**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 project will address this knowledge gap by providing a molecular map of physical activity responses and will connect them with health outcomes, thus enabling personalized exercise prescriptions. Based on our extensive prior work, we will be ideally positioned to establish and run a bioinformatics center for MoTrPAc that will interface with other consortium sites; coordinate the data flows and support data analysis; establish and maintain the data resource; develop tools for other investigators to be able to access data and samples for further studies; provide needed input into data analysis design; and conduct preliminary analyses using these data. Specifically, the bioinformatics center will accomplish the following:

**Aim 1 (informatics infrastructure).** Develop an information infrastructure based on Linked Open Data standards and deploy it as a cloud-based service.

1. Provide a database for storage and integration of clinical physiological and multiple types of “omics” data using Linked Open Data standards.
2. Develop a framework for integrating pipelines and tools for analysis and visualization of data and for provenance tracking using the W3C PROV standard.
3. Implement rapid access to accumulated data and tools through the use of cloud-based computing.

**Aim 2 (data processing).** Develop data processing standards, deploy pipelines, and process the data.

1. Oversee standardization of data and metadata across the consortium and develop a consortium-wide data release policy.
2. Develop data processing procedures and pipelines, implement quality control standards.
3. Process data and metadata, generate automated reports, populate the database, provide controlled data sharing via a portal, and submit the data to GEO and dbGaP archives.

**Aim 3** **(data analysis).** We will construct a 'molecular map' of physical activity in humans. We will normalize and register processed multi-omics data between time points and between individuals. We will perform tissue deconvolution to identify the roles of constituent cell-types. We will analyze and cluster gene expression data in acute and durable responses to exercise (for both perturbations and time courses). We will build integrative models using multi-omics data in addition to other available data in order to predict response to physical activity based on molecular biomarkers. We will integrate with genetic information to identify quantitative trait loci associated with response to exercise in order to predict an individual's response.

**Aim 4 (consortium-wide activities).** Participate in consortium-wide activities.

1. Provide expertise in data management and analytics.

Support integration of animal study data and develop plans for replication studies.

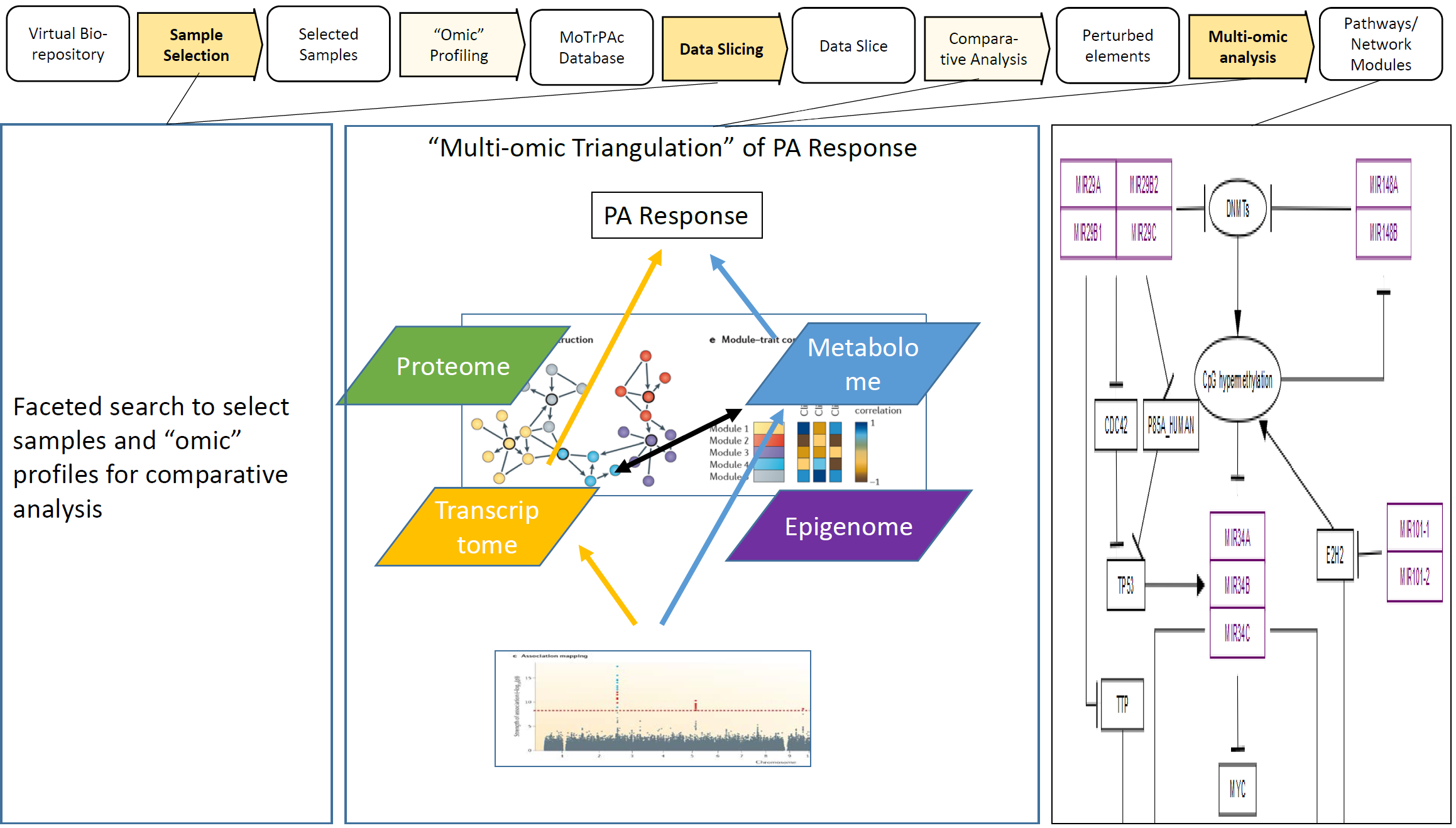
**Research Strategy**

**Significance**

The human organism is adapted for physical activity. The widely prevalent physical inactivity due to the modern way of life is identified as a contributing factor to many unhealthy conditions and chronic diseases (Booth et al., 2000). Exercise is a powerful physiological stimulus that contributes to disease prevention and is used as an effective disease treatment. The protective effects of exercise are well known for many chronic diseases including metabolic, neurodegenerative, cardiovascular diseases and cancer; its use as a therapeutic intervention is common in diabetes and cardiovascular disease. While scientists acknowledge the extensive benefits of exercise, there is still very little understanding about the molecular mechanisms by which exercise prevents disease and improves health across diverse organ systems. The traditional risk factor measurements such as reductions in blood pressure and blood lipids mediate only half of the protective effects of exercise (Joyner and Green, 2009). This presents the following three major challenges for the design of personalized exercise programs that maximize the dose/response specific to the long-term health objectives (Neufer 2013): (1) knowing exactly what to measure, (2) knowing when to make the measurement, and (3) connecting that measurement to a well-defined personal health outcome taking into account genetic and other modulating factors. The MoTrPAc project will address this knowledge gap by providing a molecular map of physical activity physical activity responses and will connect them with health outcomes, thus enabling personalized exercise prescriptions.

To characterize physical activity responses, the MoTrPAc consortium will generate a large volume of highly diverse molecular profiles, assimilate them into a publicly accessible database, and enable their integrative analysis using cloud-hosted analysis and visualization tools. Based on our extensive prior work and record of accomplishment in these areas, we believe that we are ideally positioned to be a bioinformatics center that will establish data flows from the chemical profiling centers; interface with the clinical and animal study sites and the coordination center; establish and maintain the data repository; develop tools for other investigators to be able to access data and samples for further studies; provide needed input into data analysis design; and conduct preliminary analyses using these data.

The construction of the 'molecular map' of physical activity in humans will have to be informed by proper conceptual and methodological frameworks. Toward this goal, the bioinformatics center is tasked with the development of analytical models and theoretical constructs to inform the interpretation of the molecular profiles and the construction of the map. Of particular relevance for this task, are novel integrative analysis methods that we previously developed and that may serve as initial methodological templates for the MoTrPAc project. Most notably, in the STRRIDE study of dose-response effects of exercise on cardiometabolic health, we have identified metabolic signatures in plasma and markers and putative mediators of insulin sensitivity response to exercise (Huffman, *Diabetes Care*, 2009 & 2011). In these papers, metabolic signatures were integrated with peripheral blood inflammatory protein markers and hormones. We have integrated skeletal muscle metabolomics and transcriptomics to identify fatty acid oxidation as important mediators of the differential effects of exercise intensity on exercise-induced responses in cardiorespiratory fitness, insulin action and serum triglyceride levels (Huffman, *Diabetologia*, 2014). We have used peripheral blood and skeletal muscle metabolomics signatures integrated with proteomics to implicate glycine metabolism as an important mediator of improvements in insulin sensitivity in middle-aged overweight and obese individuals undergoing a combined aerobic and resistance exercise program (Glynn et al Diabetologia 2015 PMID: 26254576). We have studied the capacity of peripheral blood metabolomics to predict changes in functional capacity and outcomes in individuals with congestive heart failure undergoing exercise training or left ventricular assist device placement (Ahmad, JACC 2016). Starting with a metabolite signature predicting mortality in individuals with coronary artery disease, we integrated metabolomics, genetics, and epigenetics to implicate the process of endoplasmic reticulum stress in adverse outcomes in this population (Kraus, PLoS Genetics, 2015; PMID: 26540294). These studies resulted in the articulation of the concept of multi-omic “triangulation of physical activity response” [ref] illustrated in **Figure 1**. We anticipate that this methodological template--along with those contributed by other consortium members--will help establish analytical models and theoretical constructs required for the construction of a 'molecular map' of physical activity in humans.



**Innovation**

* Extend and generalize the novel “PA response triangulation” template and develop informatics systems to scale up its application.
* Apply the novel Epigenomic Deconvolution method to deconvolute molecular response of constituent cell types from molecular profiles of complex tissues
* Other analysis innovations…
* Enable development, sharing, and reuse for analysis of visualization of novel network modules and pathways relevant for PA activity within and beyond the consortium using a “PA pathway” section within WikiPathways and Cytoscape.
* Using novel W3C PROVO, LDP 1.0,and other Linked Data standards for metadata we will ensure transparency and reproducibility; implement the novel concept of dynamic metadata to enable collaborative and continuous metadata development; and enable virtual integration of physically distributed data collections.

We will develop novel methodologies for analyzing the all functional genomic data being generating by the consortium to identify both acute and durable molecular responses to PA. We will also build integrate mathematical models to predict response of a subject to PA. We will integrate these results with genetic variation information in order to identify quantitative trait loci (QTLs) associated with response to PA.

**Approach**

**Aim 1 (informatics infrastructure).** **Develop an open interoperable informatic infrastructure and deploy it as a cloud-based service.**

In this aim we focus on two key components of this infrastructure: metadata flow and deployment of pipelines and tools for data analysis and visualization on the context of networks and pathways.

**Aim 1(a) Deploy an open interoperable system and database for managing data and metadata through the use of cloud-based computing.**

**Preliminary studies:**

Milosavljevic’s laboratory has hosted data coordination centers for major projects including the NIH Roadmap Epigenomics project[30] and the Extracellular RNA Communication Consortium[39] and has developed the GenboreeKB system for modeling, storage and processing of metadata and other Linked Data. GenboreeKB is 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 metadata tracking systems that supports data modeling and validation using the new W3C consortium Linked Data standards including RDF and JSON-LD. These standards define a common semantic representation of metadata that can be exposed via Linked Data APIs for indexing and searching. GenboreeKB has deployed and tested in Roadmap Epigenomics and ExRNA projects. Cheung and Gerstein’s laboratory at Yale have pioneered application of Linked Data technologies and ontology-based tool development[27],[12],[18] and have established a record of productive collaboration in metadata modeling with Milosavljevic within the Extracellular RNA Communication Consortium[39]. The established collaborative team and the extensive tested functionality of GenboreeKB will accelerate initial deployment and minimize risk while enabling application of PROVO and related Linked Data standards that provide a uniquely suitable solution for integrating highly heterogeneous, dynamic and distributed metadata that will be generated by the MoTrPAc project.

**Proposed Plan:**

MoTrPAc project will generate highly heterogeneous metadata about various objects, including human subjects, samples, experimental procedures and data analysis steps. Metadata associated with these objects will facilitate multiple steps in the processing and analysis (Figure 1, Panel A), including sample retrieval and profiling; navigation of profiles via faceted search (Figure 1 panel B); and tracking of analyses and visualization in the context of networks and pathways (Figure 1, panel D). We will adopt established data models and formats for genomic, epigenomic, transcriptomic, metabolomic, and proteomic data. In contrast, the metadata will pose significant challenge due to heterogeneity of MoTrPAc data, dynamic metadata revision and curation, and distributed processing. We plan to address those challenges by developing user-friendly interfaces for metadata modeling, entry and curation and by employing W3C standards including PROVO and Linked Data Platform 1.0. This strategy will deliver semantically interoperable ontology-annotated metadata suitable for access via Linked Data APIs for access by applications both within and outside the consortium.

**Addressing data heterogeneity.** To engage a wider number of the 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.

**Dynamic metadata.** 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. We will address this gap by employing the newly developed W3C PROV-O provenance standard to track pre-submission history and post-submission modifications of metadata documents in GenboreeKB, thus implementing the concept of *dynamic* metadata to enable collaborative and continuous metadata development.

As illustrated in **Figure 2**, PROVO support will be a natural extension of the GenboreeKB metadata modeling system because the design of both PROVO and GenboreeKB was guided by W3C recommendations and best practices. Specifically, as indicated in the figure, the document model of GenboreeKB lends itself to PROVO annotations and both the document and its PROVO annotations serialize as RDF/JSON-LD.

**Distributed processing.** We note that it may not be possible or optimal to physically integrate all the MoTrPAc consortium data within a single physical location. One reason for this may be the need for research groups and organizations to maintain physical custody over protected identifiable information about human subjects while sharing only a subset of the data with other consortium members and the public. Another reason is the benefit of utilizing any locally available cluster resources for local data processing. To address this issue, we will be employing Linked Data technologies including REST APIs, JSON-LD, RDF and the emerging Linked Data Platform 1.0 (LDP 1.0) standard. LDP 1.0 features support for integration of physically distributed donor information, thus enabling controlled sharing of personal information and faceted search and single-point (“virtually integrated”) access to physically distributed data collections.

Because the exact requirements of the consortium are not known at this point, we plan on supporting both centralized and distributed modes of operation (using local or commercial cloud or their combination). The configuration will be designed at the time of award in collaboration of consortium members in order to meet consortium requirements and make optimal use of available resources. This flexibility will be facilitated by the fact that both GenboreeKB and Genboree Workbench (discussed below) are deployable as virtual machines on both local and commercial clouds using the Open Virtualization Framework (OVF) standard and that distributed instances can inter-operate (virtually integrate) using Linked Data standards.

**Potential Problems & Alternative Approaches:**

**Aim 1(b) Deploy an open interoperable framework for cloud-based access to pipelines and software tools for analysis and visualization of data in the context of networks and pathways.**

[[describe tool integration via Genboree Workbench and integration of “omic” tools and downstream network and pathway visualization tools. This will be a relatively straightforward extension of our current exRNA work, including pathway/nework visualization work with Alex Pico. Discuss cloud deployment.]]

**Expected Outcomes:**

**Potential Problems & Alternative Approaches:**

**Timetable:**

**Aim 2 (data processing). Develop data and metadata standards, deploy processing pipelines, and process the data.**

**Research Design**:

**(a) Oversee standardization of data and metadata across the consortium and develop a data release policy.**

**Preliminary Results:**

[[Extensive work on metadata standardization on Epigenome Roadmap, International Human Epigenome Consortium, exRNA project by Milosavljevic an Kei-Hoi. Semantic Web work by Kei-Hoi and Gerstein. Work on developing data release policies for consortia.]]

**Proposed Plan:**

To facilitate cross-consortium collaboration, we will form a Metadata/Data Standards Working Group that will serve as a forum for selecting existing or developing new data and metadata standards for profiling data, samples and other resources. Because most profiling technologies have been extensively used, we envision adopting existing data formats and standards. Metadata standards will require significantly more development and ongoing incremental improvement.

The metadata standards will cover data elements describing donors, biosamples, experiments, studies and analysis steps. The metadata will enable efficient selection of samples of interest (e.g. specific health condition of the donor, biofluid or cell/tissue type, or PA intervention type) for integrative analyses and for integration with important external data sets, such as ENCODE. The metadata will help organize the profiling data for efficient interactive access via a faceted search interface at the MoTrPAc Data Portal as well as for programmatic access via REST Application Programming Interfaces (APIs) and Linked Data technologies.

In addition, ontological relationships between concepts will pave the way for knowledge-based data discovery, integration and analysis. Specifically, transitive relations such as ‘‘is-a’’ and ‘‘part-of’’ can be traversed in order to group samples and experiments into more broad categories for the purpose of retrieval and integrative analyses. Also, non-hierarchical relationships (e.g. inhibit, interact and regulate) can be used to implement expressive semantic data queries.

The WG will consider biomedical ontologies available via the BioPortal (1) developed by the National Center for Biomedical Ontology (NCBO), Open Biological and Biomedical Ontology (OBO) Foundry (2), Ontobee (3) and Ontology Lookup Service (4). To link our metadata model to ontologies, we will start with the ontologies adopted by the ENCODE Data Coordination Center or DCC (6,7) as both the MoTrPAc Consortium and ENCODE include “omic” profiling data from human cell and tissue types. These ontologies include Uber Anatomy Ontology (UBERON)(8) for tissues and Foundation Model of Anatomy (FMA)(9) for biofluids, Cell Ontology (CL) (10) for primary cell types and Experimental Factor Ontology (EFO) (11) for immortalized cell lines. While this initial set of ontologies serves as a good starting point, additional ontologies will need to be included since the MoTrPAc Consortium is more clinically focused. Because some PA interventions will include samples from subjects affected by disease (e.g. cancer, Alzheimer’s disease, etc.), additional ontologies such as the Disease Ontology (DOID) (12) will be required to capture the disease terms of interest.

[[more regarding clinical ontologies and ontologies describing PA interventions?]]

As illustrated in Fig. 1a, samples may be selected for profiling based on associated metadata. Ideally, the metadata describe donors, samples,and PA interventions will highly specific terms that are most informative but not suitable for retrieval via query tools. A set of general terms that covers all samples referred to as an ‘‘ontology slim’’ is generally useful to group samples at the top level. For example, the general term ‘‘central nervous system neoplasm’’ may provide a useful grouping of samples, including those that are annotated by highly specific terms such as ‘‘glioblastoma’’ or ‘‘metastatic CNS neoplasm.’’ These general terms are inferred by traversing ontologies from more specific terms up the hierarchy until the ‘‘slim’’ terms are encountered, as illustrated in Fig. 1b. We note that the general terms are useful for grouping and retrieval, while the most specific terms are still available for drilling-down and sub-selection.

[[data release policy paragraph]]

**Potential Problems & Alternative Approaches:**

It is important to emphasize that ontologies are by definition “works in progress” as biological knowledge is extended and may not suffice in their current state for the purpose of the MoTrPAc project. In this case, we will extended them using the collaborative process we established in the context of Roadmap Epigenomics and ExRNA Communication projects. For example, at the outset of the ExRNA project we could not find a widely used ontology that would provide sufficient coverage of the different types of microvesicles and other exRNA-containing entities that will be profiled by the ERC Consortium. To address this gap, we have established a joint project with the Gene Ontology (GO) (24) Consortium to add new exRNA-related terms to GO. We have also been engaging with several other scientific consortia and societies, including the External RNA Controls Consortium, International Society for Extracellular Vesicles and the American Society for Exosomes and Microvesicles, to ensure a broad consensus around the proposed ontology concepts describing exRNA sources. We have subsequently deployed the newly agreed terms to organize search along the “RNA source“ facet at the ExRNA Atlas portal (ExRNA Atlas is accessible via a quick link at exrna.org ).

**(b) Deploy processing pipelines, implement QC metrics, and automate report generation.**

**Preliminary Results:**

**Transcriptome Analysis:** We have extensive experience in the analysis of RNA-Seq data; as part of the ENCODE and modENCODE consortia we helped develop analysis pipelines for processing RNA-Seq data \cite{22955620,25164755}. We have also developed a small RNA-Seq pipeline (exceRpt) as part of the Extracellular RNA Communication Consortium (ERCC) for the analysis of extracellular small RNAs. The exceRpt pipeline quantifies all species of endogenous RNAs including miRNAs, snoRNAs, tRNAs, piRNAs etc. The exceRpt pipeline can also be used to identify exogenous (including bacterial) RNAs that are not due to human sources which might be of non-human origin. We have also developed a number of tools and data formats to handle large quantities of data generated by RNA-Seq experiments \cite{19015660,21134889}. These tools consist of a set of modules that perform common tasks, such as calculating gene and exon expression values, generating signal tracks of mapped reads, and segmenting that signal into actively transcribed regions. These tools employ a special sequence read format we have developed that can dissociate genome sequence information from RNA-Seq signal, maintaining the privacy of test subjects. We have also has extensive experience analyzing and identifying non-coding transcription as well as novel transcribed elements in the genome and assigning these elements putative functions \cite{15539566,21177971,21765801,25164755}. Our Database of Annotated Regions with Tools (DART) package contains tools for identifying unannotated genomic regions that are enriched for transcription, as well as a framework for storing and querying this information \cite{17567993}.

**ChIP-Seq analysis of TFs and chromatin marks:** ChIP-Seq is a mainstream experimental method for genome-wide identification of transcription factor (TF) binding and chromatin modification sites. We developed PeakSeq \cite{19122651}, a versatile tool for identification of TF binding sites and a standard peak calling program used by the ENCODE and modENCODE consortia for ChIP-Seq datasets \cite{19122651}. More recently, we developed MUSIC, a peak caller that performs multiscale decomposition of ChIP signals to enable simultaneous and accurate detection of enrichment at a range of narrow and broad peak breadths \cite{22955619}. This tool is particularly applicable to studies of histone modifications and previously uncharacterized transcription factors, both of which may display both broad and punctate regions of enrichment.

**DNA Methylation analysis:** [[Aleks to write]].

**Proteomics Analysis:** We have experience with the analysis of proteomic data

\cite{17923450,19817483,22583803} and with the integration of proteomics with other genomics data such as the combination of Gas Chromatography Mass Spectrometry (MS) proteomic and transcriptomic data {17519225,25349915}. We have developed a method to exploit the single-nucleotide, unbiased nature of RNA-sequencing \cite{19015660} to construct an ‘expected’ reference proteome to improve the resolution and throughput of MS proteomic analyses.

**Metabolomics Studies:** We have collaborated over the past 10 years, developing integrative physiology studies with metabolomics as a key component; these collaborations have resulted in over 19 manuscripts in metabolic profiling in skeletal muscle and blood related to exercise, caloric restriction, aging and physical function, heart failure and therapies, among others. We have extensive experience in integrating metabolic profiling into multidimensional platforms — genetics, transcriptomics, epigenetics (methylomics), proteomics — with physiologic and clinical outcomes data. These studies have allowed us to identify specific metabolites as biomarkers and biologic mediators of mitochondrial function and intermediary metabolism in disease and physiologic adaptation to exercise in human subjects.

**Proposed Plan:**

**Development and evaluation of uniform data processing methods for different platforms.** To facilitate the integrative analysis of the datasets produced by the consortium, we will develop and evaluate methods for uniform data processing for all common platforms, e.g., ChIP-Seq, RNA-Seq and MS proteomic data. We will evaluate the existing pipelines for the analysis of high throughput functional genomic and proteomic datasets from existing consortia such as the ENCODE Project and the Epigenome Roadmap Project. We will modify and update these pipelines as necessary for the data to be generated by this consortium. We will help implement and deploy these analysis pipelines at the DCC so that they can be run in a streamlined uniform fashion on the all the data. Adopting pipelines similar to other consortiums will facilitate the results to be better integrated and used by members of the community.

We will develop and evaluate different analysis methods for frequently performed analysis tasks, provide sound statistics for selecting among them, and work with the analysis working group (AWG) to ensure uniformity in subsequent processing of each data type using the selected methods. The majority of the consortium data is based on deep sequencing, a technology that presents potential biases and errors not yet completely understood, and thus we discuss plans for assessing the quality of consortium data. We will track metrics to quantify the quality of the data being produced for each for each of the high-throughput data types. Using these metrics we in consultation with members of the consortium we will develop quality control (QC) standards to ensure that the data being generated is of sufficiently high quality to be of general utility to the greater scientific community. For example, we will developed metrics for assessing the performance of ChIP-Seq data by comparing the agreement between genome-wide motif occurrences (based on comparative genomics) and genome-wide peaks of predicted occupancy (based on ChIP-Seq data), and by comparing the reproducibility of called peaks from independent biological replicates.

**Potential Problems & Alternative Approaches:**

If the pipelines we develop take too long to process all the data especially if the data gets submitted by members of the consortium in large boluses we will investigate alternative faster algorithms; we will also try ensure that the consortium data is submitted as a steady stream. If the a significant portion of the data that gets submitted does not meet the agreed upon quality control standards we will investigate the cause to ensure that the standards are in line with equivalent high quality data being generated by the wider genomics community and other consortia.

**(c) Process data and metadata, populate the database, provide controlled data sharing via a portal, and submit the data to GEO and dbGaP archives.**

**Preliminary Results:**

**Proposed Plan:**

As illustrated in Fig. 1a, samples will be selected based on associated metadata for profiling using specific pipelines. Sample profiling includes both experimental assays and subsequent computational steps. The computational pipelines will accept data in FASTQ, BAM, or other sequencing or array formats. Data submission for processing will be accompanied by relevant metadata in JSON (JavaScript Object Notation; www.json.org/) or predefined tabbed value formats. Experimental metadata fields will utilize the Ontology for Biomedical Investigations (OBI) (13) for experimental assays and Chemical Entities of Biological Interest (ChEBI) (14) for chemical treatments. The metadata will be validated against ontologies dynamically using the BioPortal (15) web service. GenboreeKB Metadata Tracking System will provide a user interface for browsing, managing, querying, viewing, uploading and downloading metadata documents.

The exceRpt small RNA-seq pipeline, accessible via the exRNA toolset in the Genboree Workbench (www. genboree.org/), profiles sequencing data using various small RNA databases including miRNAs from miRBase (16), tRNAs from gtRNAdb (17), piRNAs from RNAdb (18) and annotations from Gencode (19). Abundance estimates for each of the genes within these RNA species are computed, as are a variety of quality control metrics such as read-length distributions, summaries of reads mapped to each library and detailed mapping information for each read mapped to each library.

The RNA species and genes are annotated using Sequence Ontology (SO) (20), thus facilitating retrieval of abundance data for genes within different RNA species. Genomic coordinates of RNA genes within each species will define a ‘‘data slice.’’ Such a ‘‘data slice’’ uniquely identifies data based on specific sample and disease ontologies, sequencing assay and experiment ontologies and the RNA species of interest. Abundancy estimates will be pre-computed in the exRNA Atlas, thus making it possible to deliver the ‘‘data slices’’ very fast. While the standardized processing will make the profiles maximally comparable, incompatibilities will naturally exist between different technologies such as qPCR and RNA-seq. This issue will be addressed in part by providing experimental metadata that is sufficient to identify comparable profiles and selection tools in the exRNA Atlas for integrative analysis. While the immediate focus will be on integrating exRNA profiles from human biofluids, the longer term goal includes enabling cross-species analyses.

The data within the exRNA Atlas may be conceptually organized along the following three dimensions illustrated in Fig. 1d: (a) donors and biosamples; (b) assays and experiments and (c) genomic coordinates of RNA genes. A query of the exRNA Atlas should provide a ‘‘data slice’’ in this three-dimensional space that is relevant for downstream analysis. As discussed above, each of the three dimensions is covered by ontologies. Similar to ontology ‘‘slims’’ for the biosample dimension (discussed in the previous section), appropriate ‘‘slims’’ will be de-fined for experimental and genomic dimensions. Ontology traversal will infer general terms along all the three di-mensions (illustrated in Fig. 1bd), thus facilitating retrieval of ‘‘data slices’’ for downstream analysis.

**Contextual interpretation of the results of integrative analyses.**

As illustrated in Fig. 1a, â€˜â€˜data slicesâ€™â€™ are used for downstream integrative analysis. Such analyses produce sets of exRNA genes with relevant profiles. For example, comparative analysis may identify exRNA species that are highly abundant in plasma samples from patients that have a particular type of disease. As another example, an unsupervised clustering may identify a group of exRNA species that show highly correlated abundance levels across a variety of samples. In both examples, the result of analysis is a set of genes, possibly ranked by a significance score.

After identifying a list of candidate exRNAs, the next step often involves relating these candidates to existing knowledge of mechanism and function. The ERC Consortium is pursuing two avenues towards this goal. First, knowledge of exRNA functions is often scattered in unstructured or semi-structured form across many online databases. These databases describe key information like expression patterns, vesicle and body fluid localization, and literature references. We will aggregate and unify these resources within BioGPS under an interface specifically targeted towards exRNAs.

Second, we will integrate external knowledge in the form of pathways and networks into the analysis (illustrated in Fig. 1e). Such knowledge is naturally represented in a machine-readable form using the network-based RDF formalism, as evidenced by databases such as WikiPathways (21) and BioPAX (22) that are now available in RDF form. To facilitate contextual interpretation of exRNA gene sets, the ERC Consortium, therefore, aims to develop a knowledge base of pathways and network modules that include exRNA genes. We have already begun to manually curate pathway information based on Consortium publications into the exRNA col-lection at WikiPathways (www.wikipathways.org/index. php/Portal:ExRNA). Furthermore, programmatic access to ‘‘data slices’’ will allow third-party tools to query and import exRNA gene sets resulting from comparative analysis and clustering. Towards this end, we plan to develop a Cytoscape app (23) to perform exRNA data overlays onto relevant networks and pathways. Within Cytoscape, researchers will also be able to generate novel protein networks based on exRNA gene sets by leveraging existing RNAprotein and protein-protein interaction database information. These network and pathway views of exRNA ‘‘data slices’’ will facilitate interpretation and hypothesis generation by providing independent biological context.

**Submit the data to GEO and dbGaP archives.** We will ensure that relevant public datasets are available in a common repository and in uniform formats to consortium members. We will also ensure that all analysis results by consortium members are shared with the larger community.

**Expected Outcomes:**

**Potential Problems & Alternative Approaches:**

**Timetable:**

**Aim 3 (data analysis). Working with consortium members, establish a 'molecular map' of physical activity in humans.**

In this aim we will describe how the MoTrPAc primary processed results (transcriptomic, chromatin, metabolomic and proteomic data) will be analyzed and integrated with other functional genomic data to build a molecular map in response to physical activity. We will first describe how the data with be normalized and analyzed to identify molecular signatures for response to exercise. We will then further cluster and apply deconvolution approaches in order to separate out the effects of constituent cell-types in complex tissues. We will build integrative models (see Fig. 1 for triangulation approach for multi-omics data) to predict outcome to exercise based on all the available function genomic and other molecular data. Lastly, we will integrate these results with genetic variation in order to identify quantitative trait loci associated with response to PA.

We will use the data generated by the MoTrPAC Consortium will be interested in using the data generated to address several questions about the molecular mechanisms that mediate the health benefits of PA.

* Are the signaling mechanisms the mediate health benefits different by intensity or amount of exercise? Are they different for resistance and aerobic exercise?
* Are they reflected in the responses to acute or to just chronic exercise exposure?
* Are there molecular markers (genetic, epigenetic, metabolomic, proteomic) predictive of the variability of the response to PA and can they be understood in the context of signaling pathways?

The outputs from all these analyses such as the molecular biomarkers for PA, QTLs and other outputs from the computation models will be submitted to the DCC in order to be provided to members of the consortia and the wider scientific community.

**Research Design:**

**(a) Normalization, Deconvolution, and Initial Data Analysis**

**Preliminary Results:**

## In order to perform further analyses on the results from high throughput functional genomic data we have developed a number of methods for the normalization and interpretation of these results. RNA-seq provides much more information than simple quantitation of gene expression levels and that there is reason to believe that alternative RNA splicing or other RNA modifications can also be transducers of PA. To investigate novel transcriptionally active regions further, we developed incRNA, a method that predicts novel ncRNAs using known ncRNAs of various biotypes as a gold standard training set \cite{21177971}. We have also developed specific tools to identify types of transcripts that are difficult to detect using standard analysis pipelines. We developed FusionSeq, a pipeline to detect transcripts that arise due to trans-splicing or chromosomal translocations \cite{20964841,21036922}. We also developed IQSeq, which is a transcript isoform quantification tool which uses a partial sampling framework. It uses an expectation-maximization algorithm to resolve the maximum likelihood expression level of individual transcript isoforms. We also developed Pseudo-seq \cite{22951037}, which addresses the issue of quantification of pseudogene expression, which is difficult to separate from the transcription of parent genes with similar sequences. Pseudoseq solves this problem by calculating the expression in terms of RPKM for pseudogenes by focusing only on those reads and regions that are uniquely mappable \cite{22951037}. We also created the Aggregation and Correlation Toolbox (ACT), which is a general purpose tool for comparing genomic signal tracks \cite{21349863}. We have also participated in the development of a classification and analysis scheme to search for patterns personal omics data with longitudinal profiles based on the presence of “spike” events \cite{22424236}.

Skeletal muscle is a complex tissue comprised of myofibers that are heterogeneous with respect to size, metabolism, and contractile function, with type I being of “slow-twitch” and type II of “fast twitch” type, and with type I and IIa exhibiting high oxidative potential and capillary supply while type IId/x, and type IIb being primarily glycolytic (Pette and Staron 2000; Saltin et al., 1977; Schiaffino and Reggiani 1996). Aerobic training causes both intra-cellular changes such as increase of mitochondria and extensive tissue remodeling including increased vascularization, while strength training increases muscle hypertrophy and satellite cell addition to pre-existing fibers (Howley). To identify both intracellular and tissue composition changes it is highly desirable to deconvolute the “omics” profiles that are obtained on heterogeneous muscle, adipose and blood samples. Toward this end, we developed Epigenomic Deconvolution (EDec), a novel *in silico* deconvolution method that provides estimates of genomic CpG methylation, gene transcription, and other “omic” profiles within a diversity of interacting cell types within complex tissues.

[[,,,,shorten edec shorten by 33%]]

EDec is based on an iterative algorithm for constrained matrix factorization using quadratic programming. The deconvolution is based on cell-type specific patterns of genomic DNA methylation acquired during normal cellular differentiation, maintained during cell division, and serve as a chemically stable marker of cellular identity. Methylation profiles are more amenable to deconvolution than gene expression because of their linearity, measurement within the complete (0-1) dynamic range, and technology-independence (including both bisulfite sequencing and array platforms). By applying EDec to breast tumors from the TCGA collection we detect a striking pattern of cell type specific differential gene expression that reveals metabolic coupling between epithelial and stromal cells within breast tumors that parallels metabolic coupling between the muscle cells that show bias toward glycolysis and those that show both glycolysis and oxidative phosphorylation.

**Proposed Research:**

Even after all the RNA-Seq, ChIP-Seq and other omics data are processed through the primary data analysis pipelines the will be a need to appropriately normalize and register the data between time points and between individuals. Normalization is critical in order to analyze the molecular map in order to identify biomarkers that are differential in response to perturbations associated with PA. Normalization is also important to correctly correlate and cluster genes that should similar activity in response to the PA perturbation regiments. We also plan to evaluate existing tools for differential expression analysis \cite{19910308,25516281} as well as develop new methods if necessary in order to perform differential expression analysis identifying 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 analytic methods to deal with the longitudinal time course multi-omics datasets (beyond simple perturbations) across a group of subjects which a variety of phenotypic and genotypic characteristics. We have experience in normalizing and defining inter-individual registers in longitudinal data. Our proposed approach normalized omics data from several experiments individually, and then accounted for the uneven sampling and time gaps using a Lomb-Scargle periodogram \cite{16303799,22424236}. Each periodogram will then be available for standard time-series analysis and data clustering. Hierarchical clustering was used to obtain common trends and assess biological relevance through such tools as Gene Ontology and through pathway analyses with Reactome and KEGG. This framework can be applied to normalize and compare many different types of ‘omics datasets. Identifying specific effects within massive quantities of longitudinal data requires power and significance testing that takes into account the auto-correlated behavior of the datapoints. We will develop tools that utilize bootstrap simulations to assess levels of autocorrelation for timepoints. These levels can be fed into the periodogram analyses described above, where the number of datapoints can be leveraged to reduce the prediction error at each individual point.

**Tissue deconvolution**

To identify both intracellular and tissue composition changes in response to exercise we will deconvolute the “omics” profiles that are obtained on heterogeneous tissue samples using EDec. EDec will simultaneously predict proportions of multiple cell types within muscle, adipose and blood tissues as well as detect changes in gene expression and other “omic” profiles within each constituent cell type. In contrast to the previous methylation-based deconvolution methods (Accomando et al., 2014; Houseman et al., 2012), EDec does not directly depend on prior knowledge of methylation profiles of constituent cell types that may be unavailable for some of the constituent cell types. [[MG2JR: Edit this down in conjunction w/ EDec sect above]] EDec instead uses lists of loci that exhibit variation in methylation levels across constituent cell types. Such lists are compiled (Figure 1a - Stage 0) from the reference methylomes produced by the NIH Roadmap Epigenomics project (Kundaje et al., 2015) and from a growing multitude of array-based profiles in the NCBI GEO and other public archives. Starting from methylation profiles of tissue homogenates we will apply EDec to estimate both cell type proportions and methylation profiles of constituent cell types (Figure 1a - Stage 1). The proportion estimates are then used as a “key” to deconvolute gene expression and other “omic” profiles of constituent cell types (Figure 1a - Stage 2).

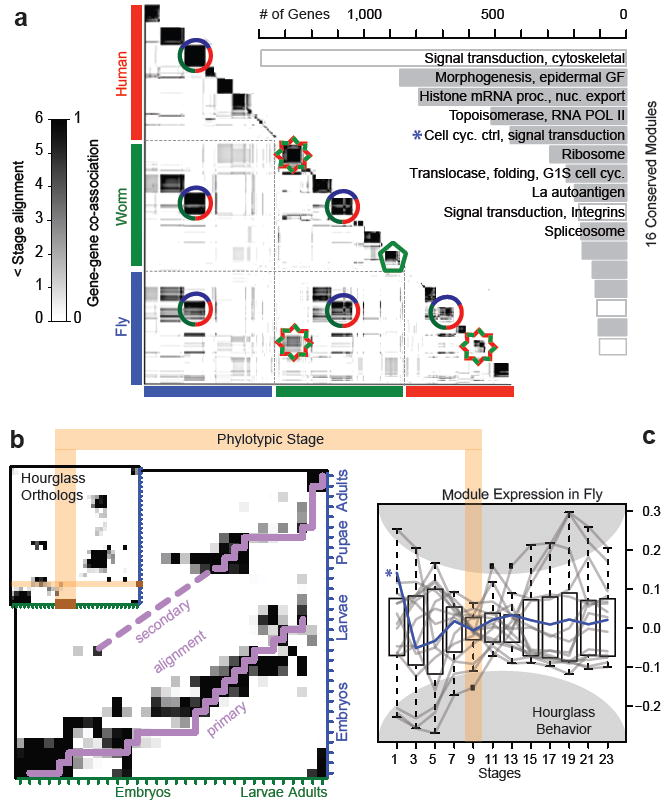
**Potential Problems & Alternative Approaches:**

If some of the time course data from certain subject is missing data from specific time point we will develop methods to predict/impute the missing data using time course profile from other matching individuals. We can also use this to identify potentially mis-labelled datasets as well as sample swaps. If we find we do not resolution in the time course data for certain analyses we will contact members of data processors the possibility of including additional time points for a subsets of the subjects being studied.

## **Aim 3b \*\* Gene expression data analysis**

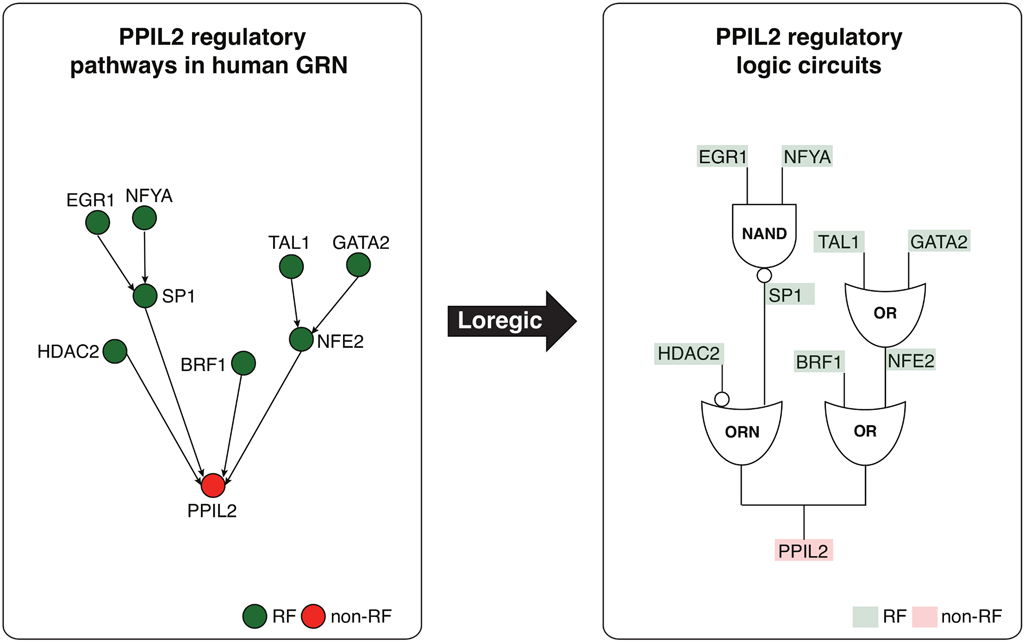
**Preliminary Results:**

[[MG2DW: trim by 15%]] We have extensive experience in analyzing gene expression data such as differential expression analysis and expression clustering. For example, we applied machine learning approaches to the co-expression networks from the prodigious amount of RNA-Seq data generated by ENCODE and modENCODE, and coupled them to orthology relationships of genes between species \cite{25249401}. This multi-layer network framework revealed conserved clusters of connected modules across human, worm and fly that are important for development. Moreover, we developed a method to quantify the differences between the inter-related networks across organisms, and used these metrics to define rates of network change. Applied across species, we used these rates to identify a consistent ordering of rewiring across different network types based on their mechanism of regulation, which thereby elucidated the regulatory mechanisms of various modules. These methods were aggregated into a tool we called OrthoClust, which can be applied to many types of interrelated networks with differing regulatory mechanisms and rates of change, such as for the diverse regulation known to contribute to changes with exercise.



**Figure.** Expression clustering by OrthoClust. Left: Human, worm, and fly gene-gene co-association matrix; darker coloring reflects the increased likelihood that a pair of genes are assigned to the same module. A dark block along the diagonal represents a group of genes within a species. If this is associated with an off-diagonal block then it is a cross-species module (e.g. a three-species conserved module is shown with a circle and a worm-fly module, with a star). However, if a diagonal block has no off-diagonal associations, then it forms a species-specific module (e.g. green pentagon). Right: The GO functional enrichment of genes within the 16 conserved modules is shown.

Gene expression is controlled by various gene regulatory factors. Those factors work cooperatively forming a complex regulatory logical circuit on genome wide. We developed Loregic, a computational method integrating gene expression and regulatory network data, to characterize the cooperativity of regulatory factors. Loregic uses all 16 possible two-input-one-output logic gates (e.g. AND or XOR) to describe triplets of two factors regulating a common target \cite{ PMID: 25884877} and is available as a general-purpose tool (github.com/gersteinlab/loregic). Using human ENCODE ChIP-Seq and TCGA RNA-Seq data, we were able to demonstrate how Loregic characterizes complex circuits involving both proximally and distally regulating transcription factors (TFs) and also miRNAs in human cancer. Besides the regulatory logics, we also developed continuous model-based approaches such as DREISS for dynamics of gene expression driven by external and internal regulatory modules based on state space model to help dissect the temporal dynamic effects of different regulatory subsystems on gene expression \cite{dreiss} [[dW: add ref later: dreiss, https://github.com/gersteinlab/Dreiss, PLoS Computational Biology, in revision).]]



**Figure Depiction of two logic circuit regulatory pathways targeting PPIL2**. Two logic circuit regulatory pathways targeting the PPIL2 gene, an important cyclophilin member in immunological suppression, are found by Loregic: **1:**  PPIL2 is co-regulated by HDAC2 and SP1 forming the triplet of (RF1 is HDAC2, RF2 is SP1, T is PPIL2), which is consistent with the “T=~RF1+RF2” gate, and SP1 is co-regulated by EGR1 and NFYA forming the triplet of (RF1 is EGR1, RF2 is NFYA, T is SP1), which is consistent with the “T=~RF1\*~RF2 (the NAND gate); **2:** PPIL2 is also co-regulated by BRF1 and NFE2 forming the triplet of (RF1 is BRF1, RF2 is NFE2, T is PPIL2), which is consistent with OR gate, and NFE2 is co-regulated by TAL1 and GATA2 forming the triplet of (RF1 is TAL1, RF2 is GATA2, T is NFE2), which is also consistent with OR gate. We replace the triplets on these pathways using matched logic gates, and depict the pathways using logic circuits to summarize the regulatory logics targeting PPIL2 at the pathway level.

In addition, we have associated the exercise phenotypes with the gene expression data. For example, we identified metabolic and genomic molecular signatures in skeletal muscle that mediated the intensity and dose specific effects of exercise training on cardiorespiratory fitness (peak VO2), insulin sensitivity and serum triglycerides \cite{25091629}.

**Proposed Research:**

**[[We need to differentiate betw. perturbation analysis & time-series]]**

In this aim, we would like to identify the greatly differentially expressed genes. Those genes can be considered to be the biomarkers either for acute or durable responses to physical exercise. We will first normalize the gene expression data based on single perturbation experiments. After normalization, we then want to analyze their time-course expression data in longitudinal study, especially for the temporal dynamics in terms of gene expression. We would like to identify the gene expression patterns along with underlying regulatory mechanisms associated with motional phenotypes. In particular, we will first construct the gene co-expression networks where genes are connected when they have correlated expression profiles across time samples. We will then cluster this network into gene co-expression modules and find modular enriched motional phenotypes such as gene expression signatures. Finally, we also want to identify the gene regulatory logics that drive the phenotypes. For instance, we can find the motional biomarker genes associated with the cardiorespiratory fitness, insulin sensitivity and serum triglycerides during exercise training that we recently discovered \cite{25091629}. We can use ENCODE or other publicly available database to construct the gene regulatory networks for those biomarker genes [[MG2DW: as described below, we will integ....]]. We will then apply Loregic to identify the gene regulatory cooperative logics that drive the expression variations of motional biomarker genes across different periods or individual athletes.

## **Aim 3c \*\* Integration with other Consortia Data & Other Integrative Models**

### We plan to integrate the MoTrPAc with external datasets from various consortia to supplement and give perspective via experiment types outside the purview of this consortium. The data integration will focus on the triangulation of multi-’omic data (Fig. 1), linking epigenome modifications to acute changes in transcript, metabolite and protein levels as they relate to a epigenome changes, tissue modification and other durable responses. This work will build on our extensive experience building mathematical models to relate these data. The integrated data will feed into a model to define the molecular map of physical activity in humans, which will be built on our past experience building such integrative models.

**Preliminary Results:**

We also have extensive experience integrating transcriptomic, metabolomic and proteomic data in the context of exercise as well as other settings. For example, we have 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 (Huffman, Diabetes Care, 2009 & 2011; Huffman, Diabetologia, 2014). We also linked unannotated metabolites, which can constitute as many as 50% of the observed spectral features in metabolomic datasets \cite{22424236}, to transcriptomics. We used compound co-occurrence profiles to couple Gas Chromatography Mass Spectrometry (MS) data to the gene expression profiles from several experimental conditions \cite{22396667}. By defining statistics to correlate the co-occurrence patterns of volatile metabolites and genes we were able to generate hypotheses for the identity of unannotated biosynthetic pathways.

We have also developed methods to integrate metabolomics data to proteomic experiments. We have used metabolomics and proteomics from blood and skeletal muscle to implicate glycine metabolism a regulator of insulin sensitivity in individuals enrolled in exercise programs \cite{26254576, Ahmad, JACC 2016}. As with the metabolomics/transcriptomics analysis, here too we have experience using these data to expand the existing knowledgebase. We developed a novel Liquid Chromatography MS assay to identify in vivo protein-hydrophobic small metabolite interactions in yeast, which allowed us to identify many novel interactions, including key regulatory proteins such as the protein kinase Ypk1 \cite{21035178}.

We extensive experience integrating datasets to form models of various types. For example, we integrated ENCODE data from transcription factor binding, histone modifications and target gene expression to establish the regulatory relationships using a probabilistic model-based method, TIP (Target Identification from Profiles) \cite{22039215}. We applied machine-learning methods to integrate multiple genomic features in order to classify regulatory regions for >100 transcription factors. In particular, we were able to identify potential enhancers from regions classified as gene-distal regulatory modules \cite{22950945}. We used these modules to quantify the relationship between TF binding and gene expression using linear and non-linear models that take the binding signals of multiple TFs in the transcription start site (TSS) proximal to genes as the input to predict protein-coding and non-coding gene expression levels as the output \cite{22955978, 22955616, 21926158}. Similarly, we constructed models to predict gene expression levels based on histone modification signals at different positions proximal to the TSS \cite{22950368, 21324173, 21177976, 22950368}. Strikingly, the models trained solely on protein-coding genes also predict the expression levels of non-coding genes, suggesting a common regulatory mechanism is shared between them, and this trend is conserved across animals.

In addition, we have related these integrated data and regulation models to changes in epigenetic profiles. For example, we demonstrated the ability to predict histone modification signals from protein-protein interaction networks and position in the hierarchy of transcriptional regulation. We grouped transcription factors into histone-sensitive and -insensitive classes that refined our ability to predict the targets of transcriptional regulation \cite{22060676}. We will use these tools to determine these parameters for the MoTrPAc data to refine our models for the predicted effects of exercise.

**Proposed Research:**

[[To build these models the first step is to integrate these data...

,,,,,to enable these integ.]] We will integrate the MoTrPAc multi-’omic triangulation of PA with phenotypic data and large datasets from other consortia. In particular, we will incorporate the ENCODE regulation data, the GTEx tissue-specific profiles and the Epigenome RoadMap effects of various epigenetic marks on transcriptional responses. We will use these data to determine how acute changes in transcript, metabolite and protein levels are related to a epigenome changes and durable responses such as tissue modification.

First we will use the logical clustering models generated in Aim 3a that organized gene expression signatures by tissue, organ and cell-specific states. We will apply the ENCODE regulatory data to model these clusters as interrelated networks. This network framework will be the multi to create ‘molecular map’ of physical activity in humans, which will model the acute responses the exercise. These data will be integrated with proteomic and metabolomic data to form the multi-'omic triangulation of PA.

Second, we will incorporate the epigenetic MoTrPAc and Epigenome Roadmap and GTeX data to model the durable responses the exercise. We will link the methylation data generated by this consortium to the extensive ChIP-Seq data generated by others to predict the changes in chromatin modification and the deconvoluted expression signals from specific tissues from Aim 3b, thereby relating short and long term effects of exercise.

## **Aim 3d \*\* Relating to Genetic Variation (eQTL analysis)**

We will relate the multi- ‘omic and integrative datasets to the changes within individuals by identifying Quantitative Trait Loci (QTLs) incorporating an individual’s genetic variation and allele information. By tailoring the ‘molecular map’ model using an individual’s genetic profile, we can gain a deeper understanding of the genetic predisposition effects of, and variation in response to, exercise.

**Preliminary Results [[ref privaseq as expertise in allelic analysis and eQTLs, in particular the alleleseq]]**

A major area of interest in RNA-Seq analysis is linking expression variation to genotype. We have expertise in this subject in the form of allelic analysis [[MG2DS: ,,,,n qtl priva]]. In particular, our AlleleSeq tool \cite{21811232} combines diploid genomic information with RNA-Seq data to identify transcripts showing allele specific expression.

These methods couple well to our extensive experience in using a network framework to integrate data from rare and somatic variants. We developed a method that creates a unified biological network of gene interactions (regulatory, genetic, phosphorylation, signaling, metabolic and physical protein-protein interactions), and then layers onto that the SNP and variant profiles from an individual. By identifying and exploiting the unique properties of loss-of-function tolerant and essential genes we built a model to predict global perturbations caused by deleterious mutations. These methods may be useful to predict the metabolic perturbations within individuals that may explain responses to interventions such as exercise. In the context of PA, we identified genomic predictors that affected the response to exercise training on cardiovascular heath \cite{21183627}, and later on those individuals that responded poorly \cite{22666405}. In addition, we integrated metabolomics and epigenetic markers with genetic variants to define metabolomic Quantitative Trait Loci (mQTLs) associated with poor coronary artery disease outcomes /cite{26540294}.

In addition, we have experience integrating data to order genomic variants in terms of their likelihood of causing significant phenotypic effects, which can be useful for reducing the dimensionality of diverse datasets. Benchmarked in tumor sequencing, this tool called Function based Prioritization of Sequence Variants or FunSeq, builds information context for variants with particular attention to non-coding somatic mutations \cite{25273974}. Importantly, the data context can be adapted to user-defined datasets, allowing for customization to the goals of the motrpac.

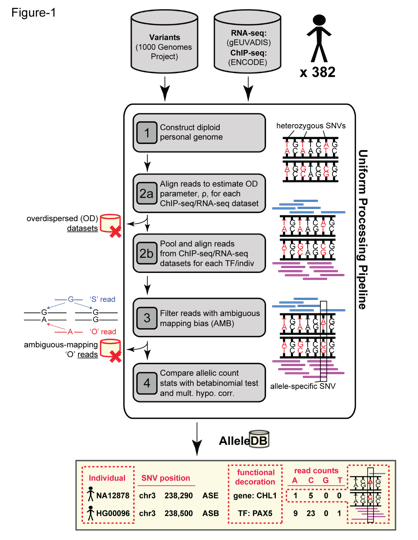
**Proposed research:**

We plan on integrating the results of analyzing the functional genomic data being generated with the genetic information from the same individuals. For example we will compare the genetic variants (such as SNPs) with the results of differential expression analysis in response to physical activity. This will allow us to identify expression quantitative trait loci (eQTLs), i.e. loci where variants show statistically significant association with the expression changes in response to exercise. The variants corresponding to eQTLs can be used to stratify the participants in these study to classify outcome in response to a physical exercise regime. This type of analysis can also be done for other types of genomics variants such as indels and structural variants (SVs) if such data is available from either from exome or whole genome sequencing for these subjects.

The strategy of identifying eQTLs can be extended to identify QTLs in response to other types of omic data (chromatin, metabolic, proteomic) being assayed. We can identify QTLs corresponding to variants that are associated with metabolic changes (i.e. mQTLs), proteomic changes (i.e. pQTLs) and chromatin or histone modification changes (i.e. hmQTLs). This can be done for differential chromatin, metabolic and proteomic changes with either acute or durable response to exercise. We will compare these different QTLs at the target gene level to identify target gene that share different types of QTLs. We will build models to cluster and classify subjects based on their response to physical activity using the various QTLs available. These may provide a more direct path to functional responses or diagnostic tests.

[[2DSnJR]] goal here is to look at all these ppl and diff rnaseq/methylation and understand how these diff relates to genetic variants. Allelic analysis is the most direct way of doing this, particularly for rare variants]]

We also plan to perform an allelic analysis - identifying genes that exhibit the phenomena of allele-specific expression or regulatory regions that show allele-specific binding using histone modification ChIP-Seq data. 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 the haplotypes in order to identify those that show an imbalance between the maternal and paternal alleles,,,,,y. We will use these allelic sites to differentiate between subjects based on their response to exercise.



**Figure.** AlleleDB workflow for determining allele-specific variants.

**Expected Outcomes:**

**Potential Problems & Alternative Approaches:**

**Timetable:**

**Aim 4. Consortium Activities: Enable Integration of Biospecimen Data and Samples across the Consortium.**

**(a) Enable integration of biospecimen data and other resources across the consortium.**

**Assumptions.** Two products of the MoTrPAC Consortium will be a publically available data resource to be used to enhance and accelerate subsequent mechanistic research on diseases and conditions affected by physical activity; and a biorepository of clinical and animal model samples to be used in furthering exercise science well into the future. We are assuming the that MoTrPAC Coordinating Center will arrange for biological samples (blood products, urine, skeletal muscle, and adipose) generated at the Clinical Sites to be stored in a central biorepository designated by the Consortium. The Bioinformatics Center will work directly with this central biorepository to tracking sample volumes and numbers; maintain a direct link between the clinical data and the biospecimens; and provide an interface to the biorepository to directly access these data. Decisions about process will be made collectively among the MoTrPAC Consortium members. If the Consortium decides to handle these issues in a different fashion, we will modify our plans accordingly. We are assuming that all clinical and physiologic data that are generated at the clinical sites will integrated into a central data repository using electronic records and files; direct web page access to database entry (á la RedCap); or through paper generated case report forms. We are assuming that these items will be handled by the Coordinating Center as part of their project grant. The Bioinformatics Center will then integrate the clinical/physiologic data, the biorepository data, and the biochemical analytic sites’ data, to create the MoTrPAC Database. Study investigators will then access this database using the tools that the Bioinformatics Center will generate. We are assuming that this resource will outlive the life of the Bioinformatics Center’s initial award, and that this resource to be used by investigators in the future will be supported through a new funding mechanism established at the conclusion of the MoTrPAC project.

**Preliminary Studies.** The Bioinformatics Center PIs have a broad collective experience working in intensive multi-site collaborative projects with clinical/physiologic data and multiplatform “–omics” data collected as part of the project; this provides us with extensive experience with the needs of the MoTrPAC project. Dr. Kraus has served in leadership position for the CALERIE (NCT00427193), STRRIDE (NCT00200993; NCT00275145get clinicaltrials.gov designation), and HF-ACTION (NCT00047437) projects. Each of these projects developed a biorepository of human samples held in a central location; tracked biological sample volumes and number, maintained linkage to the study data; and maintained access to data and samples by collaborating investigators. For these, the informatics platform PEDIGENE® — held at Duke University — served as the integrating information technology resource. Drs. Gerstein and Milosavljevic have similarly worked together on collaborative projects of similar scope, the most recent of which was ExRNA, a project designed to ….. The data sharing component of this project was operationalized through the ExTNA Virtual Biorepository, implemented using Linked Data standards and GenboreekKB. This latter platform will be used in this project to track samples and data and to provide access to information through a web-based interface.

**(b) Provide expertise in data management and analytics.**

We have extensive experience conducting integrated analyses, developing tools and databases to mine genomic data in humans and other organisms in the context of large-scale collaborations, including the ENCODE \cite{22955616} and modENCODE \cite{19536255} consortia, the 1000 Genomes Project \cite{20981092,26432245}, the BrainSpan ([http://www.BrainSpan.org](http://www.brainspan.org)) and PsychENCODE \cite{26605881} projects, Extracellular RNA Communication program (<http://commonfund.nih.gov/Exrna/>), and the DOE Knowledgebase, Kbase (http://kbase.us/) We have also conducted extensive studies of the relationship between ChIP-Seq data for localization of transcription factors and histone modifications and gene expression through RNA-Seq \cite{22955619,22955978}. Currently, we are active participants of the BrainSpan project as well as of the PsychENCODE project and we lead the data integration and analysis component of the data management and resource repository for the NIH ExRNA Communication program. We are also active leaders of PCAWG (Pan Cancer Analysis of Whole Genomes).

We have participated in a number of research activities that link genetics to exercise, while not publishing extensively on the genomics of exercise physiology. For example, we recently looked at the genomics of extreme individuals. We have published a number of preliminary papers highlighting the most prominent genes that influence human athletic ability \cite{22762737, SF Chronicle, May 2, Page E-4}, and advocated the linking of genetic factors and physical phenotypes in relation to high performance athletes \cite{DOI: 10.1126/science.1245795}. In addition, we published an opinion about how the genomic era affect sports industry \cite{Wall Street Journal, July 24, Page A13}. Actually, advocating sports genetics is also one of the most positive and easy ways for teaching genetics, especially in the classroom study. It is free of many problematical considerations; e.g., rare disease genetics potentially raises privacy concerns in a negative context.

[[Some of our experience in providing analysis support to Epigenome Ropadmap and ExRNA consortia via workshops, on-line materials, an done-on-one collaborations should go here.]]

**Potential Approaches.**  One such approach that can be taken is as follows. Imagine wanting to describe a biological molecular pathway from predictive gene to exercise-induced change in a physiologic health parameter: for instance, cardiorespiratory fitness or insulin sensitivity. If one were to work forward from gene to outcome through gene expression to protein to metabolites, one would confront an unmanageable number of statistical comparisons (20,000 genes, 67 metabolites on 10 potential outcomes = ~1013 tests. (can insert figure) Rather than working forward, one might move in the inverse manner, from outcome to metabolites that predict change in the physiologic health parameter to proteins correlated with the metabolites, and so forth, to predictive genes. By selecting candidates of interest progressively back from outcome through metabolite principal components (factor analysis) and gene expression to gene and genetic modification, one substantially reduces the number of comparisons and the statistical tractability of the problem. A similar approach can be taken for signatures in multiple tissue components (blood, skeletal muscle, and adipose) to describe a systemic molecular network describing the molecular physiology of exercise responses.

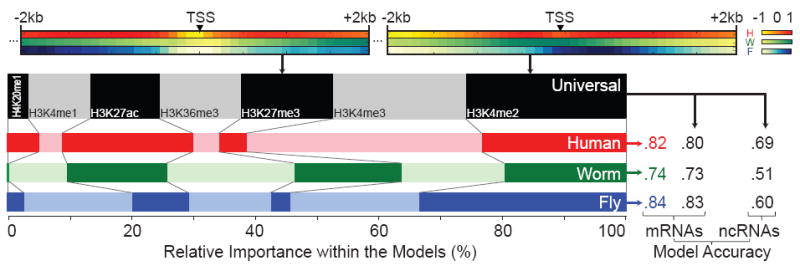
**(c) Support integration of animal study data and develop plans for replication studies.**

**Assumptions.**  An important component of the MoTrPAC project are the Preclinical Animal Study Sites (PASS). The strategy is for the PASS sites to spend the first two years of their projects developing animal models that replicate the ongoing studies being undertaken in humans at the Clinical Sites. The PASS sites will study tissues in animal models unavailable to human studies (e.g., liver, pancreas, brain). Molecular signatures will be developed on these issues; physiologic data; molecular data and sample repositories will be created as for the human studies. In the second phase (3 years) of the project, the PASS sites will design and conduct replication mechanistic studies in animal models to further explore the findings in the human work.

We have great experience inter-relating model organism to humans. We played an important role in modencode. In particular, we played an important role in the analysis of human and model organisms transcriptomes \cite{21177976,25164755,22955620} within the mod/ENCODE consortium. We led the comparison of the worm and fly to the human \cite{}(add ref).

[DW to move some stuff here from aim 3]][[DW: done. See below]]

[[also in 3c (need to modify): we constructed models to predict gene expression levels based on histone modification signals at different positions proximal to the TSS \cite{22950368, 21324173, 21177976, 22950368}.]]



**Figure.** Histone Models for Gene Expression. Top: Normalized correlations of two representative histone marks with expression. Left: Relative importance of the histone marks in organism-specific models and the universal model. Right: Prediction accuracies (Pearson correlations all significant, p<1e-100) of the organism-specific and universal models.

We constructed statistical models to quantify the effects of transcription factor binding and histone modification in mouse embryonic stem cells. We found spatial differences between these regulatory types: TFs were predictive very locally and HMs across a wider region \cite{21926158}. We also developed a novel clustering method, OrthoClust to cluster across-species gene networks \cite{25249401}. We applied OrthoClust to simultaneously cluster model organisms and human gene co-expression networks, and discovered novel human genomic functions.

**Approach.**  We will create a data handling resource for the PASS studies as we have proposed to do (Aim 4.a) for the human Clinical Sites. Having the animal data in the same platform as the human data with the same metadata elements will facilitate integration of the PASS and human MoTrPAC data. We will work with the Consortium at common meetings to develop analytic strategies that will promote replication studies, as described.

**Expected Outcomes:**

**Potential Problems & Alternative Approaches:**

**Timetable:**

**Management, multi-PI sections, Data Management, other sections**

**LEFTOVERS**

By challenging homeostasis, PA causes both transient responses upon each successive exercise bout as well as adaptive responses to long-term exercise training. PA responses involve networks that span multiple levels of organization, from intra-cellular to tissue-level to systemic. At the intra-cellular level, endurance-based exercise training elicits increase the mitochondrial protein content and respiratory capacity of the trained myofibers that result in less lactate production at a given submaximal power output or speed (Holloszy 1967). At the tissue level, exercise elicits extensive tissue remodeling, including increased muscle mass and vascularization. Finally, systemic responses include cardiovascular, respiratory, and metabolic changes that increase fuel and oxygen supply of the contracting muscles to sustain higher level of activity. The responses at all three levels are mediated by layers of molecular transducers that may be measured by epigenomic, transcriptomic, proteomic, and metabolomic profiling and are modulated by genetic and other factors such as age, gender, physical fitness, drugs, and nutrition.

**Preliminary Study:** ExRNA Virtual Biorepository implemented using Linked Data standards and GenboreeKB.

based on W3C Linked Open Data standards:

Metadata standards

Ids

Data elems

Formats

Apis

**Plan:**

A number of resources will be shared across the consortium. For example, the central biorepository will distribute sample aliquots to profiling centers; the profiling centers will then submit the profiles to the data coordination center. There will be a need to track the biosample aliquots and resulting profiles across the consortium. Cross-consortium collaborations may also benefit from tracking of other shared resources.

Enable consortium members to share information about biospecimen aliquots and resulting molecular profiles using GenboreeKB and Linked Data Technologies. The tracking system will be generic and will enable sharing of other resources across the consortium, thus providing an essential informatic backbone for collaboration within the consortium and a nucleation point for broadening collaboration beyond the consortium.

**PIs and Key Personnel**

**Gerstein:** Develop data processing pipelines. Define data Quality Control metrics. Perform knowledge integration. Participate in cloud computing implementation. Develop models for understanding the multidimensional, multimodality data, and developing high quality connectivity maps showing the interrelations between the various data types. Design and implement integrative analysis strategies.

**Kraus:**

1. Participate in metadata standards development (RFA: “coordinate implementation of data and ontology-based metadata standards”; “Selecting or developing common data elements to enable uniform aggregation of data”) 0.2 FTE
2. Perform metadata “wrangling” and curation (RFA: “Developing, or using pre-existing and well-defined, data curation standards and methods”; “Accumulating, integrating, and storing as necessary physiological, metabolic, and metadata from the Clinical Centers and PASS, and metabolomic, proteomic, genomic, and transcriptomic data from the Chemical Analysis Sites”) 1.5 FTE
3. Act as liaison with clinical studies – and co-develop plans for replication and validation studies (RFA: “Working with the Steering Committee to develop plans for replication/validation studies as needed.”) 0.1FTE
4. Participate in the integration of biospecimen data 0.1 FTE
5. Co-develop data portal, 0.1 FTE
6. Co- develop models for understanding the multidimensional, multimodality data, and participate in developing connectivity maps showing the interrelations between the various data types 0.6 FTE
7. Participate in preliminary data analysis (RFA: “and conduct preliminary data analysis of the diverse datasets submitted by other MoTrPAC elements.”) 0.4 FTE

**Milosavljevic:** Co-develop data and metadata standards. Develop a “virtual biorepository” to integrate biospecimen data across the consortium. Implement the infrastructure for metadata and data processing. Deploy data processing pipelines. Perform data processing. Enable data sharing. Implement provenance tracking. Perform data deposition into public archives. Develop on-line supporting material sand implement user training. Develop methods and tools to facilitate and participate in integrative analysis in the context of background knowledge of pathways and networks.

**Other Key Personnel:**

Alex Pico (UCSF, networks, pathways, visualization)

Sai Lakshmi Subramanian (BCM, data processing)

Kei-Hoi Cheung (Yale, ontologies, Linked Data)

Hongyu Zhao (Yale, statistics)

Kim Huffman (Duke, metabolomics, exercise, physiology, metadata gene expression)

Svati Shah (Duke, metabolomics, exercise, QC, metadata)