Specific Aims

Working with consortium members, we will develop a 'molecular map' of physical activity in humans. We will implement an informatics infrastructure to support consortium's effort to develop a public data resource that any researcher can access to develop hypotheses regarding the molecular mechanisms through which physical activity can improve or preserve health. We will validate the resource by conducting preliminary data analysis of both acute and adaptive response to exercise using the diverse datasets submitted by the consortium members.

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

(a) Provide a database for storage and integration of clinical physiological and multiple types of "omics" data using Linked Open Data standards.
(b) Develop a framework for integrating pipelines and tools for analysis and visualization of data and for provenance tracking using the W3C PROV standard.
(c) 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.

(a) Oversee standardization of data and metadata across the consortium and develop a consortium-wide data release policy.
(b) Develop data processing SOPs and pipelines, implement QC metrics and automated report generation.
(c) Process data and metadata, populate the database, provide controlled data sharing via a portal, and submit the data to GEO and dbGaP archives.

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

(a) Working with consortium members, develop analytical models and theoretical constructs for a 'molecular map' of physical activity in humans.
(b) Develop methods and deploy the tools for analysis and visualization of molecular responses to physical activity in the context of pathways and networks.
(c) Conduct preliminary data analysis of (i) acute and (ii) durable responses to exercise using the diverse datasets submitted by the consortium members..

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

(a) Track biospecimen processing and provide an integrated view of other resources across the consortium.
(b) Provide expertise in data management and analytics.
(c) Support integration of animal study data and develop plans for replication studies.

Research Strategy

Significance

Scientific Premise:

<Literature and Preliminary Studies in Support of Aims>

Significance of the Expected Research Contribution:

Innovation

Approach

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

<introductory paragraph>

Research Design:

(a) Provide a database for storage and integration of clinical physiological and multiple types of "omics" data using Linked Open Data standards.

(b) Develop a framework for integrating pipelines and tools for analysis and visualization of data and for provenance tracking using the W3C PROV standard.

(c) Implement rapid access to accumulated data and tools through the use of cloud-based computing.

Expected Outcomes:

Potential Problems & Alternative Approaches:

Timetable:

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

<introductory paragraph>

Research Design:

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

(b) Develop data processing SOPs and pipelines, implement QC metrics and automated report generation.

Preliminary Results:

Transcriptome Analysis: The Gerstein lab has developed a number of tools and data formats to handle large quantities of data generated by RNA-Seq experiments. We have developed a suite of tools (RSEQtools) that uses this format for the analysis of RNA-Seq experiments \cite{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. Moreover, the components of RSEQtools can readily be assembled and extended to build customizable RNA-Seq workflows. This tool employs a special sequence read format we have developed that can dissociate genome sequence information from RNA-Seq signal, maintaining the privacy of test subjects. 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}. We have also developed the exceRpt pipeline for the analysis of small RNA-Seq data quantifying 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 metagenomic origin.

ChIP-Seq and Epigenomics Analysis: 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.

Proteomics Analysis: We have developed a novel method (that we are proposing to further develop) to exploit the single-nucleotide, unbiased nature of RNA-sequencing to construct an 'expected' reference proteome for the subject and tissue of interest in an effort to improve the resolution and throughput of LC-MS/MS proteomic analyses. RNA-seq can potentially predict the entire proteome without requiring any a priori information regarding the transcriptome sequence \cite{19015660}, and we have already shown that this approach is capable of predicting the proteome with very high resolution. We have shown through preliminary investigations that the number of identified proteins can be greatly increased (by up to 150% compared to SWISS-Prot) through the use of a tissue-specific reference proteome derived directly from an RNA-seq experiment. We have also found that the coverage of the detected proteins, in terms of the number of distinct peptides observed for each, is consistently greater when using a specific proteome reference compared to a generic reference such as SWISS-Prot, leading to more accurate protein-quantitation and increased likelihood of validating variants, allele-specific expression, or RNA-editing at the whole-protein level.

Metabolomics
[Bill to fill in]

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. We will continue to evaluate possible sources of bias in existing methods, and continue to test new methods being developed that improve upon accuracy and speed of analysis.

(c) Process data and metadata, populate the database, provide controlled data sharing via a portal, and 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.

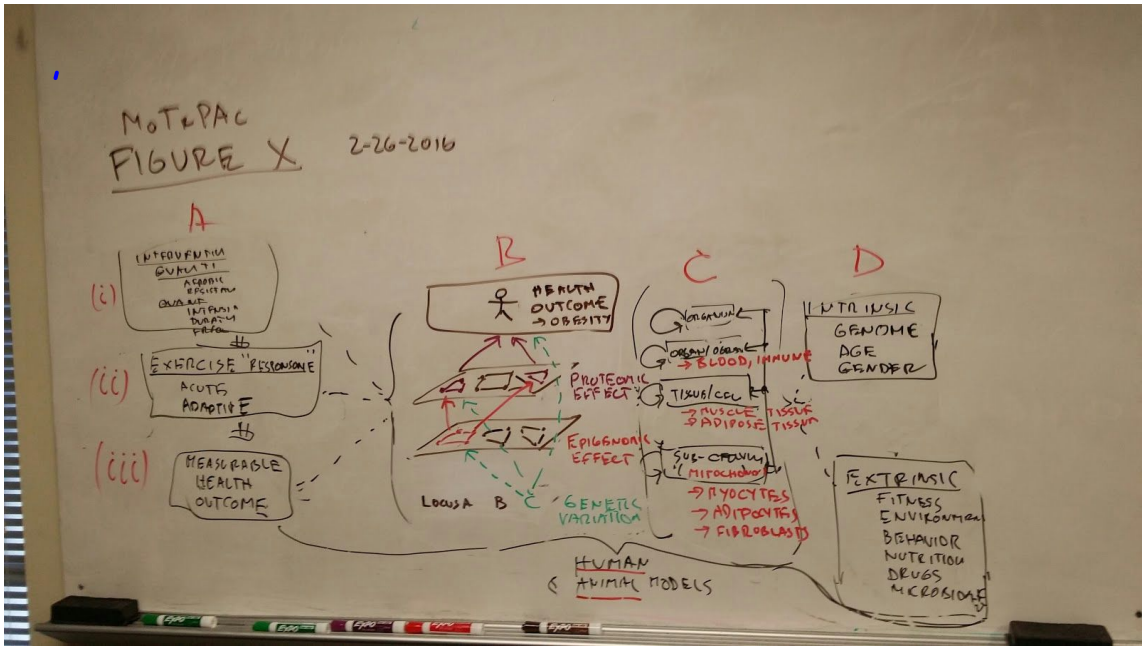
Expected Outcomes:

Potential Problems & Alternative Approaches:

Timetable:

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

This review paper: **PMID:24296534** provides some structure that may help organize this aim.



[[
Rewrite aim 3
JR .75 pg on normalization & initial clustering & proc of RNA data
[orthoclust, chip normalization, cmptxn, incRNA,]] clustering modules, quantitation and normalization

DW 1pg on deconvolution & dynamic models , driess & loregic
DS 1 pg on integrating w other consortia & integrative models relating genome to epigenme [[CC'stuff]]
DS .5 pg on relating genetic variation (QTLs)
]]

<introductory paragraph>

Significance of the Expected Research Contribution:

Literature and Preliminary Studies in Support of Aims

Research Design:

Aim 3a ** Normalization and Initial Downstream Analysis of Consortium Data

JR .75 pg on normalization & initial clustering & proc of RNA data
[orthoclust, chip normalization, cmptxn, incRNA,]] clustering modules, quantitation and normalization

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. 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, 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 genome 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}.

Proposed Research:

One of the main analysis problems will be to develop analytic methods to deal with the longitudinal time series omics datasets across a group of subjects which a variety of phenotypic and genotypic characteristics. Even after 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 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 physical activity. Normalization is also important to correctly correlate and cluster genes that should similar activity in response to the physical activity regiment.

The Gerstein lab has 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. Each periodogram was then 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.

Aim 3b ** Deconvolution and Dynamic Models

DW 1pg on deconvolution & dynamic models , driess & loregic

Preliminary Results:
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 (https://github.com/gersteinlab/Dreiss, PLoS Computational Biology, in revision).

In addition, TFs and histone modifications are two interrelated components that regulate the transcriptional output of a gene. To quantify the relationship between TF binding and gene expression, we have constructed 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 have also constructed models to predict gene expression levels based on histone modification signals at different positions proximal to the TSS of different genes \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.

Proposed Research: DW 1/2 + developing approaches for tissue de-convolution**

In this aim, we want to identify cell type signatures in terms of gene expression from athletes tissue samples such as muscle, and find the motional biomarker genes from the signatures that can most discriminate the benefits of exercise. We assume that the mixed effects from various related cell types determine the gene expression from each athlete. We will apply both linear

and nonlinear approaches to capture the mixed effects as follows. Moreover, we want to identify gene regulatory mechanisms such as regulatory logics that drive these signatures.

We first try the linear models that will be computationally efficient. Given the gene expression levels and cell type fractions of the muscle sample for each athlete, we can use a linear matrix model to identify cell types gene expression signatures. For instance, the athlete's *i*th gene expression level can be modeled as a linear superposition of the same gene's expression levels of multiple cell type signatures; i.e., the *i*th gene expression level of *k*th individual person, *x(i,k)* is the linear combination of this gene's expression levels of different cell type signatures; i.e., , where *s(i,j)* is the *i*th gene's expression level in the *j*th cell type, and *w(i,k)* is the contributing weight of *j*th cell type to *k*th person, which can be the *j*th cell type fraction of *k*th person. If we rewrite this linear model in a matrix form, we have that $X=SW$, where X is the gene expression matrix whose the rows and columns represent genes and persons, W is the cell type fraction matrix whose rows and columns represent cell types and persons, and S is the cell type signature matrix whose the rows and columns represent genes and cell types. The cell type fractions in W can be provided by clinicians. The RNA-seq data will provide matrix X, so we need to find the optimal S to minimize $||X-SW||_F$ given X and W. The optimal solution $S=XW^*$, where W^* is pseudo inverse of W s.t., $WW^*=I$ identity matrix.

We then try to apply advanced models to capture nonlinear effects from different cells to gene expression. For example, we can use the Denoising Autoencoder (DA), an unsupervised machine-learning framework to extract and characterize cell type signatures. DA is able to discover non-linear expression features from gene expression data using sigmoid transformation. We will apply DA to different patients clusters and compare their non-linear features, and find the genes that have features to most discriminate different groups of athletes.

Finally, we also want to identify the gene regulatory logics that drive the cell-type signatures, especially for those motional biomarker genes. We will first construct the gene regulatory networks for the cell types by integrating their ChIP-seq data in ENCODE. 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

DS 1 pg on integrating w other consortia & integrative models relating genome to epigenome [[CC'stuff]]

We plan to integrate the motrpac data with other datasets from various consortia both to help give perspective and to supplement the motrpac with experiment types outside the purview of this consortium. The data integration will focus on relating multi-'omic to epigenome modification to determine how acute changes in transcription and metabolite and protein levels are related 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 of statistically rigorous network modeling.

Preliminary Results:

We have extensive experience in performing integrative analyses both within and between large consortia, including ENCODE, modENCODE, 1000 Genomes, KBase and Brainspan. For example, we constructed highly-integrated regulatory networks for humans and model organisms using the ENCODE and modENCODE datasets. We applied machine learning approaches to the co-expression networks from the prodigious amount of RNA-Seq data generated by these consortia, and coupled them to orthology relationships of genes between species. 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.

We also have extensive experience in combining these integrated data into mathematical models. For example, we established the regulatory relationships between transcription factors and target genes using a probabilistic model-based method, TIP (Target Identification from Profiles) \cite{22039215}. We have applied machine-learning methods that integrate multiple genomics features to classify regulatory regions from ENCODE data of >100 TF binding sites. In particular, we were able to identify potential enhancers from regions classified as gene-distal regulatory modules \cite{22950945}. This was coupled to the construction of statistical models to predict gene expression levels based on TF binding and HM signals proximal to transcription start site (TSS) \cite{22060676,21177976,21324173,21926158,22955978}. We constructed linear and nonlinear models that utilize TF binding signals as input to predict the transcriptional output of a gene \cite{22955978} and applied these methods on a diverse set of model organisms from yeast to humans \cite{22060676,21177976,21324173,21926158} and achieved high predictive expression levels based on binding signals of 40 TFs in K652 \cite{22955978}.

In addition, we have related these integrated data and regulation models to changes in epigenetic profiles. For example, working in yeast we demonstrated the ability to predict histone modification signals from protein-protein interaction networks and position in the hierarchy of

transcriptional regulation. These grouped transcription factors into histone-sensitive and -insensitive classes that refined our ability to predict the targets of transcriptional regulation \cite{22060676}. Further, 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 will use these tools to determine these parameters for the motrpac data to refine our models for the predicted effects of exercise.

- **Proposed Research:**

** Build models (DS) - 1pg

We will integrate the motrpac multi-'omic and phenotypic data with that of other large consortia, particularly the ENCODE regulation data, the GTEx tissue-specific profiles and the Epigenome RoadMap effects of various epigenetic marks on transcriptional response. We will use these data to determine how acute changes in transcription and metabolite and protein levels are related to a epigenome changes, tissue modification and other durable responses.

First we will use logical clustering to organize gene expression signatures by clinical states, e.g. pre-exercise, two hours post exercise, etc. Then we will then use regulatory data to organize the clusters into networks and modules. The generation of this network is the entrypoint into many integrative analyses and models. Other data types will be layered onto this construct for further integration.

We will use these integrated networks create a 'molecular map' of physical activity in humans, which will model the responses the exercise. Importantly, changes in the network will focus on the relationship of the data to epigenetic effects that have been shown to be important in an individual's response to exercise.

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

DS .5 pg on relating genetic variation (QTLs)

We will relate the multi- 'omic and integrative datasets to the changes within individuals by identifying Quantitative Trait Loci ("star-QTL") 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 effects of, and variation in response to, exercise.

Preliminary Results

- integration of different types of omics data with genetic variants ("star-QTL") & allelic analysis [.5 pg]

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. 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 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 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 will build allelic profiles of individuals using their genomic data and cluster these with exercise outcomes to build quantitative trait loci. While this may be related to particular SNPs or expression profiles, it need not be limited to genomic and transcriptomic information. We propose to build protein and metabolite QTL's (pQTLs and mQTLs) that may provide a more direct path to functional responses or diagnostic tests.

Expected Outcomes:

Potential Problems & Alternative Approaches:

Timetable:

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

<introductory paragraph>

Research Design:

(a) Track biospecimen processing and provide an integrated view of other resources across the consortium.

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.

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

Plan: 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.

(b) Provide expertise in data management and analytics.

Preliminary Study: Data Analysis consultation within the ExRNA Consortium, enabled by shared communication (Redmine/Genboree Commons forums), shared metadata (GenboreeKB), data and tools (Genboree Workbench), and on-line materials.

Plan: Enable collaboration, sharing data, tools and pipelines. Define best data analysis practices and “methodological templates”. Organize workshops to educate consortium members about analysis methods. Develop on-line materials and resources that may be used by both consortium members and the general scientific community.

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

Preliminary study: modENCODE?

Plan:

Expected Outcomes:

Potential Problems & Alternative Approaches:

Timetable:

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)

Initial Thoughts on an Outline:

Grant Outline
12 pages overall (6 - 4 - 2 effort division)

* 1 pg intro

* 6 pg to DCC
(incl. 1/3 page to be slotted in on MG experience in privacy, cloud computing, integration w/ lit.)

* 4 pg to DAC

- Prelim. Results [2 pg]
- Upstream processing [1 pg]
+ Setting up standards & pipelines [.5 pg]
+ Integration w/ encode, exRNA [.5 pg]
(incl. sensitivity to extracellular & cellular information)
- developing integrative omics models & tools for these [1pg]
+ developing approaches for normalization, registration & analysis of temporal data
+ developing tools to analyze dynamic data & longitudinal variation
+ developing approaches for tissue de-convolution (⅛ page from Aleks)
- integration of different types of omics data with genetic variants ("star-QTL") & allelic analysis [.5 pg]

* 1 pg on consortium activities