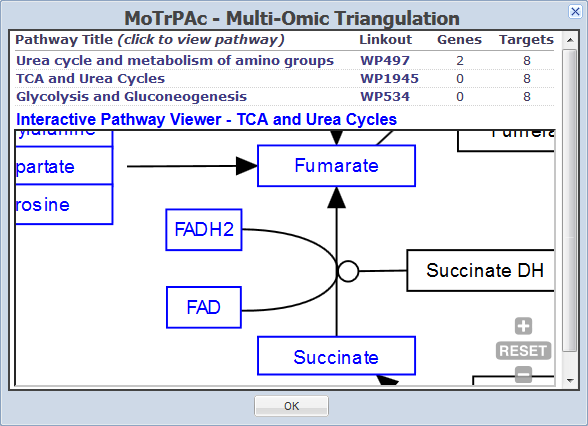
**Aim 1(a). Develop a cloud-based infrastructure for data processing and analysis in the context of networks and pathways.**

**Preliminary studies:** The Genboree Workbench is a web-based front end to a rich set of RESTful Genboree web services that make data, pipelines and tools available for diverse “omic” analyses. Tool inputs and storage output areas may involve different hosts, permitting distributed storage and execution of pipelines and tools on a combination of commercial and local clusters. The Workbench contains tools for genome sequencing[[15](#_ENREF_15), [28](#_ENREF_28)] microbiome analysis[[82](#_ENREF_82)], and processing of tens of thousands of diverse molecular profiles for the NIH Roadmap Epigenome project[[6](#_ENREF_6), [23](#_ENREF_23), [83](#_ENREF_83)], as well as for processing of exRNA profiles as part of the Extracellular RNA Communication project[[99](#_ENREF_99)]. Using the Workbench, an exRNA researcher may upload and map RNA-seq reads obtained from exosomal RNA found in the cerebro-spinal fluid of glioblastoma patients, identify miRNAs specific to patients, use the Pathway Finder tool to find pathways enriched in the miRNAs and their targets, identify relevant network modules, and visualize them in Cytoscape or WikiPathways[[78](#_ENREF_78), [88](#_ENREF_88)]. Tutorials, videos, and Wiki pages guide the ExRNA consortium members and the wider scientific community on how to use the Workbench on their privately hosted or public data.

**Proposed plan:** The Genboree Workbench will provide single-point access to tools for data and metadata submission, pipelines for data processing (described in **Aim 2(a)**), analysis, and visualization. Having selected molecular profiles of interest using faceted search (depicted in **Fig. 1B**), users would apply a number of tools within the Workbench, such as the Pathway Finder, to identify network modules or pathways to be visualized (depicted in **Fig. 3-right**).The Center will prioritize tool integration based on the Consortium’s requirements.



**Figure 3. Workbench tools for integrative analysis of molecular profiles.**



Our previously published analyses of the storage, compute, and I/O resources required for sequencing thousands of genomes using local cluster computing vs. Amazon Web Services[[28](#_ENREF_28)] will inform the final infrastructural design in Year 1. We anticipate a hybrid solution in which most computing performed at the Bioinformatics Center while peak demand being is met by “elastic” CPU hours from Rackspace or Amazon clouds. Large-volume data submissions will be facilitated by our existing bulk file transfer tool built on FTP or by Aspera Fastp. The largest volume of data will come from clinical studies and will be archived at dbGaP, while animal-study data will be archived at NCBI GEO, and metabolomics/proteomic profiles at HMDB and the Scripps METLIN. We have extensive experience jointly developing metadata with GEO, dbGaP, and non-NCBI archives, and with submitting tens of thousands of records for the Roadmap Epigenome and other projects. To complement this long-term archival storage, we plan on providing local off-line storage on detachable NAS devices such as Synology within a dedicated storage room in our data center, covering all BAM and other raw data files generated, about 2PB of data. Conservatively assuming 10% increase in storage density per year, 12 NAS devices will be required at *one-time hardware cost* of about $70K, or $0.035 per GB which is *less than half* the *recurrent yearly cost* of $0.084 for Amazon's most economical Glacier service. Data access service will be provided at cost by the Bioinformatics Research Laboratory Core at BCM (directed by Milosavljevic).

**Potential Problems & Alternative Approaches:** The design of the compute infrastructure for the project is contingent on resources available at participating sites and on possible constraints that are currently unknown. We will be well prepared to virtually integrate any storage and compute resources offered by consortium members. Specifically, by adopting the Linked Data Platform 1.0 (LDP 1.0) standard as proposed in **Aim 1(b)**, distributed hosts will be able to expose their MoTrPAc-related data and compute resources within the Workbench, thus eliminating the strict requirement for centralized warehousing. For example, the Workbench will identify resources available at specific hosts by LDP requests and will then present a unified view of geographically distributed resources to the user (**Fig. 3-left).**

**Aim 1(b). Develop an open interoperable data storage system for cloud-based management of data and metadata based on Linked Open Data standards.**

**Preliminary studies:** We have developed GenboreeKB, a free open source Redmine plugin developed using Ruby-on-Rails framework and MongoDB as a back-end database. GenboreeKB is one of the first systems to support metadata modeling, validation, and tracking using the new Linked Data standards including JSON-LD and RDF[[99](#_ENREF_99)]. Gerstein, Cheung, Pico and Milosavljevic have jointly published a large-scale “omic” use case[[99](#_ENREF_99)] for Linked Data technologies and have established a record of productive collaboration in metadata modeling within the Extracellular RNA Communication Consortium. The established team and the tested functionality of GenboreeKB will accelerate initial deployment and minimize risk, while enabling PROV-O and Linked Data standards application in a uniquely suitable solution for integrating highly heterogeneous, dynamic, and distributed data and metadata that will be generated by the MoTrPAc project.

**Proposed Plan:** Data models and formats for genomic, epigenomic, transcriptomic, metabolomic, and proteomic data are generally well established and will be adopted without major modifications. In contrast, the metadata will pose significant challenge due to heterogeneity of MoTrPAc data and the dynamic nature of metadata revision and curation. Rich metadata will play a key role in multiple steps in the processing and analysis (**Fig. 1A**), including sample retrieval and profiling; navigation of profiles via faceted search (**Fig. 1B**), linear tree drill-down diagrams, tabular formats, and grid views (such as the ones we developed for the exRNA Atlas); and in tracking of analyses and visualization in the context of networks and pathways (**Fig. 1D**).

To engage a wider number of consortium members in the continuous process of metadata capture and modeling, we will implement user-friendly interfaces on top of GenboreeKB. One interface will support metadata entry and curation and the other will support the design of *new* metadata models, not just preconfigured models, without programming and low-level database design. By lowering the engineering skill required we will engage a larger number of consortium members in this critical collaborative process (**Aim 4c)**.

Metadata is currently communicated to central repositories as *static* snapshot that contains information accumulated up to the submission date and is typically not updated as new valuable information becomes available after that date. One of the reasons for the static mode of communication is the current difficulty of collecting and communicating metadata history pre-submission and tracking its revisions post-submission. Distributed nature of the process may also pose a challenge. We will address these challenges by employing the newly developed W3C PROV-O[[7](#_ENREF_7)] and Linked Data Platform (LDP) 1.0[[98](#_ENREF_98)] standards.

As illustrated in **Fig. 4A**, LDP containers afford a means for discovering and organizing metadata documents exposed by a site. LDP requests reveal the top-level containers, any sub-containers, and the types of member documents (or the documents themselves). LDP can be used as a self-description of resources at a particular site, but as a linked data format, the objects of containment statements are URIs and thus may be stored elsewhere. GenboreeKB natively organizes metadata documents into collections, similar to “Samples” and “Studies” *ldp:DirectContainers* shown in **Fig. 4A**, and aggregates such related collections into a ‘knowledgebase’ which would be analogous to the top-level “EVB” *ldp:BasicContainer*. LDP affords *dynamic metadata discovery* by “drill down” from top-level containers, to collection-containers that organize the site’s resources, and to the schemas of the individual metadata documents.

Mass Spec

Analysis

Measurement

duke.edu

Sample

used

wasAssociatedWith

wasGeneratedBy

generated

**EVB**

*ldp:DirectContainer*

*ldp:DirectContainer*

*ldp:BasicContainer*

**Samples**

**Studies**

**. . .**

Sample Metadata

Documents

*(ldp#RDFSource)*

*lpd:contains*

**Figure 4. (A) LDP for dynamic discovery of available metadata and (B) PROV-O for tracking metadata history and modification.**

**A**

**B**

PROV-O also fits well with the GenboreeKB metadata modeling system and will be implemented as a natural extension of existing concepts of actors (users or software) and entities (documents) together with the existing document-level history tracking (**Fig. 4B**). Creation and modification of documents will be communicated via the PROV-O standard, as will the provenance of documents (and non-RDF resources) *used* by *activities*, thereby generating a complete provenance chain back to original measurement collections.

**Potential Problems & Alternative Approaches:** Because MoTrPAc involves clinical studies involving thousands of human subjects, there may be the need for research groups and organizations to maintain custody over protected identifiable information (using either local or cloud hardware) while sharing only a subset of non-identifiable data with other consortium members and the public. We will address this situation by a combination of local software deployment and virtual integration. Specifically, both GenboreeKB and Genboree Workbench will be deployed as virtual machines on local machines or commercial clouds using the Open Virtualization Framework (OVF) standard and the LDP-based aggregation features will support controlled sharing of physically-distributed non-identifiable information.

**Aim 2. Create the database of molecular profiles.**

The center will develop “omics” pipelines to process large volumes of data and will operate them during the course of the project to create the database of molecular profiles. To accomplish these goals, the center will combine the extensive and complementary experiences of Gerstein’s laboratory in developing computational methods, tools and pipelines for ENCODE and modENCODE projects, with Milosavljevic’s record of accomplishment in providing data coordination for the NIH Roadmap Epigenomics project. This project will extend their ongoing collaboration on the Extracellular RNA Communication project.

**(a) Develop data processing protocols and pipelines for genomic, epigenomic, transcriptomic, metabolomic, and proteomic data and implement associated quality control standards.**

**Preliminary studies:**

**Transcriptome Analysis.** We have extensive experience in developing RNA-Seq processing pipelines[[103](#_ENREF_103)] as part of the mod/ENCODE consortia [[26](#_ENREF_26), [34](#_ENREF_34)]. We have developed tools for identifying non-coding transcription and novel transcribed elements[[9](#_ENREF_9), [22](#_ENREF_22), [34](#_ENREF_34), [65](#_ENREF_65), [87](#_ENREF_87)], and contributed to the ExceRpt[[2](#_ENREF_2)] pipeline for the analysis of extracellular small RNA-Seq experiments. We have also developed a number of tools and data formats to handle large quantities of data generated by RNA-Seq experiments[[41](#_ENREF_41), [103](#_ENREF_103)].

**ChIP-Seq analysis of TFs and chromatin marks.** We have developed PeakSeq[[86](#_ENREF_86)], a tool for the genome-wide identification of TF binding sites from ChIP-Seq data, which is used by ENCODE. More recently, we developed MUSIC[[32](#_ENREF_32)], a peak caller that performs multiscale decomposition of ChIP-Seq signal, which is applicable to studies of histone modifications enabling detection of broad and punctate regions of enrichment.

**DNA Methylation Analysis.** We developed criteria for the design of epigenome-wide association studies[[44](#_ENREF_44), [67](#_ENREF_67)] and algorithms for whole-genome bisulfite sequencing[[24](#_ENREF_24), [60](#_ENREF_60)], and applied them to analyze over 100 methylomes as part of the NIH Roadmap Epigenomics project[[83](#_ENREF_83)]. We have performed integrative extensive analyses of the methylome, genetic variation[[23](#_ENREF_23), [62](#_ENREF_62)], and chromatin marks[[83](#_ENREF_83)].

**Proteomics Analysis.** We have experience with the analysis of proteomic data[[90](#_ENREF_90), [97](#_ENREF_97), [101](#_ENREF_101)] and its integration with genomic data such as the combination of mass spectrometry (MS) proteomic and transcriptomic data[[56](#_ENREF_56), [104](#_ENREF_104)]. We developed miBAT on top of RNA-Seq to improve the reference proteome [[55](#_ENREF_55)].

**Metabolomics Studies.** Over the past decade we have published nineteen studies of integrative physiology with metabolomics as a key component; including metabolic profiling in skeletal muscle[[46](#_ENREF_46) , [57](#_ENREF_57) ] and blood[[95 2010](#_ENREF_95)] related to exercise, caloric restriction[[47 2012](#_ENREF_47), [80](#_ENREF_80) ], aging and physical function[[66](#_ENREF_66)], and heart failure[[5](#_ENREF_5) ] and therapies. We have extensive experience in integrating metabolic profiling with genetics, transcriptomics, epigenetics (methylomics), proteomics and with physiologic and clinical outcome data, in order to identify biomarkers and mediators of physiologic adaptation to exercise[[10](#_ENREF_10) , [40 205](#_ENREF_40), [48](#_ENREF_48) , [49](#_ENREF_49) , [59](#_ENREF_59), [70](#_ENREF_70) , [94](#_ENREF_94) , [96](#_ENREF_96) ].

**Proposed Plan:** To facilitate the integrative analysis of the datasets produced by the consortium, we will evaluate the existing pipelines for the analysis of high throughput functional genomic and proteomic datasets from existing consortia such as ENCODE, Roadmap Epigenomics, ExRNA Communication, and 1000 Genomes Projects. We will modify and update these pipelines as necessary for the data commonly generated by MoTrPAc. We will help implement and deploy these analysis pipelines at the DCC so that they can be run in a streamlined uniform fashion. Using best practices and pipelines developed for the 1000 Genomes project, we will develop pipelines to call variants from WGS data. We will develop and evaluate different analysis methods for frequently performed analysis tasks, provide sound statistics for selecting among them, and work with the Data Working Group (**Aim 4c**) to ensure uniformity in the subsequent processing of each data type. We will develop and track metrics to quantify the quality of the data being produced for each of the high-throughput data types. Using these metrics in consultation with members of the consortium we will develop quality control (QC) standards to ensure that the generated data is of sufficiently high quality to useful to the greater scientific community. For example, the performance of ChIP-Seq data will be assessed using occurrences of sequence motifs within peaks, and the reproducibility of biological replicates.

**Potential Problems & Alternative Approaches:** If the pipelines we develop take too long to process all the data generated by MoTrPAc (especially if consortium data is submitted in large boluses), we will investigate faster algorithms; we will also attempt to ensure that there is a steady submission of data. If a significant portion of the submitted data does not meet the agreed upon QC standards we will investigate the cause to ensure that the standards are in line with equivalent high quality data from other consortia.

**(b) Organize standardization of data and metadata; process the raw data to populate the MoTrPAc database; and implement access-controlled sharing, analysis, distribution and archiving.**

**Preliminary Results:** The Baylor DCC will employ extensive data coordination experience and infrastructure for high-throughput data processing developed during the NIH Roadmap Epigenomics and ExRNA Communication projects. The projects involved construction of the Epigenome Atlas and exRNA Atlas portals for the data, large data submissions for archiving into GEO, SRA, and dbGaP archives and integration into the ENCODE portal and UCSC Browser Hubs.

**Proposed Plan:** Upon molecular profiling at Chemical Profiling Centers, the raw data and associated metadata will be submitted for inclusion into the MoTrPAc database using a FTP or Aspera Fastp data submission pipeline in BAM and other raw data formats. Sample-level quality control (QC) metrics (developed by the data Work Group, see **Aim 3c**), abundance estimates for various molecular species, and detailed alignment information will be computed and made available for visualization and validation. The metadata for Biosamples, Donors, Experiments, Runs, Studies and Analyses will be provided in JSON (JavaScript Object Notation[[50](#_ENREF_50)]) or predefined tabbed value formats and deposited into the GenboreeKB Metadata Tracking System for validation, tracking and curation. Upon metadata validation, the raw data will be processed using automated pipelines on the Bioinformatics Center cluster and—in case of spikes of computing demand—using “elastic” cloud computing services at Rackspace or Amazon Web services. The processing statistics and agreed upon QC metrics will be used to generate regular reports at various levels of granularity. To generate the reports we will make use of a modified version of an existing pipeline reporting system that we developed in the context of Roadmap and ExRNA projects.

Data processing will produce molecular profiles that will be conceptually organized within the MoTrPAc database along the following three high-level dimensions: (a) donors, PA interventions, and biosamples; (b) molecular profiling assays; (c) metabolites or genomic coordinates of genes and other genomic elements. Querying the database will yield a ‘‘data slice’’ of this three-dimensional space that is relevant for downstream analyses using tools integrated within the Genboree Workbench. Each of the three dimensions is covered by ontologies developed by the Data Work Group (**Aim 4c**). Ontology ‘‘slims’’, consisting of appropriate higher-level terms (illustrated in **Fig. 1B**) will facilitate the retrieval of ‘‘data slices’’ for downstream analysis. Abundance estimates for various elements, such as protein-coding genes, regulatory elements and miRNAs will be pre-computed, making it possible to deliver the ‘‘slices’’ very rapidly.

**Data release policy and submission to archives.** We have lead the development of strategies for data and metadata production and sharing, data release policies, and their implementation within the NIH Roadmap Epigenomics, NIH Extracellular RNA Communication, ENCODE, IHEC, modENCODE, and 1000 Genomes projects. We have extensive experience in jointly developing metadata standards, and have submitted tens of thousands of records for the Roadmap Epigenome and other projects to NCBI GEO and dbGaP archives. Using this experience we will work on developing policies that maximize the impact of data resources through early release of quality data while protecting the publication and attribution rights of authors. The largest volume of clinical data (WGS) will be archived in dbGaP, while animal-study data will be archived in NCBI GEO and metabolomics/proteomic profiles in HMDB and METLIN. All molecular profiles and analysis results are shared within the consortium and with the broader research community via the MoTrPAC portal (**Aim 4(a)**).

**Aim 3. Perform data analysis and construct response models.**

To construct physical activity response models, the center will analyze the molecular profiles obtained in **Aim 2** and integrate them with other functional genomic data. The center will synthesize the results into response models and address key questions about molecular mechanisms that mediate health benefits of PA, such as: (1) Do the relevant signaling mechanisms differ by intensity or amount of exercise? Are they different for resistance and aerobic exercise? (2) Are they reflected in the responses to acute exercise or to just chronic exercise? (3) Are there molecular markers (genetic, epigenetic, metabolomic, proteomic) predictive of the PA response variability? The results of these analyses (e.g. the molecular biomarkers for PA, QTLs, and the outputs from computational models will be submitted to the DCC for sharing with the consortium and the wider scientific community. Network modules and PA pathways will be disseminated via WikiPathways (**Aim 4(e)**).

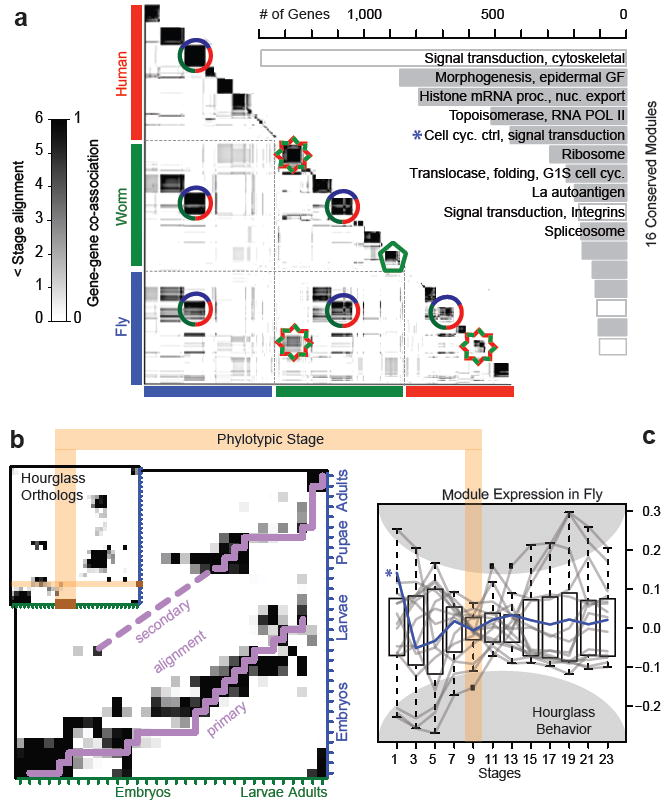
**(a) Normalize processed multi-omics profiles between individuals; deconvolute profiles of complex tissues; and register time points for time-course analyses.**

**Preliminary studies:** We have developed a number of advanced methods for normalization, analysis, and comparison of RNA-seq profiles. In particular: (1) incRNA, a method that predicts novel ncRNAs using known ncRNAs of various biotypes as a training set[[65](#_ENREF_65)]; (2) FusionSeq, a pipeline to detect transcripts that arise due to trans-splicing or chromosomal translocations[[77](#_ENREF_77), [89](#_ENREF_89)]; (3) IQSeq, a transcript isoform quantification tool that uses an EM algorithm to resolve the maximum likelihood expression level of individual transcript isoforms; (4) Pseudo-seq[[75](#_ENREF_75)], which addresses the issue of quantification of pseudogene and repetitive region expression; and (5) Aggregation and Correlation Toolbox (ACT), which is a general purpose tool for comparing genomic signal tracks[[52](#_ENREF_52)]. In addition, we contributed to the development of a classification and analysis scheme for “spike” event patterns in omics data with longitudinal profiles[[16](#_ENREF_16)].

Molecular profiles from many human and animal tissues will be collected as part of the MoTrPAC project. Skeletal muscle is a complex tissue comprised of myofibers that are heterogeneous in size, metabolism, and contractile function[[76](#_ENREF_76), [92](#_ENREF_92)]. To identify both intracellular and tissue composition changes in muscle and other profiled tissues it is highly desirable to deconvolute the “omics” profiles of heterogeneous tissue samples. Toward this end, we have developed and experimentally validated Epigenomic Deconvolution (in review), a novel *in silico* deconvolution method that provides estimates of genomic CpG methylation, gene transcription, and other “omic” profiles within a diversity of constitutive cell types. The method employs an iterative algorithm for constrained matrix factorization using quadratic programming and extends the related method of Houseman[[45](#_ENREF_45) ] by deconvoluting gene expression and other “omic” profiles in addition to just CpG methylation profiles.

**PROPOSED PLAN:** The molecular profiles obtained by RNA-Seq, ChIP-Seq, and other omics primary data processing pipelines will be normalized and registered between time points and between individuals. Normalization is critical in order to identify differential biomarkers in response to PA. We will also evaluate existing tools for differential “omic” analysis[[64](#_ENREF_64), [84](#_ENREF_84)] as well as develop new methods if necessary in order to identify the molecular biomarkers that show significant differences either as acute or durable responses to PA.

One of the main analysis problems will be to develop methods to deal with longitudinal time course multi-omics datasets. Toward this end, we will normalize omics data from several experiments individually, and then account for uneven sampling and time gaps using a Lomb-Scargle periodogram[[16](#_ENREF_16), [39](#_ENREF_39), [100](#_ENREF_100)]. Each periodogram will then be available for standard time-series analysis and data clustering such as the hierarchical clustering used to obtain common trends and assess biological relevance using such tools as Gene Ontology, Reactome, KEGG and WikiPathways for pathway analysis[[16](#_ENREF_16), [33](#_ENREF_33), [35](#_ENREF_35)]. This framework will normalize and compare many different types of ‘omics datasets. To identify specific effects within massive quantities of longitudinal data we will develop tools that use bootstrap simulations to assess power and significance, taking into account the auto-correlated behavior of the data-points and periodogram analyses described above, where the number of datapoints can be leveraged to reduce the prediction error at each individual point.



**Figure 5.** **Expression clustering by OrthoClust.** **Left**: Human, worm, and fly gene-gene co-association matrix; darker coloring reflects increased likelihood that a pair of genes are assigned to the same module. A dark block along the diagonal represents a gene group within a species. **Right**: The GO functional enrichment of genes within the 16 conserved modules is shown.

To identify both intracellular and tissue composition changes in response to exercise we will apply the Epigenomic Deconvolution method, which utilizes lists of loci exhibiting variation in CpG methylation levels across constituent cell types compiled from reference methylomes produced by the NIH Roadmap Epigenomics project[[83](#_ENREF_83)] and from a growing multitude of array-based profiles in NCBI GEO and other public archives. Starting from methylation profiles of tissue homogenates we will estimate both cell type proportions and methylation profiles of constituent cell types. The proportion estimates will then be used as a “key” to deconvolute gene expression and other “omic” profiles of constituent cell types.

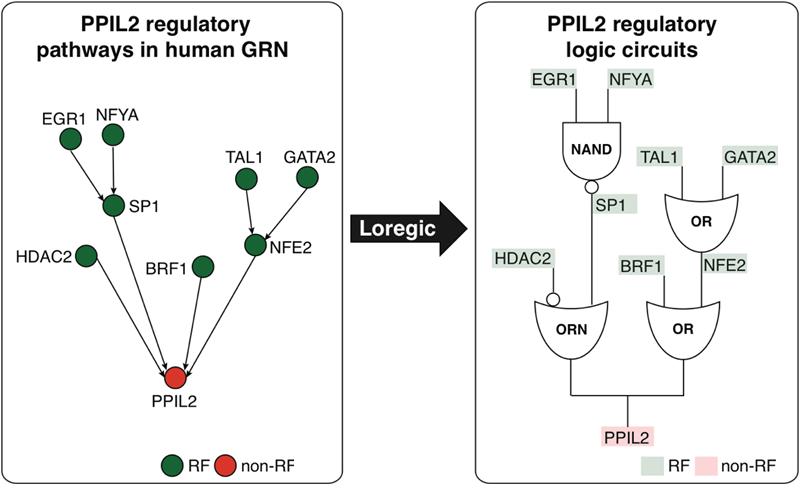
**Potential Problems & Alternative Approaches:** If some of the time course data from certain subjects is missing, we will develop methods to impute the data using time course profiles from other matching individuals. The imputation method will also identify potentially mis-labelled datasets and sample swaps. If we find that some time course data lack resolution for certain analyses, we will contact the data submitters about the possibility of including additional time points.

**Aim 3(b). Analyze and cluster gene activity data to identify coordinated acute and durable response modules for both single-perturbation and time-course experiments.**

**Preliminary studies:** We have extensive experience in analyzing gene expression data to identify co-regulated modules. For example, we recently developed a new method, OrthoClust, to simultaneously cluster cross-species gene networks into gene modules[[105](#_ENREF_105)]. We applied OrthoClust to human and cross-species gene co-expression networks from the mod/ENCODE projects and discovered novel human developmental transcriptional programs (**Fig. 5**).

Gene regulatory factors work cooperatively, forming a complex regulatory circuit controlling gene expression. We developed Loregic, a general-purpose method to characterize the cooperativity of regulatory factors[[102](#_ENREF_102)]. Using ENCODE ChIP-Seq and TCGA RNA-Seq data, we demonstrated how Loregic characterizes complex circuits involving TFs and miRNAs in human cancer. We developed continuous model-based approaches such as DREISS[[4](#_ENREF_4)] to identify gene expression dynamics driven by external and internal regulatory modules, helping dissect the temporal dynamic effects of different regulatory subsystems on gene expression. In addition, we associated exercise phenotypes with gene expression data. For example, we identified metabolic and genomic molecular signatures in skeletal muscle that mediate the intensity and dose specific effects of exercise training on cardiorespiratory fitness (peak VO2), insulin sensitivity, and serum triglycerides[[46](#_ENREF_46)].

**Proposed PLAN:** We will analyze both single-perturbation and temporal dynamic patterns from longitudinal time-course expression data and identify expression patterns associated with PA response and their regulatory mechanisms. In particular, we will construct the gene co-expression networks and find modules (with associated expression signatures) enriched in exercise phenotypes. Finally, we will identify gene regulatory logics driving PA response phenotypes [[46](#_ENREF_46)] via Loregic **(Fig. 6)**. Using ENCODE and other publicly available molecular profiles such as ChIP-seq data we will construct the regulatory networks for biomarker genes.



**Figure 6. Depiction of two logic circuit regulatory pathways.** Two logic circuit regulatory pathways targeting the PPIL2 gene, an important cyclophilin member in immunological suppression, are found by **Loregic**.

**Aim 3(c). Build integrative models and identify biomarkers that predict response to physical activity based on multi-omics profiles in the context of other relevant public data.**

We aim to integrate MoTrPAc data with relevant external datasets to construct an integrative molecular map of human physical activity, and will leverage our prior accomplishments in creating such models.

**Preliminary studies:** We have comprehensive experience integrating transcriptomic, metabolomics, and proteomic data in the context of exercise as well as other settings. For example, we integrated skeletal muscle metabolomics with RNA-Seq to identify fatty acid oxidation as an important mediator of the effect of exercise on insulin sensitivity and clinical indicators of cardiovascular health [[46](#_ENREF_46), [48](#_ENREF_48), [49](#_ENREF_49)]. We also integrated unknown metabolites, which can constitute as much as 50% of spectral features[[16](#_ENREF_16)], with transcriptomics profiles from different experimental conditions[[36](#_ENREF_36)]. By defining statistics to correlate the co-occurrence patterns of metabolites and genes we generated hypotheses about the identities of unannotated biosynthetic pathways. In addition, we have experience with the analysis of proteomic data and its integration with transcriptomics [[56](#_ENREF_56), [90](#_ENREF_90), [97](#_ENREF_97), [104](#_ENREF_104)]. This allowed us to identify previously uncharacterized proteins in a temporally and spatially resolved manner[[104](#_ENREF_104)].

We have extensively used machine-learning to generate models from integrated datasets. For example, we integrated ENCODE data on transcription factor (TF) binding, histone modifications, and target gene expression to establish regulatory relationships using a probabilistic model we named TIP (Target Identification from Profiles)[[19](#_ENREF_19)]. We identified potential enhancers from distal gene regions and we used these modules to quantify the relationship between TF binding and gene expression [[17](#_ENREF_17), [18](#_ENREF_18), [25](#_ENREF_25), [106](#_ENREF_106)]. We integrated these data types with protein-protein interaction and transcriptional regulation networks [[20](#_ENREF_20), [21](#_ENREF_21), [27](#_ENREF_27), [33](#_ENREF_33)]. This allowed us to group TFs into histone-sensitive and -insensitive classes that refined the prediction of gene-regulation targets and effects. Finally, we were able to build cross-organism integrative chromatin models (see **Aim 4(f)**)[[105](#_ENREF_105)].

**Proposed PLAN:** To create the PA molecular map, we will integrate transcriptomic and epigenetic MoTrPAc data with large datasets from other consortia. Specifically, we will incorporate the ENCODE TF data, the GTEx tissue-specific profiles, and the epigenetic marks of transcriptional regulatory elements from the Epigenome Roadmap. From this dataset we will construct integrative models relating epigenetics and transcriptomics using our previously developed machine learning approaches. Briefly, combined sets of genomic features in small (100bp) bins will be correlated with expression values over those regions. We will then generate statistical models relating epigenetic marks, TF binding, and gene expression, and further extend these models to incorporate proteomic and metabolomic data. To build our integrated models, the proteomic and metabolomic data will be combined with pathway information, such as KEGG and WikiPathways[[97](#_ENREF_97)]. These pathways will be linked to transcriptomic data through their associated genes, using the same machine learning approaches to relate transcriptional activity to protein and metabolite abundances. Thus we can integrate metabolomic and proteomic with epigenetic and regulatory data. Finally, the large depth and coverage of transcriptomic experiments will be leveraged to develop integrative models of PA response (e.g. increased metabolite or protein quantities). Using the deconvoluted data from **Aim 3(a)**, these models will be scaled to determine responses at the cellular, tissue, and organ levels.

**Aim 3(d). Integrate genetic information into exercise response models to identify the variants associated with individual response variation.** The center will correlate genetic variation with molecular and phenotypic effects to identify the variants that predict an individual’s response to exercise. To this end, we will “triangulate” integrative response models described in Aim 3c with genetic variation by identifying Quantitative Trait Loci (QTLs) that inter-relate molecular profiles and individual’s response to exercise.

**Preliminary studies:** We have extensive experience relating genetic variation to multi-omic datasets in the context of exercise and in other settings, generating extensive maps of metabolomic and expression Quantitative Trait Loci (mQTLs and eQTLs). In the context of PA, we have identified genomic predictors of the cardiovascular response to exercise training[[13](#_ENREF_13)], and variants associated with poor response to training[[12](#_ENREF_12)]. In addition, we integrated metabolomics and epigenetic markers with genetic variants to identify mQTLs associated with poor coronary artery disease outcomes[[59](#_ENREF_59)]. Finally, we have experience linking eQTLs to phenotype data by quantifying the amount of information necessary to identify an individual[[42](#_ENREF_42)].

eQTLs and mQTLs can only be derived for common variants (given sufficient statistical power). As one of the PIs on the Yale Center for Mendelian Genomics (CMG), which focuses on studying rare variants, Dr, Gerstein has developed methods for analyzing rare variants from WGS data. A key approach for analyzing rare variants is the framework of burden tests and for this we have developed the LARVA tool[[63](#_ENREF_63)]. Furthermore, in the case of rare genetic variants, allelic analyses also provide a primary method for connecting genetic information with functional genomics. This is done by comparing either the RNA-Seq or ChIP-Seq data from loci that contain a heterozygous SNP and using the reads to differentiate between the maternal and paternal haplotypes in order to identify those loci that show an imbalance between the two alleles (see Figure). We have pioneered the development of tools to analyze (AlleleSeq[[85](#_ENREF_85)]) and catalogue allelic events across large number of individual genomes using available functional genomic data. (AlleleDB[[51](#_ENREF_51)])

We developed a method that integrates a unified biological network of various gene-gene interactions (regulatory, genetic, phosphorylation, signaling, metabolic, and physical protein-protein interactions) together with the set of variants from an individual in order to predict potentially deleterious mutations[[54](#_ENREF_54)]. To this end we have built a tool called FunSeq (“Function based Prioritization of Sequence Variants”)[[29](#_ENREF_29), [54](#_ENREF_54)] that prioritizes rare genomic variants in terms of their likelihood of causing significant phenotypic effects.

**Proposed PLAN:** We will integrate processed functional genomic data with the genetic information from the same individuals in order to identify eQTLs, and will extend this strategy to identify mQTLs and variants associated with proteomic changes (pQTLs). To characterize the rare variants within the individuals studied we will perform burden tests with LARVA to identify genomic regions that are over or under represented in terms of the number of rare variants. Furthermore, we will perform an allelic analysis (using AlleleSeq) of available functional genomic data to identify allelic heterozygous variants within each individual. For the rare detected allelic variants we will compare these to the PA response in order to identify potential associations. Finally, we will use FunSeq in order to integrate rare variants and associated functional genomic data to rank those that are most likely to be significant for either acute or durable responses to PA.

**Aim 4. Organize and provide informatics support for cross-consortium projects.**

Cross-consortium activities will be essential for MoTrPAc success, as they will produce the public portal for the project; enable tracking of biospecimens as they traverse across consortium sites; ensure quality and accessibility of molecular profiles; engage the consortium in their analysis; organize the network and pathway knowledge and integrate this knowledge into the public domain for use by the research community; and engage the consortium in the design of animal models and follow-up clinical studies.

**Aim 4(a). Deploy a consortium portal for use by the Coordination Center and other consortium members.** Working with the Coordination Center and other MoTrPAC sites, the Baylor DCC will establish the portal, modeled on the ExRNA portal (exrna.org) we designed for the ExRNA Communications project. Briefly, the portal will be commercially hosted using WordPress for 24/7 up-time and backed by our Redmine content management system, which will also serve as an integration point with Genboree-hosted content (GenboreeKB being a Redmine plugin). The DCC will provide tools for the management of both structured and unstructured content, collect access statistics, and work to ensure an excellent visitor experience and project visibility.

**Aim 4(b). Develop a system for tracking animal and human specimens across the consortium sites.** Working with the Coordination Center, the host of the MoTrPAC biorepository, the Baylor DCC will establish a “virtual biorepository” system for tracking biosamples across the Clinical and Chemical Characterization sites, maintaining a direct link between the clinical data and the molecular profiles. For this purpose we will reuse the GenboreeKB-based “Extracellular RNA Virtual Biorepository” system (EVB) that we recently developed for the exRNA project. This system will greatly augment the value of the biosamples, a key reusable resource developed by MoTrPAC. The reusability the biosamples will be enhanced through the molecular profiling and accessibility provided via faceted search and other query modalities provided by the virtual biorepository.

**Aim 4(c). Organize a data working group (DWG) to engage the consortium in the development of the MoTrPAc database.** The MoTrPAc database will be organized along multiple high-level dimensions. A metadata-enabled query (**Fig. 1B**) will provide a ‘‘data slice’’ in this space that is relevant for downstream analyses. To enhance the quality and utility of this resource, a data working group (DWG) will be created within MoTrPAc. The DWG will help establish data quality and standards by reusing standard “omics” formats and data elements from the NIH CDE library, and will also develop rich metadata along multiple dimensions to augment the accessibility of this resource via various manual and programmatic modalities. Dr. Kim Huffman will coordinate the DWG effort, with contribution from other Duke CIMP members and will be supported by Dr. Kei-Hoi Cheung (DCC member from Yale) and by the GenboreeKB metadata system hosted at the Baylor DCC. Dr. Cheung is past chair of the BioRDF task force as part of the Semantic Web for Health Care and Life Science Interest Group. This group is chartered by the W3C tasked with identifying issues and proposing best practices for converting biomedical datasets into RDF. Dr. Cheung is also a co-PI of the CEDAR project, which is incorporating RDF and JSON-LD standards into metadata representations and he will be ensuring metadata interoperability between our projects (see letter of collaboration from Dr. Musen, CEDAR PI).

# **Aim 4(d). Provide expertise in data management and analytics and facilitate communication about data analysis methods and tools by organizing an analysis working group (AWG).** In collaboration with the Duke CIMP, the Yale DAC will organize the AWG, which is tasked with combining analytical know-how across the consortium to perform integrative analyses of molecular profiles. The Yale DAC has extensive experience leading and participating in similar AWGs, including: the ENCODE [[25](#_ENREF_25)], modENCODE [[14](#_ENREF_14)], the 1000 Genomes Project [[30](#_ENREF_30), [31](#_ENREF_31)], BrainSpan [[93](#_ENREF_93)], PsychENCODE [[79](#_ENREF_79)] projects; the Extracellular RNA Communication program (<http://commonfund.nih.gov/Exrna/>); and the DOE Knowledgebase Kbase (<http://kbase.us/)>. The Duke CIMP will provide complementary clinical perspective using their prior experience with exercise­induced molecular and health effects in animal models [[57](#_ENREF_57) , [58](#_ENREF_58) , [69](#_ENREF_69) , [70](#_ENREF_70) , [71](#_ENREF_71) ]. The AWG will aim to replicate Duke team’s successful experience in crossing the boundary between exercise physiology and “omic” biology, and in facing the challenge posed by the volume and heterogeneity of molecular profiling data collected in a clinical context. The early focus will be on data management; for instance, identification of correct data ontologies and evaluating fidelity of metadata elements. Later, the focus will be on building integrative models across tissues and animal models. Finally, the focus will be on specific scientific questions generated by the consortium members.

**Aim 4(e). Organize the development by the consortium of a section within WikiPathways devoted to molecular pathways relevant for physical activity.** We will organize a consortium-wide effort to develop a WikiPathways section devoted to network and pathway models relevant PA biology. We will use these models for computational pathway enrichment analysis, Cytoscape visualization, and integration with the BD2K, WikiData, Bio-GPS, and GeneWiki databases. Dr. Pico, co-founder and developer of WikiPathways and his group at Gladstone/UCSF will oversee the development of required features at WikiPathways, will organize MoTrPAC pathway curation workshops and will provide help desk and quality control for MoTrPAC curators. WikiPathways represents an early success case for crowdsourcing biomedical knowledge. In 2015, WikiPathways surpassed KEGG in terms of unique gene content and pathway diversity [[61](#_ENREF_61)]. As a co-PI of the National Resource for Network Biology, Dr. Pico will contribute algorithms for pathway enrichment for exercise responses. Dr. Pico is a core developer of Cytoscape, which employed as a platform for visualization and analysis of omics datasets in the context of biological networks and pathways. Dr. Andrew Su, a BD2K project PI, will collaborate on using RDF as a means of integrating the pathway knowledge with WikiData and gene-centric knowledge and profiles via Bio-GPS and GeneWiki (see letter of collaboration from Dr. Su).

**Aim 4(f). Support integration of animal study data and develop plans for replication studies.** We will work with Consortium members at common meetings and on AWG calls to develop analytic strategies that will promote replication studies in humans and animal models. The Duke CIMP will lead this effort and lend their extensive prior experience with exercise­induced molecular and health effects in both clinical studies and animal models [[57](#_ENREF_57) , [58](#_ENREF_58) , [69](#_ENREF_69) , [70](#_ENREF_70) , [71](#_ENREF_71)]. This expertise will be complemented by the cross-species molecular models developed by the Yale DAC, including modENCODE comparison of human and model organism transcriptomes[21, [26](#_ENREF_26), [27](#_ENREF_27), [33](#_ENREF_33), [34](#_ENREF_34)] and cross-species modeling of co-expression networks using OrthoClust [105]. This combined expertise will enable the group to suggest credible and useful platforms for testing and replication of physiologic questions from model system to man.

**Software Dissemination.** All software tools developed by the Center will be distributed as free open source via GitHub. The MoTrPAC Data Analysis Center (led by the Gerstein lab at Yale) also operates a centralized repository for software tools and resources (<http://info.gersteinlab.org/Resources>, accessed >7000 times) many of which are already distributed via GitHub (github.com/gersteinlab). We will make the tools general purpose so that they can be used in a variety of contexts. We have pursuing that practice with our other tools; for example, OrthoClust can cluster any multi-layer network besides cross-species gene co-expression networks. Some tools have already been downloaded thousands of times and widely used by other researchers; e.g., our paper on the PeakSeq [86] has been cited over 450 times by the end of 2015. GenboreeKB, Genboree Workbench, and other software developed by the Data Coordination Center (Milosavljevic laboratory at Baylor) will continue to be distributed under a GNU GPL license and in the form of OVF compliant virtual appliances for easy installation at consortium sites and on commercial clouds.

**Milestones and Timeline**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Milestones** | | | | | | | | **Y1** | **Y2** | **Y3** | **Y4** | **Y5** | **Y6** |
| **Aim 1.** **Develop an informatics infrastructure.** | | | | | | | |  |  |  |  |  |  |
| **a.** Develop a cloud-based infrastructure for data processing & analysis. | | | | | | | |  |  |  |  |  |  |
| **b.** Develop data storage system for cloud-based management of data & metadata. | | | | | | | |  |  |  |  |  |  |
| **Aim 2 Create the database of molecular profiles.** | | | | | | | |  |  |  |  |  |  |
| **a.** Develop data processing protocols and pipelines. | | | | | | | |  |  |  |  |  |  |
| **b(i).** Organize standardization of data and metadata. | | | | | | | |  |  |  |  |  |  |
| **b(ii).** Process the raw data to populate the MoTrPAc database. | | | | | | | |  |  |  |  |  |  |
| **b(iii).** Implement sharing of the molecular profiles. | | | | | | | |  |  |  |  |  |  |
| **b(iv).** Perform archiving of the molecular profiles. | | | | | | | |  |  |  |  |  |  |
| **Aim 3 Perform data analysis and construct response models.** | | | | | | | |  |  |  |  |  |  |
| **a.** Normalize processed multi-omics profiles. | | | | | | | |  |  |  |  |  |  |
| **b.** Cluster gene activity data to identify response modules. | | | | | | | |  |  |  |  |  |  |
| **c.** Build integrative models and identify biomarkers. | | | | | | | |  |  |  |  |  |  |
| **d.** Integrate genetic information into exercise response models. | | | | | | | |  |  |  |  |  |  |
| **Aim 4.** **Organize and provide informatics support for cross-consortium projects.** | | | | | | | |  |  |  |  |  |  |
| **a.**  Deploy a consortium portal. | | | | | | | |  |  |  |  |  |  |
| **b.** Develop a system for tracking animal and human specimens. | | | | | | | |  |  |  |  |  |  |
| **c.** Organize a data working group (DWG). | | | | | | | |  |  |  |  |  |  |
| **d.** Organize an analysis working group (AWG). | | | | | | | |  |  |  |  |  |  |
| **e.** Develop PA pathway section within WikiPathways. | | | | | | | |  |  |  |  |  |  |
| **f.** Support integration of animal study data & develop plans for replication studies. | | | | | | | |  |  |  |  |  |  |
|  | = Planning |  | = In Progress |  | = Completing |  | = No Activity | | | | | | | |

**BIBLIOGRAPHY**

1. *exRNA Research Portal.* (URL). [**http://www.exrna.org**](http://www.exrna.org).

2. *extra-cellular RNA processing toolkit (exceRpt).* (URL). [**http://github.gersteinlab.org/exceRpt/**](http://github.gersteinlab.org/exceRpt/).

3. *Comprehensive genomic characterization defines human glioblastoma genes and core pathways.* Nature, 2008. **455**(7216): p. 1061-8.

4. *DREISS.* PLoS Comput Biol, (in revision). **https://github.com/gersteinlab/Dreiss**.

5. Ahmad, T., et al., *Prognostic Implications of Long-Chain Acylcarnitines in Heart Failure and Reversibility With Mechanical Circulatory Support.* J Am Coll Cardiol, 2016. **67**(3): p. 291-9.

6. Amin, V., et al., *Epigenomic footprints across 111 reference epigenomes reveal tissue-specific epigenetic regulation of lincRNAs.* Nat Commun, 2015. **6**: p. 6370.

7. Belhajjame, K., et al., *PROV-O: The PROV Ontology.* W3C Recommendation, 2013. **Apr**.

8. Bernstein, B.E., et al., *The NIH Roadmap Epigenomics Mapping Consortium.* Nat Biotechnol, 2010. **28**(10): p. 1045-8.

9. Bertone, P., et al., *Global identification of human transcribed sequences with genome tiling arrays.* Science, 2004. **306**(5705): p. 2242-6.

10. Bhattacharya, S., et al., *Validation of the association between a branched chain amino acid metabolite profile and extremes of coronary artery disease in patients referred for cardiac catheterization.* Atherosclerosis, 2014. **232**(1): p. 191-6.

11. Booth, F.W., et al., *Waging war on modern chronic diseases: primary prevention through exercise biology.* J Appl Physiol (1985), 2000. **88**(2): p. 774-87.

12. Bouchard, C., et al., *Adverse metabolic response to regular exercise: is it a rare or common occurrence?* PLoS One, 2012. **7**(5): p. e37887.

13. Bouchard, C., et al., *Genomic predictors of the maximal O(2) uptake response to standardized exercise training programs.* J Appl Physiol (1985), 2011. **110**(5): p. 1160-70.

14. Celniker, S.E., et al., *Unlocking the secrets of the genome.* Nature, 2009. **459**(7249): p. 927-30.

15. Challis, D., et al., *An integrative variant analysis suite for whole exome next-generation sequencing data.* BMC Bioinformatics, 2012. **13**: p. 8.

16. Chen, R., et al., *Personal omics profiling reveals dynamic molecular and medical phenotypes.* Cell, 2012. **148**(6): p. 1293-307.

17. Cheng, C., et al., *Understanding transcriptional regulation by integrative analysis of transcription factor binding data.* Genome Res, 2012. **22**(9): p. 1658-67.

18. Cheng, C. and M. Gerstein, *Modeling the relative relationship of transcription factor binding and histone modifications to gene expression levels in mouse embryonic stem cells.* Nucleic Acids Res, 2012. **40**(2): p. 553-68.

19. Cheng, C., R. Min, and M. Gerstein, *TIP: a probabilistic method for identifying transcription factor target genes from ChIP-seq binding profiles.* Bioinformatics, 2011. **27**(23): p. 3221-7.

20. Cheng, C., et al., *Genome-wide analysis of chromatin features identifies histone modification sensitive and insensitive yeast transcription factors.* Genome Biol, 2011. **12**(11): p. R111.

21. Cheng, C., et al., *A statistical framework for modeling gene expression using chromatin features and application to modENCODE datasets.* Genome Biol, 2011. **12**(2): p. R15.

22. Clark, M.B., et al., *The reality of pervasive transcription.* PLoS Biol, 2011. **9**(7): p. e1000625; discussion e1001102.

23. Coarfa, C., et al., *Analysis of interactions between the epigenome and structural mutability of the genome using Genboree Workbench tools.* BMC Bioinformatics, 2014. **15 Suppl 7**: p. S2.

24. Coarfa, C., et al., *Pash 3.0: A versatile software package for read mapping and integrative analysis of genomic and epigenomic variation using massively parallel DNA sequencing.* BMC Bioinformatics, 2010. **11**: p. 572.

25. Consortium, E.P., *An integrated encyclopedia of DNA elements in the human genome.* Nature, 2012. **489**(7414): p. 57-74.

26. Djebali, S., et al., *Landscape of transcription in human cells.* Nature, 2012. **489**(7414): p. 101-8.

27. Dong, X., et al., *Modeling gene expression using chromatin features in various cellular contexts.* Genome Biol, 2012. **13**(9): p. R53.

28. Evani, U.S., et al., *Atlas2 Cloud: a framework for personal genome analysis in the cloud.* BMC Genomics, 2012. **13 Suppl 6**: p. S19.

29. Fu, Y., et al., *FunSeq2: a framework for prioritizing noncoding regulatory variants in cancer.* Genome Biol, 2014. **15**(10): p. 480.

30. Genomes Project, C., *A map of human genome variation from population-scale sequencing.* Nature, 2010. **467**(7319): p. 1061-73.

31. Genomes Project, C., et al., *A global reference for human genetic variation.* Nature, 2015. **526**(7571): p. 68-74.

32. Gerstein, M.B., et al., *Architecture of the human regulatory network derived from ENCODE data.* Nature, 2012. **489**(7414): p. 91-100.

33. Gerstein, M.B., et al., *Integrative analysis of the Caenorhabditis elegans genome by the modENCODE project.* Science, 2010. **330**(6012): p. 1775-87.

34. Gerstein, M.B., et al., *Comparative analysis of the transcriptome across distant species.* Nature, 2014. **512**(7515): p. 445-8.

35. Giaever, G., et al., *Functional profiling of the Saccharomyces cerevisiae genome.* Nature, 2002. **418**(6896): p. 387-91.

36. Gianoulis, T.A., et al., *Genomic analysis of the hydrocarbon-producing, cellulolytic, endophytic fungus Ascocoryne sarcoides.* PLoS Genet, 2012. **8**(3): p. e1002558.

37. Gibbs, R.A., et al., *Evolutionary and biomedical insights from the rhesus macaque genome.* Science, 2007. **316**(5822): p. 222-34.

38. Gibbs, R.A., et al., *Genome sequence of the Brown Norway rat yields insights into mammalian evolution.* Nature, 2004. **428**(6982): p. 493-521.

39. Glynn, E.F., J. Chen, and A.R. Mushegian, *Detecting periodic patterns in unevenly spaced gene expression time series using Lomb-Scargle periodograms.* Bioinformatics, 2006. **22**(3): p. 310-6.

40. Glynn, E.L., et al., *Impact of combined resistance and aerobic exercise training on branched-chain amino acid turnover, glycine metabolism and insulin sensitivity in overweight humans.* Diabetologia, 2015. **58**(10): p. 2324-35.

41. Habegger, L., et al., *RSEQtools: a modular framework to analyze RNA-Seq data using compact, anonymized data summaries.* Bioinformatics, 2011. **27**(2): p. 281-3.

42. Harmanci, A. and M. Gerstein, *Quantification of private information leakage from phenotype-genotype data: linking attacks.* Nat Methods, 2016. **13**(3): p. 251-6.

43. Harris, R.A., J. Rogers, and A. Milosavljevic, *Human-specific changes of genome structure detected by genomic triangulation.* Science, 2007. **316**(5822): p. 235-7.

44. Harris, R.A., et al., *Comparison of sequencing-based methods to profile DNA methylation and identification of monoallelic epigenetic modifications.* Nat Biotechnol, 2010. **28**(10): p. 1097-105.

45. Houseman E A, et al., *Reference-free deconvolution of DNA methylation data and mediation by cell composition effe.* bioRxiv, 2016.

46. Huffman, K.M., et al., *Metabolite signatures of exercise training in human skeletal muscle relate to mitochondrial remodelling and cardiometabolic fitness.* Diabetologia, 2014. **57**(11): p. 2282-95.

47. Huffman, K.M., et al., *Caloric restriction alters the metabolic response to a mixed-meal: results from a randomized, controlled trial.* PLoS One, 2012. **7**(4): p. e28190.

48. Huffman, K.M., et al., *Relationships between circulating metabolic intermediates and insulin action in overweight to obese, inactive men and women.* Diabetes Care, 2009. **32**(9): p. 1678-83.

49. Huffman, K.M., et al., *Exercise-induced changes in metabolic intermediates, hormones, and inflammatory markers associated with improvements in insulin sensitivity.* Diabetes Care, 2011. **34**(1): p. 174-6.

50. International, E., *ECMA-404 The JSON Data Interchange Standard.* (URL), 2013. [**http://json.org**](http://json.org)

(October).

51. J, C. and e. al., *A uniform survey of allele-specific binding and expression over 1000-Genomes-Project individuals.* Nat Commun. **(in press)**.

52. Jee, J., et al., *ACT: aggregation and correlation toolbox for analyses of genome tracks.* Bioinformatics, 2011. **27**(8): p. 1152-4.

53. Joyner, M.J. and D.J. Green, *Exercise protects the cardiovascular system: effects beyond traditional risk factors.* J Physiol, 2009. **587**(Pt 23): p. 5551-8.

54. Khurana, E., et al., *Integrative annotation of variants from 1092 humans: application to cancer genomics.* Science, 2013. **342**(6154): p. 1235587.

55. Kitchen, R. and e. al., *miBAT: Multi-modal profiling of the translatome at isoform resolution.* (in preparation).

56. Kitchen, R.R., et al., *Decoding neuroproteomics: integrating the genome, translatome and functional anatomy.* Nat Neurosci, 2014. **17**(11): p. 1491-9.

57. Kovalik, J.P., et al., *Metabolic remodeling of human skeletal myocytes by cocultured adipocytes depends on the lipolytic state of the system.* Diabetes, 2011. **60**(7): p. 1882-93.

58. Kraus, W.E., T.S. Bernard, and R.S. Williams, *Interactions between sustained contractile activity and beta-adrenergic receptors in regulation of gene expression in skeletal muscles.* Am J Physiol, 1989. **256**(3 Pt 1): p. C506-14.

59. Kraus, W.E., et al., *Metabolomic Quantitative Trait Loci (mQTL) Mapping Implicates the Ubiquitin Proteasome System in Cardiovascular Disease Pathogenesis.* PLoS Genet, 2015. **11**(11): p. e1005553.

60. Kunde-Ramamoorthy, G., et al., *Comparison and quantitative verification of mapping algorithms for whole-genome bisulfite sequencing.* Nucleic Acids Res, 2014. **42**(6): p. e43.

61. Kutmon, M., et al., *WikiPathways: capturing the full diversity of pathway knowledge.* Nucleic Acids Res, 2016. **44**(D1): p. D488-94.

62. Li, J., et al., *Genomic hypomethylation in the human germline associates with selective structural mutability in the human genome.* PLoS Genet, 2012. **8**(5): p. e1002692.

63. Lochovsky, L., et al., *LARVA: an integrative framework for large-scale analysis of recurrent variants in noncoding annotations.* Nucleic Acids Res, 2015. **43**(17): p. 8123-34.

64. Love, M.I., W. Huber, and S. Anders, *Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.* Genome Biol, 2014. **15**(12): p. 550.

65. Lu, Z.J., et al., *Prediction and characterization of noncoding RNAs in C. elegans by integrating conservation, secondary structure, and high-throughput sequencing and array data.* Genome Res, 2011. **21**(2): p. 276-85.

66. Lum, H., et al., *Plasma acylcarnitines are associated with physical performance in elderly men.* J Gerontol A Biol Sci Med Sci, 2011. **66**(5): p. 548-53.

67. Michels, K.B., et al., *Recommendations for the design and analysis of epigenome-wide association studies.* Nat Methods, 2013. **10**(10): p. 949-55.

68. Milosavljevic, A., *Putting epigenome comparison into practice.* Nat Biotechnol, 2010. **28**(10): p. 1053-6.

69. Muoio, D.M., et al., *Fatty acid homeostasis and induction of lipid regulatory genes in skeletal muscles of peroxisome proliferator-activated receptor (PPAR) alpha knock-out mice. Evidence for compensatory regulation by PPAR delta.* J Biol Chem, 2002. **277**(29): p. 26089-97.

70. Muoio, D.M., et al., *Muscle-specific deletion of carnitine acetyltransferase compromises glucose tolerance and metabolic flexibility.* Cell Metab, 2012. **15**(5): p. 764-77.

71. Muoio, D.M., et al., *Peroxisome proliferator-activated receptor-alpha regulates fatty acid utilization in primary human skeletal muscle cells.* Diabetes, 2002. **51**(4): p. 901-9.

72. Muzny, D.M., et al., *The DNA sequence, annotation and analysis of human chromosome 3.* Nature, 2006. **440**(7088): p. 1194-8.

73. Neufer, P.D., et al., *Understanding the Cellular and Molecular Mechanisms of Physical Activity-Induced Health Benefits.* Cell Metab, 2015. **22**(1): p. 4-11.

74. Overby, C.L., et al., *Providing Access to Genomic Variant Knowledge in a Healthcare Setting: A Vision for the ClinGen Electronic Health Records Workgroup.* Clin Pharmacol Ther, 2016. **99**(2): p. 157-60.

75. Pei, B., et al., *The GENCODE pseudogene resource.* Genome Biol, 2012. **13**(9): p. R51.

76. Pette, D. and R.S. Staron, *Myosin isoforms, muscle fiber types, and transitions.* Microsc Res Tech, 2000. **50**(6): p. 500-9.

77. Pflueger, D., et al., *Discovery of non-ETS gene fusions in human prostate cancer using next-generation RNA sequencing.* Genome Res, 2011. **21**(1): p. 56-67.

78. Pico, A.R., et al., *WikiPathways: pathway editing for the people.* PLoS Biol, 2008. **6**(7): p. e184.

79. Psych, E.C., et al., *The PsychENCODE project.* Nat Neurosci, 2015. **18**(12): p. 1707-12.

80. Redman, L.M., et al., *Effect of caloric restriction with and without exercise on metabolic intermediates in nonobese men and women.* J Clin Endocrinol Metab, 2011. **96**(2): p. E312-21.

81. Rehm, H.L., et al., *ClinGen--the Clinical Genome Resource.* N Engl J Med, 2015. **372**(23): p. 2235-42.

82. Riehle, K., et al., *The Genboree Microbiome Toolset and the analysis of 16S rRNA microbial sequences.* BMC Bioinformatics, 2012. **13 Suppl 13**: p. S11.

83. Roadmap Epigenomics, C., et al., *Integrative analysis of 111 reference human epigenomes.* Nature, 2015. **518**(7539): p. 317-30.

84. Robinson, M.D., D.J. McCarthy, and G.K. Smyth, *edgeR: a Bioconductor package for differential expression analysis of digital gene expression data.* Bioinformatics, 2010. **26**(1): p. 139-40.

85. Rozowsky, J., et al., *AlleleSeq: analysis of allele-specific expression and binding in a network framework.* Mol Syst Biol, 2011. **7**: p. 522.

86. Rozowsky, J., et al., *PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls.* Nat Biotechnol, 2009. **27**(1): p. 66-75.

87. Rozowsky, J.S., et al., *The DART classification of unannotated transcription within the ENCODE regions: associating transcription with known and novel loci.* Genome Res, 2007. **17**(6): p. 732-45.

88. Rubin, R., *Despite Potential Health Benefits of Maternity Leave, US Lags Behind Other Industrialized Countries.* JAMA, 2016. **315**(7): p. 643-5.

89. Sboner, A., et al., *FusionSeq: a modular framework for finding gene fusions by analyzing paired-end RNA-sequencing data.* Genome Biol, 2010. **11**(10): p. R104.

90. Sboner, A., et al., *Robust-linear-model normalization to reduce technical variability in functional protein microarrays.* J Proteome Res, 2009. **8**(12): p. 5451-64.

91. Scherer, S.E., et al., *The finished DNA sequence of human chromosome 12.* Nature, 2006. **440**(7082): p. 346-51.

92. Schiaffino, S. and C. Reggiani, *Molecular diversity of myofibrillar proteins: gene regulation and functional significance.* Physiol Rev, 1996. **76**(2): p. 371-423.

93. Science, A.I.f.B., *BrainSpan Atlas of the Developing Human Brain.* (URL), 2015. [**http://brainspan.org**](http://brainspan.org).

94. Shah, A.A., et al., *Metabolic profiles predict adverse events after coronary artery bypass grafting.* J Thorac Cardiovasc Surg, 2012. **143**(4): p. 873-8.

95. Shah, S.H., et al., *Association of a peripheral blood metabolic profile with coronary artery disease and risk of subsequent cardiovascular events.* Circ Cardiovasc Genet, 2010. **3**(2): p. 207-14.

96. Shah, S.H., et al., *Baseline metabolomic profiles predict cardiovascular events in patients at risk for coronary artery disease.* Am Heart J, 2012. **163**(5): p. 844-850 e1.

97. Smith, A., et al., *Leveraging the structure of the Semantic Web to enhance information retrieval for proteomics.* Bioinformatics, 2007. **23**(22): p. 3073-9.

98. Steve Speicher, et al., *Linked Data Platform 1.0 (Working Draft 16 September 2014).* <http://www.w3.org/TR/ldp/>, 2014. **Sep**.

99. Subramanian, S.L., et al., *Integration of extracellular RNA profiling data using metadata, biomedical ontologies and Linked Data technologies.* J Extracell Vesicles, 2015. **4**: p. 27497.

100. Van Dongen, H.P., et al., *Searching for biological rhythms: peak detection in the periodogram of unequally spaced data.* J Biol Rhythms, 1999. **14**(6): p. 617-20.

101. Vidal, M., et al., *The human proteome - a scientific opportunity for transforming diagnostics, therapeutics, and healthcare.* Clin Proteomics, 2012. **9**(1): p. 6.

102. Wang, D., et al., *Loregic: a method to characterize the cooperative logic of regulatory factors.* PLoS Comput Biol, 2015. **11**(4): p. e1004132.

103. Wang, Z., M. Gerstein, and M. Snyder, *RNA-Seq: a revolutionary tool for transcriptomics.* Nat Rev Genet, 2009. **10**(1): p. 57-63.

104. Wu, L., et al., *Global survey of human T leukemic cells by integrating proteomics and transcriptomics profiling.* Mol Cell Proteomics, 2007. **6**(8): p. 1343-53.

105. Yan, K.K., et al., *OrthoClust: an orthology-based network framework for clustering data across multiple species.* Genome Biol, 2014. **15**(8): p. R100.

106. Yip, K.Y., et al., *Classification of human genomic regions based on experimentally determined binding sites of more than 100 transcription-related factors.* Genome Biol, 2012. **13**(9): p. R48.