http://bit.ly/mglab-DIRC

Notes

http://info.gersteinlab.org/Summaries
https://grants.nih.gov/grants/guide/rfa-files/RFA-RM-18-031.html
https://commonfund.nih.gov/pain/https://www.nih.gov/news-events/news-releases/nih-research-program-explore-transition-acute-chronic-pain

9th - today
Check in call on F or M
12 Fri - Alex gdoc
15 Mon. - hard done word doc (they'll do references)
17 Wed. (late)
18 Thurs.
Due at NIH 10/24

Word Count (Mon. afternoon before MG edits)

Specific Aims = 614
Intro = 747
Innovation = 186
Aim 1 = 2739
Aim 2 = 1527 (w/o Imaging tools)
Aim 3 = 3132 (?)
Total = (8945) - 7500 is 10pg

TXN
PRO
MET
AIM1
Y?1
AIM2
Y?1
Y?2
CLUSTOLOGY
Proomics
AIM3
Y2
Y3

{{Julie START ==>

Specific Aims

We propose to execute the Data Integration and Analysis Component (DIAC) of the Data Integration and Resource Center (DIRC) for the Common Fund Acute to Chronic Pain Signatures (A2CPS) Program for this consortium. This component will function in order to help the overall consortium with its data integration and analysis needs. We anticipate that these needs and the component's responses will be broken into three aims.

Aim 1. Construction of high-throughput pipelines for the analysis of transcriptomic, proteomic, metabolomic and lipidomic data. First of all, we will set up a number of high-throughput pipelines at the DCC to help the consortium process large scale data that will be generated. In particular, we will set up a number of pipelines to enable the processing of transcriptomic and proteomic data, and we will also manage and develop additional large-scale pipelines for other -omics data types (eg metabolomics and lipidomics), as they become necessary. As we will demonstrate, we have substantial experience in setting up high-throughput pipelines in the context of other consortia.

Aim 2. Development of analysis tools for visualization and identification of acute to chronic pain signatures. We aim to build a number of tools that combine the output of results from the pipelines from aim 1 that comprise a variety of available data types and begin to develop candidate signatures of the acute to chronic transition. We anticipate, given the multi-omics and multi-center nature of the A2CPS consortium, that making these tools available will be extremely useful to members of the consortium and the wider scientific community. We specifically foresee the need to combine the various -omics data type with neuroimaging data. We will leverage our experience in developing tools that can cluster the transcriptomics data in terms of a variety of simple phenotypic and genotypic changes. The neuroimaging aspect of the data integration will provide a multi-modal analytic platform along with the ability to map both transcriptomic and neuroimaging features into the same atlas space. This will enable discovery of brain-wide data-driven markers of chronic pain signatures, which can in turn be merged with other -omics datasets. These tools will be constructed collaboratively with members of the consortium based on specific priorities as directed by the Analysis Working Group (AWG). Such tools will enable the identification of signatures and potential biomarkers that distinguish acute from chronic pain individuals using the available -omics datasets from the cohorts under investigation. We will also develop tools to integrate -omics data with available electronic health records (EHR) data from the cohorts studied.

Aim 3. Perform and publish integrative analyses investigating acute to chronic pain. In the third aim, we will help lead large-scale integrative analysis efforts on the data from the A2CPS consortium. These analyses would be based on our prior experience conducting integrative analyses for other consortia (eg ENCODE and ERCC). We anticipate that the analysis will involve connecting the consortium's data with many complementary data types from external sources. Essential external data sources include databases of common and disease variation, as well as phenotypic characterization. We will describe our large-scale experience for these types of data sets, with the expectation that such integration constitutes a major part of the DIRC DIAC endeavor. We will help organize the Analysis Working Group (AWG) and lead the consortium in publishing integrative analyses using -omics data to investigate the onset of acute from chronic pain.

Overall, we will demonstrate that, as relevant to the mandate of the DIRC DIAC, we have extensive experience in performing integrative analyses and leading the publication of these results for several large genomics consortia. Furthermore, we aim to show that our response to the data challenges presented by the consortium will be both comprehensive and state-of-the-art.

Introduction

Acute pain caused by injury, surgery or disease may persist as chronic pain after the initial trauma. Such a transition of acute to chronic pain poses a major burden on pain care and management, and is particularly crucial for post-injury interventions, but the mechanism of development of chronic pain is currently poorly understood. Consequently, the A2CPS Program aims to collect extensive data on the transition from acute to chronic pain. Such an endeavor demands a concurrent drive towards the integration of the data in a coherent, interpretational framework. In light of this, we propose to execute the Data Integration and Analysis Component (DIAC) of the Data Integration and Resource Center (DIRC) for the Common Fund A2CPS Program. This component will function in order to help the overall consortium with its data integration and analysis needs.

The transition from acute pain to chronic pain

The arousal of chronic pain may associate with neuroplastic changes in the central nervous system (CNS), and has little relevance to the nature of the original stimuli. The amplification of neural signalling in the nociceptive system within the CNS, namely central sensitization, leads to heightened pain sensitivity after being triggered by the initial injury or inflammation \cite{3220875, 3268359}. Specifically, central sensitization causes previously subthreshold synaptic inputs, which do not normally drive any output, to generate increased or augmented action potential output \cite{2750819}. In a broader sense, researchers have proposed that transition to chronic pain involves continuous neural reorganizations of the CNS

\cite{18952143}. These changes may be detected and characterized by transcriptomic alterations in CNS tissues, peripheral extracellular contexts, as well as the circulating system.

Altered transcriptional regulation related to chronic pain

Transcriptome profiling has enabled the characterization of several differentially expressed genes associated with chronic pain in dorsal root ganglion and spinal cord tissue of rats and mice after nerve or inflammatory surgery \cite{24472155, 21561713}. It has also been observed that several types of chemokines are significantly upregulated over a time scale of two weeks in peripheral tissues of the femorotibial joint in rats after induced chronic joint pain \cite{3835139}. A recent study has characterized over 8,000 eQTLs associated with susceptibility and maintenance of chronic pain in human dorsal ganglia \cite{28564610}.

Epigenetic modifications also play a role in regulating the expression of genes related to the transition from acute to chronic pain. This includes methylation and downregulation of genes associated with accelerated disc generation \cite{21867537}, and demethylation-induced aberrant production of cytokine in osteoarthritis patients \cite{2788707}. Studies of expression and regulation of genes related to chronic pain development may provide diagnostic markers and targets for personalized intervention.

Circulating RNAs as potential predictors

Circulating RNA markers may be used as a source for non-invasive biological signatures related to acute to chronic pain transition. Circulating, or extracellular, RNA refers to a group of RNAs detected outside the cellular context especially body fluids. Several studies have identified differential expression of some circulating RNAs, especially miRNAs in body fluids, related to the development and treatment of chronic pain. Researchers have found that mice after spinal nerve ligation surgery display increased or decreased expression of several miRNAs in serum, and some of their target genes relate to the activation of cell signalling associated with nervous lesions \cite{25274330}. Altered miRNA profiles are also detected in the cerebrospinal fluids for patients with fibromyalgia, a disorder characterized by chronic pain and related to central sensitization \cite{24205312}.

Some plasma miRNAs were also found to have commonly altered expression levels for patients after treatment with opioids, which is generally effective for chronic pain, and may serve as diagnostic markers for clinical outcomes \cite{4110167}. Generally, a systematic study of the significance of circulating RNAs in the development of chronic pain is still lacking. The function and origin of RNAs detected in body fluid requires further study. Large-scale analysis of transcriptomics of RNAs from body fluids in larger cohorts would further facilitate an understanding of the role of circulating RNAs in the development of chronic pain.

Importance of neuroimaging techniques

Neuroimaging has enabled noninvasive investigation of abnormally altered activities in the CNS. It has been observed that several types of chronic pain are associated with regional changes in gray matter density \cite{20236763}, abnormal interactions between gray and white matter

\cite{19038215, 19035484}, altered functions in various brain regions \cite{22961548, 9252330, 18184777}, and altered connectivity in the default-mode networks (DMN) \cite{18256259, 20506181}. It should be noted that some of these studies show variable or even opposite changes for different types of chronic pain, suggesting high variability of the syndromes and complexity of the function of the CNS. Accumulation of high spatial and temporal resolution imaging data and incorporation of novel pattern recognition methods would help to identify neurological signatures and facilitate our understanding of the role of the CNS in the development of chronic pain \cite{5289824}, and could aid our integrative analysis of the cellular and extracellular transcriptomics.

Innovation

Prediction of the risk of transition into chronic pain is crucial for personalized prevention, and calls for further detailed investigation of the underlying mechanisms. This requires the accumulation and processing of large amounts of integrative data from multiple genomic sources and the integrative analysis of these data. Thus, the DIRC DIAC, as part of the Acute to Chronic Pain Signatures (A2CPS) Program, plans to identify biological signatures of patient susceptibility, the biological processes and pathways related to the development of chronic pain, and potential treatment targets by integrating diverse datasets including health records, brain imaging and other omics studies. The scale of the data proposed to be generated by the A2CPS consortium has to date never been studied by the pain research community and is in of itself significantly innovative. In addition, the vast amounts of the diverse data and the breadths of the explored systems in the body will demand innovative interpretational frameworks and analysis tools.

Aim 1) Running Pipelines

1.1 Preliminary Results

Transcriptomics

We have extensive expertise with transcriptome analysis and in developing a wide range of customized tools, as well as building standardized pipelines for analysis and uniform processing of both long and short RNA-Seq data. These tools have been evaluated and implemented in several major consortia, including long RNA-Seq analysis tools (in modENCODE \cite{19536255}) and short RNA-Seq pipelines for the analysis of small extracellular RNA-Seq data (in the Extracellular RNA Communication Consortium (ERCC) \cite{26320938}).

For general RNA-Seq analysis, we have developed an efficient in-house data processing workflow for long RNA-Seq data that includes data organization, format conversion, and quality assessment. RSEQtools (http://rseqtools.gersteinlab.org/), is a computational package that enables expression quantification of annotated RNAs, as well as identification of splice sites and gene models \cite{21134889}. Comparisons between RNA-Seq samples, and to other genome-wide data, are facilitated in part by our Aggregation and Correlation Toolbox (ACT), which is a general purpose tool for comparing genomic signal tracks \cite{21349863}. An important challenge in RNA-Seq analysis is detecting unannotated transcription that may be hard to distinguish from noise. 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}. To investigate newly 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, since these transcripts could be important biomarker for diseases such as various types of cancer and mental diseases. To address this, we created FusionSeq to detect transcripts that arise due to trans-splicing or chromosomal translocations \cite{20964841}. Our lab has also constructed IQSeq \cite{22238592}, which calculates the relative and absolute abundance of contributing transcript isoforms to a gene from RNA-seq data using a fast algorithm based on the Fisher information matrix. We developed Pseudo-seq \cite{22951037} to addresses the issue of quantification of pseudogene expression, which is difficult to separate from the transcription of parent genes with similar sequences. Another 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.

We recently developed the extracellular RNA processing toolkit, exceRpt (http://github.gersteinlab.org/exceRpt/), a set of tools and a pipeline designed for comprehensive analysis of small RNA-Seq datasets: read preprocessing, filtering and alignment, biotype abundance estimation, visualization and quality assessment. It is specifically designed to handle technical issues that are often characteristic of small RNA-Seq samples, such as those obtained from extra-cellular preparations. The exceRpt pipeline is used for uniform processing of hundreds of RNA-Seq datasets submitted to the exRNA Atlas (http://exrna-atlas.org/) repository.

[[consortium experience in txn??]]
We also have extensive experience conducting integrative analyses of large sets of RNA-seq data. We have worked on the development and analysis of multiple RNA-seq flows in the context of large consortia, including the implementation of tools we developed and other popular tools such as Bowtie and Tophat. We describe our consortium experience further in aim 3. We played a lead role in the analysis of model organisms (such as C. elegans) and human transcriptome studies \cite{22955620, 21177976} within the ENCODE consortium

\cite{22955616}. Currently, we are active participants of the Brainspan project, which profiles RNAs in different parts of the human brain (http://www.brainspan.org), as well as the PsychENCODE project \cite{26605881}. We are involved in the coordination of the RNA-seq working group activities for the ENCODE project; and we lead the data integration and analysis component (DIAC) of the data management and resource repository (DMRR) for the NIH ERCC (http://commonfund.nih.gov/exrna/).

Proteomics

We have substantial experience with the analysis of proteomic data \cite{19817483, 17923450, 22583803} and its integration with genomic data, such as the combination of mass spectrometry (MS) proteomic and transcriptomic data \cite{25349915, 17519225}. Specifically, we have constructed a web tool called PARE (Protein Abundance and mRNA Expression; http://proteomics.gersteinlab.org), to correlate these two quantities \cite{17718915}. We also published the tool EMpire \cite{30125121}, which uses transcript-level RNA-seq expression as a prior likelihood and enables protein isoform abundances to be directly estimated from LC–MS/MS, an approach derived from the principle that most genes appear to be expressed as a single dominant isoform in a given cell type or tissue. We have also led studies interpreting protein-protein interactions based on data from proteomic experiments \cite{15491499, 14564010}. We have been members of numerous NIH proteomics projects and consortia, including the Northeast Structural Genomics Consortium, the NHLBI Proteomics Center and the Yale/NIDA Neuroproteomics Center, and have conducted analyses on the large scale proteomic data generated by these consortia \cite{12952525, 17923450}.

Metabolomics

The Metabolomics Consortium Data Repository and Coordinating Center (DRCC) at the University of California San Diego (PI: S. Subramaniam) has recently processed and curated it's 1,000th metabolomics study. This collection of experimental datasets contains submissions from over 200 different institutions around the world and represents over 70 different species with the majority coming from human (47%) and mouse (31%) sample sources. Analytical methods used in these studies include untargeted/targeted LC-MS (67%), GC-MS (21%) and NMR (12%). Studies are available for browsing, analysis and download (subject to embargo release) in the NIH Data Repository section of the website. The DRCC is actively accepting metabolomics data for small and large studies on cells, tissues and organisms via the Metabolomics Workbench. The Metabolomics Workbench serves as a national and international repository for metabolomics data and metadata and provides analysis tools and access to metabolite standards, protocols, tutorials, training, and more (**Figure 1**).

The UC San Diego Center for Computational Biology & Bioinformatics (CCBB) (Executive Director: Dr. K. Fisch, compbio.ucsd.edu), established in 2014 by the UC San Diego School of Medicine and UC San Diego Clinical & Translational Research Institute (NIH/NCATS CTSA), will leverage the Metabolomics Workbench [1] and XCMS [2] to develop open source,

automated and reproducible primary and secondary analysis pipelines for the A2CPS DIRC and will be responsible for data coordination and QC of lipidomics and metabolomics data. The CCBB provides investigators with bioinformatics expertise to analyze large molecular datasets in the areas of genomics, systems biology and translational medicine. The CCBB will bring systems biology and machine learning techniques to analyze and integrate metabolomics data with outcomes, EHR data, imaging data and multi-omics data to prioritize clinically relevant genes and generate novel biological insights. The UCSD CCBB has completed 285 investigator-initiated collaborative projects resulting in 43 peer-reviewed publications leveraging the scalable cloud-computing resources of Amazon Web Services, including a metabolomics analysis of rheumatoid arthritis [3] (**Figure 2**).

In summary, we have carried out a number of research and clinical studies that establish our expertise in the field of metabolomics, scientific ability as well as our capacity both technical and instrumental to successfully perform accurate and precise metabolomics measurements on a large scale and in high throughput settings.

Sequential references for Metabolomics under “Ref” at the end of the document

Lipidomics

The University of California, San Diego LIPID MAPS Lipidomics Core has been focusing on developing the field of lipidomics, especially targeting bioactive lipid mediators and biomarker development (1). The complexity of the lipidome both in dynamic range and structural diversity represents a major analytical challenge. To address this challenges, The LIPID MAPS Consortium (Director: Dr. E.A. Dennis; Scientific Officer and Investigator: Dr. O. Quehenberger) was created in 2003 as a multi-institutional effort to quantify all of the major and minor lipid species of the mammalian lipidome. When it ended in 2013, we leveraged all these technologies and established the LIPID MAPS Lipidomics core at UCSD (Director: Dr. O. Quehenberger, http://www.ucsd-lipidmaps.org).

We established the first comprehensive human lipid profile in plasma and identified and quantified some six hundred distinct lipid molecular species across all mammalian lipid categories (2). Immunologically-activated macrophages were also profiled and over 500 discrete lipid species were measured and associated pathways were mapped, integrating transcriptomics, proteomics and lipidomics (3). Our laboratory now routinely profiles plasma, urine, bronchial alveolar lavages, cerebral spinal fluid and various other tissues of both human and animal origin for biomarker discover and for indicators of abnormal lipid metabolism. More recently, we established lipid profiles of liver biopsy specimen and plasma from individuals with non-alcoholic fatty liver disease for biomarker development (4). Our lipidomics platform for monitoring over 200 oxidation and signal transduction consequences is the most developed platform to emerge in the metabolomics area (5-8). Pertinent to this application, we established

that inflammatory hyperalgesia induced bioactive eicosanoid production and inhibition of the underlying enzymatic systems in the spinal cord attenuated NSAID-unresponsive hyperalgesia in a rat pain model (9, 10) (**Figure 1**). We used the same platform to profile plasma from individuals with non-alcoholic liver disease of various severities and established an eicosanoid biomarker panel that is able to discriminate between steatosis and steatohepatitis (11). Similar approaches were used to identify eicosanoid targets in various bacterial and viral infectious diseases including Lyme disease and influenza (12, 13).

The UC San Diego Center for Computational Biology & Bioinformatics (CCBB) (Executive Director: Dr. K. Fisch, compbio.ucsd.edu), established in 2014 by the UC San Diego School of Medicine and UC San Diego Clinical & Translational Research Institute (NIH/NCATS CTSA), will leverage the expertise of the LIPID MAPS Lipidomics core to develop open source, automated and reproducible primary and secondary analysis pipelines for the A2CPS DIRC and will be responsible for data coordination and QC of lipidomics and metabolomics data. The CCBB provides investigators with bioinformatics expertise to analyze large molecular datasets in the areas of genomics, systems biology and translational medicine. The CCBB will bring systems biology and machine learning techniques to analyze and integrate lipidomics data with outcomes, EHR data, imaging data and multi-omics data to prioritize clinically relevant genes and generate novel biological insights. The UCSD CCBB has completed 285 investigator-initiated collaborative projects resulting in 43 peer-reviewed publications leveraging the scalable cloud-computing resources of Amazon Web Services.

In summary, we have carried out a number of research and clinical studies that establish our expertise in the field of lipidomics, scientific ability as well as our capacity both technical and instrumental to successfully perform accurate and precise lipidomic measurements on a large scale and in high throughput settings.

Sequential references for Lipidomics under “Ref” at the end of the document

1.2 Proposed Research

We will develop and test pipelines for the various omic data types generated by the A2CPS consortium including, in particular, pipelines for processing transcriptomics data (RNA sequencing), proteomics, metabolomics and lipidomics data types. We will evaluate existing published pipelines, as well as compare with current best practice approaches and pipelines used by other larger genomic consortia to process these data. Compatibility of our approaches with existing analysis pipelines from other relevant genomic consortia (such as the Extracellular RNA Communication Consortium) will enable easier integration with external data sources. The evaluation of these processing pipelines will be conducted under the supervision of the Analysis Working Group (AWG) of the consortium.

Analysis pipelines for the various omic data types will then be deployed at the DCC for processing of the data generated by the consortium. The DIRC DIAC will provide the pipelines

to the DCC in the form of a Github repository, as well as dockerized images for deployment. The DIAC will help support these pipelines, and modify and update them as needed to fulfill the potential evolving needs of the consortium.

The DIAC will also assess existing standards and, if necessary, develop new quality control (QC) metrics for evaluating the data being generated, in agreement with members of the consortium. The DIAC will incorporate these QC metrics as output from the analysis pipelines, and will routinely assess such output for the data processed by the DCC.

Metabolomics: Primary Data Analysis

Targeted metabolomics datasets will be analyzed using XCMS-MRM and METLIN-MRM, which are a cloud-based data-analysis platform and a public library, respectively, to perform signal processing to detect, integrate and align peaks across samples [4]. Untargeted LC/MS-based metabolomics data will be processed using XCMS for peak-picking and alignment, followed by peak annotation. Peak annotation includes peak grouping, using ion adducts to annotate features, making use of pathway information, integrating MS/MS data and incorporating retention time [5].

Metabolomics: Secondary Analysis

XCMS will be used to perform statistical testing between groups to identify biomarkers using Welch's *t*-test with unequal variances and the "HPLC/Q-TOF" parameter [2]. Dysregulated metabolic pathways will be identified using mummichog [6] using the entire metabolic feature table. Integrative analysis with other omics data, such as proteomics and transcriptomics will be employed using the autonomous multimodal metabolomics data integration approach described in_[7].

Sequential references for Metabolomics under "Ref" at the end of the document

Lipidomics: Primary Data Analysis

All data will be acquired by liquid chromatography and mass spectrometry (LC-MS) using various acquisition modes. The samples will be extracted using established procedures and then analyzed by LC-MS, essentially as described (2). The raw data will be labeled using the nomenclature that was established as a standard system by The LIPID MAPS Consortium and that allows integration of current data sets into databases (14, 15). The second step of data processing is to normalize the data via a set of internal standards. The third step of data processing performs quantitation if authentic standards are available for the measured metabolites. The fourth step will normalize the data to the amount of input material and the results will be expressed as concentration units (e.g., pmol/ml plasma). In some cases, absolute quantification may not be possible due to lack of authentic standards. In that case, we will express the data as ratio between measured lipid metabolite and corresponding internal standard. The ratios will be normalized to and expressed as ratio per ml plasma. As such, they represent relative concentrations and can be used for direct comparison of individual lipid metabolites between different samples as well as different metabolites in the same sample.

Lipidomics: Secondary Data Analysis & Coordination

For dissemination, we will create a table that contains all sample identification numbers, name of the metabolites that were detectable in the sample and the associated absolute concentrations. This table and raw data will be submitted to DCC by the UCSD CCBB. Lipid biomarkers of acute to chronic pain will be computed by taking the log10 quantity ratios between conditions, calculating statistical significance adjusting for multiple testing and pathway analysis will be performed to identify active pathways that are dysregulated. Pathway analysis will be performed by computing a Z-score for each weighted pathway, based on the molecular concentration of lipid species across all possible lipid pathways from Reactome following the methods in (16). To integrate lipidomics with other omics data, we will employ both cluster based and network based approaches (17).

Lipidomics: Quality Control Analysis

Following the guidelines outlined in Good Laboratory Practice Standards (USEPA), validation assays are performed regularly for all the lipids using routine analytical preparation procedures. Every thirty sample, a quality control sample will be analyzed. The quality control sample will consist of the human plasma standard reference material SRM 1950, collected by NIST in collaboration with NIH. We previously established a comprehensive and quantitative lipid profile that covered over 600 lipid species. As we recorded the exact concentrations of these lipid species, repeated analysis of the reference material will serve as a quality control of our data set. As additional quality controls, we will use the retention times and mass spectral intensities of the internal standards. As an example, for the analysis of eicosanoids we will use 26 deuterated internal standards. The deuterated standards are easily identifiable in the spectra and will be used to gauge potential retention time drift in the chromatogram. To prevent misidentification of endogenous metabolites due to retention time drifts, all spectra will be aligned based on the retention times of the internal standards. We will add the internal standards to all samples at exactly the same amounts. Thus, they will serve as additional quality control to compensate for any variations in analytical sensitivity of the mass spectrometer. For quantification, we will create standard curves with quantitative standards that also contain the internals standards at the same concentrations as the samples. Our eicosanoid standard library for quantitative analysis contains over 140 authentic eicosanoid. Nine point standard curves will be generated and used to calculate the exact concentrations of the endogenous metabolites in the samples.

Sequential references for Lipidomics under “Ref” at the end of the document

Aim 2) Building Tools

2.1 Preliminary Results

Transcriptome tools [[for clustering & sig. analysis??]]

We will leverage our extensive experience processing and analyzing transcriptomic data in addressing the aims of the DIRC DIAC. In the interests of constructing the most robust set of tools for transcriptome analysis, we evaluated 24 protocol variants of 14 independent computational methods for exon identification, transcript reconstruction and expression level quantification from RNA-seq data \cite{24185837}. Our results characterize the strengths and weaknesses of these methods, which would aid the design of analytical strategies.

~~We have designed a series of RNA-seq processing frameworks that enable accurate profiling of the transcriptome. We have also developed several transcriptome analysis tools, ranging in scope from dynamic gene expression to orthology comparison. These tools extract simple signatures on various levels from the transcriptome that may be associated with diseases or disorders of interest. One such suite of tools, RSEQtools, uses the Mapped Read Format (MRF) for the analysis of RNA-Seq experiments, and performs common tasks such as calculating gene expression values, generating signal tracks of mapped reads, and segmenting that signal into actively transcribed regions \cite{21134889}. These tools can readily be used to build customizable RNA-Seq workflows. In addition to the anonymization afforded by MRF, this format also facilitates the decoupling of the alignment of reads from downstream analyses. Another tool at our disposal is IQSeq (Isoform Quantification in next-generation Sequencing), which tackles the problem of gene isoform quantification\cite{22238592}. To measure the accuracy of an isoform quantification result, one estimates the average variance of the estimated isoform abundances for each gene by resampling the RNA-seq reads, using the Fisher Information Matrix to calculate this . The tool is available at archive.gersteinlab.org/proj/rnaseq/IQSeq.~~

Following transcriptomic data processing in the aforementioned manner, several downstream analyses can be conducted to identify the functional and regulatory implications of the observed gene expression patterns. We developed a computational method (DREISS, dreiss.gersteinlab.org) for analyzing the “Dynamics of gene expression driven by Regulatory networks, both External and Internal, based on State Space models” \cite{27760135}. This tool evaluates the temporal dynamics of subnetworks of genes, by differentiating between internal and external sources of impact on these subnetworks. DREISS employs dimensionality reduction to help identify canonical temporal expression trajectories (e.g., degradation, growth and oscillation) representing the regulatory effects emanating from various subsystems (**Aim 2 Figure 1**). Another such tool, Loregic (github.com/gersteinlab/loregic), is a computational

method integrating gene expression and regulatory network data to characterize the cooperativity of regulatory factors \cite[25884877}. Loregic use all 16 possible two-input-one-output logic gates (e.g. AND or XOR) to describe triplets of two factors regulating a common target. The tool finds the gate that best matches each triplet's observed gene expression pattern across many conditions (**Aim 2 Figure 2**). Loregic is able to characterize complex circuits involving both proximally and distally regulating transcription factors (TFs) and also miRNAs, as demonstrated on human ENCODE ChIP-Seq and the Cancer Genome Atlas (TCGA) RNA-Seq data. Additionally, cross-species data can be exploited by OrthoClust, a computational framework for simultaneously clustering data across multiple species \cite{25249401}. It integrates the co-association networks of individual species by utilizing the orthology relationships of genes between species. It outputs optimized modules that are fundamentally cross-species, which can either be conserved or species-specific (**Aim 2 Figure 3**). The application of OrthoClust was demonstrated using RNA-Seq expression profiles of Caenorhabditis elegans and Drosophila melanogaster from the modENCODE consortium. A potential application of cross-species modules is to infer putative analogous functions of uncharacterized elements like non-coding RNAs based on guilt-by-association.

Tools for the Deconvolution of Tissue-level Data

Deconvolution refers to the decomposition of a dataset into its constituent components, such as functional modules or cell types in bulk tissue. In exRNA studies, deconvolution methods can help us identify fractions in the bulk expression data associated with specific cell types, and their corresponding characteristic expression patterns. We have previously employed several deconvolution analysis methods that can be integrated into the exRNA pipeline, in order to specify subtypes of cells associated with signatures of interest.

We have employed two approaches to the bulk tissue deconvolution problem \cite{Wang et al 2018 (capstone4)}: an unsupervised approach, non-negative matrix factorization (NMF); and a supervised approach, cell-signature-based decomposition. In the NMF approach, the bulk tissue gene expression matrix X (dimensions = N by M, where M = number of samples and N = number of selected genes (e.g., biomarker genes)), is decomposed into the product of two matrices, H and V: H is a K by M matrix with the (i,j) element describing the contribution of the j^{th} NMF "top component" (NMF-TC) to the ith sample, K is the number of selected NMF-TCs (e.g., equal to the number of selected cell types), and V is an N by K matrix with the (i,j) element being the expression level of the j^{th} gene in the i^{th} NMF-TC. Conceptually, V describes the gene expression pattern of each characteristic "top component", while H provides the weight of each "top component" in the observed samples of X. We found that NMF-TCs recovered the expression patterns of different cell types in bulk RNA-Seq data on brain cell population. This suggests that it is highly likely that a linear combination of single-cell components contributes to the overall expression pattern of the sample. Therefore, we aimed to more accurately identify the fractions that determine the sample expression. We then applied a supervised approach that uses single-cell expression signatures to find the fractions of different cell types.

In particular, we defined the sample gene expression matrix B (N by M) for a phenotype/disorder, where M and N are defined above), and C is the fraction gene expression matrix is C (N by K), where K is the number of selected cell types. We used the non-negative least square method to find a non-negative K by M matrix, W: the (i,j) element of W represents the linear combination coefficient of the i^{th} fraction to the j^{th} sample expression. Applying this method to bulk RNA-Seq data on a brain cell population, we identified cell-fraction changes associated with different traits (**Aim 2 Fig 4**). For example, there were different fractions of particular types of excitatory and inhibitory neurons in male and female samples, with the fraction of In6 being significantly higher in females. We then validated the method on an independent subset of samples to predict cell population fractions, finding that our estimations were close to the experimental fractions.

Network Analysis and Visualization Tools

Cmpreg: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4336544/
Encodenets: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4154057/

[[we have many papers on the developing a heir struct of the regulatory network both for txn & post-txn reg. We hace a number of alg to build the heir including sim anneal (encodenet), Hy(bfs), CC...
]]

We have demonstrated experience in biological network science. Following the identification of functional and regulatory networks from aforementioned pipelines and tools, the properties of these networks will be quantified and visualized to identify possible signatures of dysregulation in the transition to chronic pain.

Our lab has developed various tools for network analysis from multiple perspectives, including tools to determine small-scale network motifs such as feed-forward loops and feedback loops as well as large-scale structures such as overall network hierarchies, center points of networks, bottlenecks of networks, and so forth. These tools have been used to analyze the human regulatory network, the network associated with cancer, the phosphorylation network in yeast, the yeast regulatory network, and other model organism networks \cite{25880651]. We have performed extensive comparisons between these regulatory networks and published many comparative network papers \cite{20439753}.

TopNet is an automated web tool designed to calculate topological parameters and compare different sub-networks for any given network \cite{14724320}. TopNet takes as an input an arbitrary undirected network and a group of node classes to create sub-networks. It then computes a variety of topological parameters, such as average degree, clustering coefficient, characteristic path length, etc., and calculates the power-law degree distribution for each

sub-network. TopNet also enables the user to explore and visualize the complex networks: all first neighbors of a certain node can be shown in a simple graph; after the user defines two nodes of interest and a maximum path length, the sub-network between these two nodes with all the nodes on the paths within the maximum path length can be drawn in an independent graph (**Aim 2 Figure 5**).

In addition, we developed TopNet-like Yale Network Analyzer (tYNA), a Web system for managing, comparing and mining multiple networks, both directed and undirected \cite[17021160}. tYNA efficiently implements methods that have proven useful in network analysis, including identifying defective cliques, finding small network motifs (such as feed-forward loops), calculating global statistics (such as the clustering coefficient and eccentricity), and identifying hubs and bottlenecks. It also allows one to manage a large number of private and public networks using a flexible tagging system, to filter them based on a variety of criteria, and to visualize them through an interactive graphical interface. A number of commonly used biological datasets have been pre-loaded into tYNA, standardized and grouped into different categories. It can be used for managing, comparing and mining multiple networks, both directed and undirected. tYNA efficiently implements methods that have proven useful in network analysis, including identifying defective cliques, finding small network motifs (such as feed-forward loops), calculating global statistics (such as the clustering coefficient and eccentricity), and identifying hubs and bottlenecks etc (**Aim 2 Figure 6**).

Imaging tools

*** Alan : ask 2 pages on tools for image proc & rel this to omics
Add this Experience relating image data to omics data.

We have a number of papers relating omics data to imaging data and we have developed a formalism using canonical correlation analysis to interlace these two quantities to find the best correlation. (Refer to the metagenomics papers and also TARA's genome pathology paper)

2.2 Proposed Research

[[
we plan to apply loregic
We plan to develop sig tool from orthoclust
We plan look at the time series txn & meta data w/ driess
We need to add Alan's stuff
]]

We plan to develop a number of tools to identify candidate biomarkers and combine them into biosignatures predictive of the susceptibility or resilience to the development of chronic pain after an acute pain event. These tools will also be helpful identifying signatures and potential biomarkers that distinguish acute from chronic pain individuals. These tools will be developed collaboratively with members of the consortium based on specific priorities as directed by the Analysis Working Group (AWG). Specifically, we will evaluate and compare a number of commonly used supervised and unsupervised data mining methods, such as Robust Feature Selection (Saeys et al., 2008), Principal Component Analysis \cite{18327243}, Support Vector Machine-Recursive Feature Elimination (Guyon et al., 2002), for the search and prioritization of biomarker candidates from proteomics, extracellular RNA, lipidomics, metabolomics, transcriptomic, and possibly other data types as determined by the consortium. This analysis will be conducted by integrating electronic health records, patient-reported outcomes, and imaging data. We will build upon our evaluation of the aforementioned approaches and develop software that provide diverse functionality for the analysis of A2CPS datasets. The software will be made modular, open-source, user-friendly, and will include appropriate documentation and easy-to-follow tutorials. It will be crafted such that it requires little external dependencies, is straightforward to set up, and can work as a stand-alone package. The tool will not duplicate existing software with similar features. It will use standard formats for data input and output to facilitate its use and interoperability with other software.

Aim 3) Integrative Analysis [3321 words]

3.1 Preliminary Results

Approach for power analysis
*** 1pge from [PE] + Power analysis [LS] - futility analysis + [JZ]

How powerful are summary-based methods for identifying expression-trait associations under different genetic architectures? https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5785784/
Interpretation of TWAS and its vulnerabilities
http://sashagusev.github.io/2017-10/twas-vulnerabilities.html

https://grants.nih.gov/grants/guide/rfa-files/RFA-RM-18-031.html
How would we approach this ?

lead the consortium in a futility analysis (when 50% of the enrolled patients complete the 6 month phenotyping) to determine whether the rate of transition to chronic pain and subject retention and patient retention is adequate to meet the assumptions of the power analysis.

- ##################### JZ newly added start #####################

JZ2MG: without the specific aim and approach found, I can only write non-specific stuff here. One key issue in transcriptome-wide expression-trait association studies is the test-statistic inflation, which can lead to significant over-estimation of associated genes. We aim to avoid such false positive inflation. Specifically, we will corrects for known biological and technical covariates by adding them into our model and also estimate unobserved covariates by adjusting the residual bias and inflation using the empirical null distribution. We will adopt a similar approach as the BACON package in R to calculate the associations. Previous studies shows that with effective bias correction, the statistical power can be as high as 0.85 with the sample size around 500 \cite{“Confounder Adjustment in Multiple Hypothesis Testing”, pubmed 28129774}.
##################### JZ newly added end #####################

As part of the integrative analyses, we aim to carry out transcriptome-wide association studies (TWAS) to identify genes whose expression patterns are significantly associated with the acute to chronic transition. However, this necessitates an assessment of the power of the study, as well as the impact on the power due to subject attrition, or a so-called futility analysis.

One key issue in transcriptome-wide expression-trait association studies is test-statistic inflation, which can lead to significant over-estimation of associated genes [9]. We aim to avoid such false positive inflation. Specifically, we will correct for known biological and technical covariates by adding them into our model, and also estimate unobserved covariates by adjusting the residual bias and inflation using the empirical null distribution. We will adopt a similar approach as the BACON package in R to calculate the associations. Previous studies shows that with effective bias correction, the statistical power can be as high as 0.85 with the sample size around 500 \cite{“Confounder Adjustment in Multiple Hypothesis Testing”, pubmed 28129774}.

This has always been a major concern in GWAS, but inflated test statistics are also observed in EWAS [10, 11]. Often the level of inflation exceeds that observed in GWAS, yet it is generally not corrected [12]. In GWAS, test-statistic inflation is commonly addressed using genomic control in which the inflated test statistics are divided by the genomic inflation factor. The genomic inflation factor estimates the amount of inflation by comparing observed test statistics across all genetic variants to those expected under the hypothesis of no effect [9]. Recent work pointed out crucial limitations of genomic control in GWAS [13, 14]. Notably, the genomic inflation factor was shown to provide an invalid estimate of test-statistic inflation when the outcome of interest is associated with many, small genetic effects [13]. In EWAS and TWAS, this is the rule rather than the exception. Moreover, test statistics may not only be subject to inflation but also to bias [15], which is not corrected for when using genomic control. Bias of test statistics leads to a shift in the distribution of effect sizes and is driven by confounding [16, 17], a prominent feature of EWAS and TWAS but much less of a concern in GWAS [18]. Thus, this calls for the development of new methods specifically designed to address test-statistic inflation and bias in EWAS and TWAS analyses.

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5785784/

We developed a statistical method, using aggregated burden tests to look for differential burdening between populations and to use this to rank genomic regions. While our approach is not striving for absolute statistical significance in differential burdening, the sample size provides an appreciable signal for ranking. For the power aspect of burden tests applied to our sample populations, we used the SKAT package (available from the R project) on several population models for genomic regions of 5000 nucleotides to estimate the sample size needed to obtain statistical power (Figure 1).

Specifically, we explored two key dimensions of the resource and the ENCODE data: breadth across cell types and depth across assays to construct a deep, integrated annotation with two key characteristics: 1) noncoding elements are compactly defined to more precisely locate functional sites, and 2) these discontinuous regulatory regions are linked to genes to form extended-gene definitions. Extended genes are highly dynamic and may change considerably across cell types (similar in fashion to cell-type specific isoforms for conventional gene structures).

Overall, our annotation is compact to maximize the statistical power in somatic burden analysis in two respects: it contains fewer total elements (because the deep integration across many assays removes many potential false positives) and each individual element tends to be shorter in length yet is more enriched in functionally relevant nucleotides. In principle, both these facts benefit statistical power through decreasing multiple testing burden or more sharply defining core regions by removing nonfunctional nucleotides in each element. We also linked together the above compact annotation elements to define extended gene structures, which may also increase power in many circumstances. It potentially enriches the number of functional sites being tested, thus increasing power. Second, it helps with the interpretation of noncoding elements by linking them to genes. Third, it allows us to subset non-coding annotations by the many well-known gene categories, for instance, cancer-associated and metabolic genes.

Integrative analysis of genomic variants

We have extensively analyzed patterns of variation in non-coding regions and their coding targets \cite{21596777}{22955619}{22950945}. In recent projects \cite{24092746,25273974}, we integrated multiple methods into a comprehensive prioritization pipeline called **FunSeq** (**Figure 2**). The pipeline identifies sensitive regions with annotations under high selective pressure, links non-coding mutations to their target genes, and prioritizes variants based on network connectivity. It also identifies deleterious variants in non-coding elements including TF binding sites, enhancers, and regions corresponding to DNase I hypersensitive sites. Using integrated data from large-scale resources (including ENCODE and 1000 Genomes Project) with cancer genomics data, FunSeq can prioritize known TERT promoter driver mutations. Recently, we developed RADAR by extending the variant prioritization framework to the RNA transcript level

\cite{Genome Biology in press}. RADAR integrates the ENCODE eCLIP data and other genomic information to pinpoint deleterious variants, such as splicing-disruptive ones.

Additionally, we have developed a variety of tools that prioritize protein-coding variants. VAT (Variant Annotation Tool) characterizes variants according to affected genes and transcript isoforms \cite{22743228}, while ALoFT (Analysis of Loss of Function Transcripts) predicts loss-of-function (LOF) mutations and their impact (**Figure 3**) \cite{28851873}. Relatedly, our netSNP biological network integration tool \cite{23505346} identifies cancer genes based on connectivity. STRESS \cite{27066750} and Frustration \cite{27915290} are two other tools we built to identify mutations that affect allosteric hotspots in proteins and identify key functional protein regions prone to genetic alterations. Our Intensification tool searches for deleterious mutations within repeat regions of proteins \cite{27939289}. Finally, we developed a computational tool to systematically annotate uORFs (upstream open reading frames) in the genome \cite{29562350}. We applied this tool to predict the consequences of genomic variants and somatic mutations for affecting uORFs.

Further, we have published methods to identify allele-specific expression patterns. The AlleleSeq pipeline quantifies allele-specific expression \cite{21811232}, which can provide a direct readout of the effects of allele-specific variants (ASVs). We also conducted a study of allele-specific activity from RNA-Seq and ChIP-Seq experiments conducted on 1000 Genomes Project individuals \cite{23128226}{27089393}. After uniformly reprocessing all datasets, including ones from the gEUVADIS \cite{24037378} and ENCODE \cite{22955616}, we detected ASVs using a beta-binomial test to correct for overdispersion. We then combined the effects of multiple ASVs to assign allelicity scores to genomic elements, indicating that these elements are sensitive to mutations \cite{27089393}.

Integrative analysis of GWAS SNPs

*** Joel G [PE] - contribute ask 1 page on integration w/GWAS

Relating omics data to various human disorders & conditions

Sequencing of thousands of tumour samples has revealed the landscape of somatic mutations in protein-coding genes and non-coding regions. Besides, it has been shown that defects in non-coding regions could cause various brain disorders, such as autism, schizophrenia and Alzheimer's disease. These human disorders and phenotypes can be an ideal application for the comprehensive genomic annotations derived from various omics datasets.

Brain disorders. We lead data analysis for the PsychENCODE Consortium and participate in the BrainSpan Consortium, with several papers currently in the revision stage \cite{capstone 4}{capstone 1}{capstone development}. In recent work, we identified functional elements, QTLs and regulatory-network linkages specific to the adult brain by integrating data from the PsychENCODE Consortium together with relevant data from ENCODE, CommonMind (CMC), GTEx, and Roadmap \cite{capstone 4}. In addition to the adult brain, we also assessed the

degree of chromatin differences between developmental stages relative to that between tissues. Furthermore, we combined these elements and networks to build an integrated deep-learning model that predicts disorder phenotypes from genotype and functional elements (**Figure 4**) \cite{capstone 4}.
[[PE - reference our other brain related papers]]

***Cancer*.** We have published work focussed on noncoding sequence variants and mutational signatures for cancer drivers. We leveraged our expertise in non-coding regions in the first whole-genome analysis of TCGA kidney cancer (KIRP) samples \cite{28358873}. Our work found significant genomic noncoding alterations beyond traditional known drivers of KIRP located within coding exons (**Figure 5**). We were also able to unveil mutation patterns, signatures and tumor evolutionary structures, which reflect the mutagenesis processes and help understand how heterogeneity arises. We developed LARVA (Large-scale Analysis of Recurrent Variants in noncoding Annotations), a statistical method for identifying significant mutation enrichments in noncoding elements \cite{26304545}. LARVA includes corrections for biases in mutation rate owing to DNA replication timing. LARVA can also be targeted exclusively to coding regions to prioritize genes. We used this tool in a pan-cancer analysis of 760 cancer whole genomes' variants spanning a number of cancer data portals and some published datasets. Furthermore, we developed MOAT (Mutations Overburdening Annotations Tool), an alternative, empirical mutation burden approach that evaluates mutation enrichments based upon permutations of the input data \cite{29121169}. More recently, we have developed a network-based annotation for cancer mutations by leveraging thousands of functional genomics datasets from ENCODE cell types \cite{submitted}. Our analysis improves the understanding of different oncogenic transformations in the context of a broader cell space. Finally, we organized the whole ENCODE resource as a coherent workflow for cancer genomics to prioritize key elements and variants.

Experience leading consortium analysis activities
We are involved in and provided leadership to several large-scale national collaborations focused on aspects of genomics and data science.

ENCODE Consortium \cite{25164755}{25164757}. As part of a multi-institutional collaboration, we are involved in annotating the human genome and developing methods for analyzing large-scale genomic experiments. To facilitate the integrative analysis of the ENCODE datasets produced by different groups, we have developed and evaluated methods for uniform high-throughput sequencing data processing for all common platforms such as RNA-Seq and ChIp-Seq. In particular, we are working extensively on pseudogene identification and annotation of the human genome in collaboration with the GENCODE team members (http://www.gencodegenes.org/). We developed Peak-Seq \cite{19122651} to elucidate transcription factor binding sites and chromatin structure based on ChIP-Seq experiments. Furthermore, we developed MUSIC \cite{25292436}, a tool for identifying enriched regions in ChIP-seq data.

The 1000 Genomes Project \cite{26028266}. NIH's marquee effort on personal genomics, the sequencing of individual people's genomes, the project aims sequenced thousands of individuals' genomes to get a sense of their variability. We developed an annotation pipeline that maps SNPs, indels and structural variations (SVs) on to protein coding genes. We also developed algorithms to identify indels and structural variations based on split-read, read-depth and paired-end mapping methods. We are also involved in the 1KG SV trio project, a plan to sequence trios of individual from multiple families to very high coverage.

PsychENCODE Consortium \cite{26605881}{29439242}. We are working on PsychENCODE, a project aimed at understanding regulatory variants in the context of their functional connections to psychiatric disease. The project's approach involves a comprehensive examination of the genome, transcriptome, epigenome, and proteome in relation to brain function and disorders.

Pan-Cancer Analysis of Whole Genomes
\cite{https://www.biorxiv.org/content/early/2017/07/12/162784}{https://www.biorxiv.org/content/early/2018/09/07/179705}. We are heavily involved in the Pan-Cancer Analysis Working Group (PCAWG), an effort to combine all TCGA and ICGC whole genome sequencing data to improve our understanding of cancer. We are co-leaders of the PCAWG-2 group, and participate in the analyses of the PCAWG-3, 8, and 11 groups.

Extracellular RNA Consortium \cite{27112789}{27076901}. We are an integral part of the exRNA Consortium, a large-scale collaboration project aimed at establishing dada standards, a data portal, and tools and reagents to the scientific community. We developed the exceRpt (extra-cellular RNA processing toolkit) (http://github.gersteinlab.org/exceRpt), a pipeline for the analysis of extracellular small RNA-Seq experiments. exceRpt was designed to handle the variable contamination and often poor-quality data obtained from low input small RNA-seq samples such as those obtained from extra-cellular preparations.

modENCODE Consortium \cite{21177976}. The goal of the modENCODE project is to provide the biological research community with a comprehensive encyclopedia of genomic functional elements in the model organisms *C. elegans* and *D. melanogaster*. We have developed and evaluated methods for uniform high-throughput sequencing data processing such as RNA-Seq and ChIp-Seq generated by modENCODE project. As part of modENCODE project, we integrated RNA-seq and ChIP-seq binding data to reconstruct different regulatory networks to understand multi-level regulation in higher eukaryotes \cite{21324173}. We also developed a machine learning method, incRNA, to identify and characterize novel ncRNAs in C. elegans by integrating conservation, secondary structure, and expression data \cite{21177971}.

BrainSpan Consortium \cite{24695229}. In collaboration with Prof. Nenad Sestan's and Flora Vaccarino's group at Yale, together with groups at USC, the Allen Brain Institute and elsewhere, we analyzed large amounts of RNA-seq data to characterize the transcriptome of the human

brain during development. The aim of this project is to create a comprehensive map of gene expression and to understand how the human brain changes throughout life. We have already developed RSEQtools \cite{21134889}, a suite of tools that performs common tasks on RNA-seq data such as calculating gene expression values, generating signal tracks of mapped reads, and segmenting that signal into actively transcribed regions.

TCGA Cancer Genomics \cite{21307934}\cite{28358873}{26536169}. We also participated in TCGA PRAD (prostate cancer) and KICH (kidney cancer) projects. One particular focus of ours was noncoding sequence variants. Our software projects included FunSeq \cite{24092746}{25273974}, a tool for functionally annotating regulatory variants in cancer genome sequences, and LARVA \cite{26304545}, a tool for detecting significant mutation burdens in noncoding elements in cancer whole genomes.

Northeast Structural Genomics Consortium \cite{18487680}. We participated in the Northeast Structural Genomics (NESG) consortium several years ago. The NESG employs both X-ray crystallography and NMR spectroscopy tovel structural information useful in modeling thousands protein domains.

3.2 Proposed Research

Normalize processed datasets and deconvolute multi-omics profiles
The molecular profiles obtained by RNA-Seq, ChIP-Seq, and other omics primary data processing pipelines will be normalized and registered between time points and between individuals. Normalization is critical in order to identify differential biomarkers of diseases or phenotypes. We will also evaluate existing tools for differential "omic" analysis \cite{25516281}{19910308} as well as develop new methods if necessary in order to identify the molecular biomarkers that show significant differences between diseases and normal conditions.

One of the main analysis problems will be to develop methods to deal with longitudinal time course in multi-omics datasets. Toward this end, we will normalize omics data from several experiments individually, and then account for uneven sampling and time gaps using a Lomb-Scargle periodogram \cite{22424236}{16303799}{10643760}. Each periodogram will then be available for standard time-series analysis and data clustering such as the hierarchical clustering used to obtain common trends and assess biological relevance using such tools as Gene Ontology, Reactome, KEGG and WikiPathways for pathway analysis \cite{22424236}{21177976}{12140549}. This framework will normalize and compare many different types of omics datasets. To identify specific effects within massive quantities of longitudinal data we will develop tools that use bootstrap simulations to assess power and significance, taking into account the auto-correlated behavior of the data-points and periodogram analyses described above, where the number of datapoints can be leveraged to reduce the prediction error at each individual point.

To identify both intracellular and tissue composition changes under disease conditions, we will apply the Epigenomic Deconvolution method, which utilizes lists of loci exhibiting variation in CpG methylation levels across constituent cell types compiled from reference methylomes produced by the NIH Roadmap Epigenomics project \cite{25693563} and from a growing multitude of array-based profiles in NCBI GEO and other public archives. Starting from methylation profiles of tissue homogenates we will estimate both cell type proportions and methylation profiles of constituent cell types. The proportion estimates will then be used as a "key" to deconvolute gene expression and other "omic" profiles of constituent cell types.

Analyze and cluster gene activity data to identify coordinated modules
We will analyze both single-perturbation and temporal dynamic patterns from longitudinal time-course expression data and identify expression patterns associated with diseases or phenotypes and their regulatory mechanisms. In particular, we will construct the gene co-expression networks and find modules (with associated expression signatures) enriched in diseases or phenotypes. Finally, we will identify gene regulatory logics driving diseases or phenotypes via Loregic \cite{25091629}. Using ENCODE and other publicly available molecular profiles such as ChIP-seq data we will construct the regulatory networks for biomarker genes.

Build integrative models and identify biomarkers for diseases or phenotypes
To create a molecular map of the samples, we will integrate transcriptomic and epigenetic data with large datasets from other consortia. Specifically, we will incorporate the ENCODE TF data, the GTEx tissue-specific profiles, and the epigenetic marks of transcriptional regulatory elements from the Epigenome Roadmap. From this dataset we will construct integrative models relating epigenetics and transcriptomics using our previously developed machine learning approaches. Briefly, combined sets of genomic features in small (100bp) bins will be correlated with expression values over those regions. We will then generate statistical models relating epigenetic marks, TF binding, and gene expression, and further extend these models to incorporate proteomic and metabolomic data. To build our integrated models, the proteomic and metabolomic data will be combined with pathway information, such as KEGG and WikiPathways \cite{17923450}. These pathways will be linked to transcriptomic data through their associated genes, using the same machine learning approaches to relate transcriptional activity to protein and metabolite abundances. Thus, we can integrate metabolomic and proteomic with epigenetic and regulatory data. Finally, the large depth and coverage of transcriptomic experiments will be leveraged to develop integrative models of diseases or phenotypes (e.g. increased metabolite or protein quantities). Using the deconvoluted data, these models will be scaled to determine the diseases or phenotypes at the cellular, tissue, and organ levels.

Integrate genotype with genomics data to prioritize the variants and genomic regions
We will integrate processed functional genomic data with the genetic information from the same individuals in order to identify eQTLs, and will extend this strategy to identify metabolomic variants (mQTLs) and variants associated with proteomic changes (pQTLs). To characterize the rare variants within the individuals studied we will perform burden tests with LARVA to identify

genomic regions that are over or under represented in terms of the number of rare variants. Furthermore, we will perform an allelic analysis (using AlleleSeq) of available functional genomic data to identify allelic heterozygous variants. Finally, we will use FunSeq in order to integrate rare variants and associated functional genomic data to rank those that are most likely to be significant for diseases or phenotypes.

==> Julie END}}

Figures

Aim 1 figures

Aim 2 figures

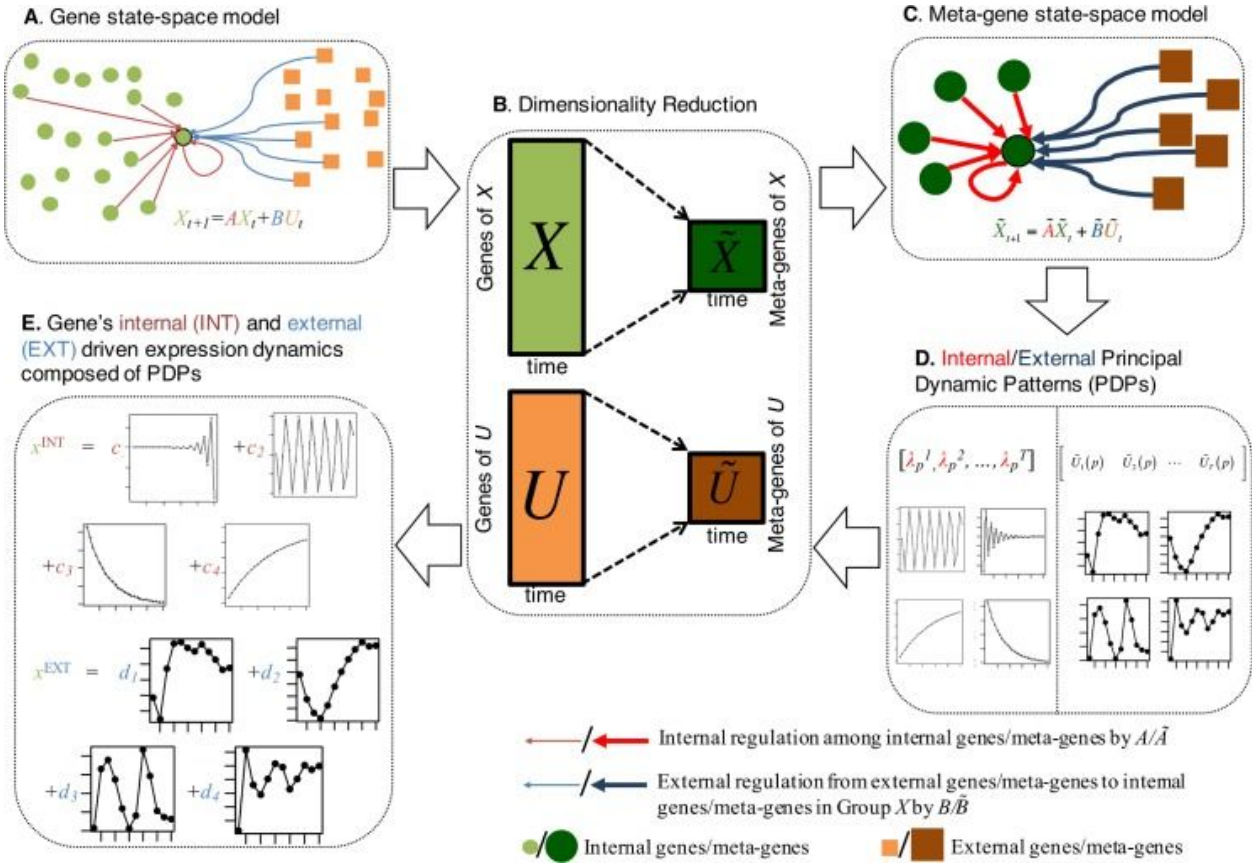


Figure 1 DREISS workflow.

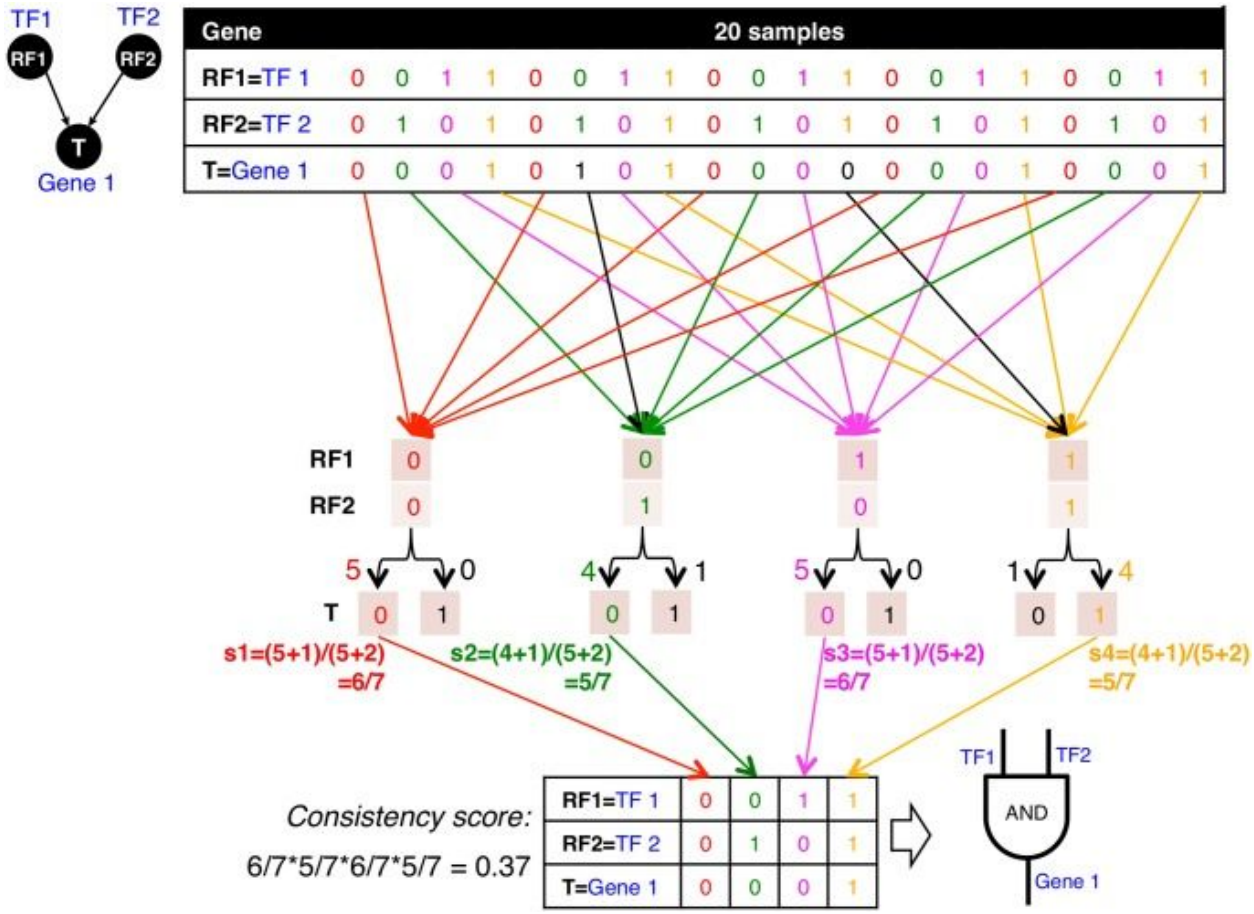


Figure 2 Procedures for mapping logic gates and calculating consistency scores.

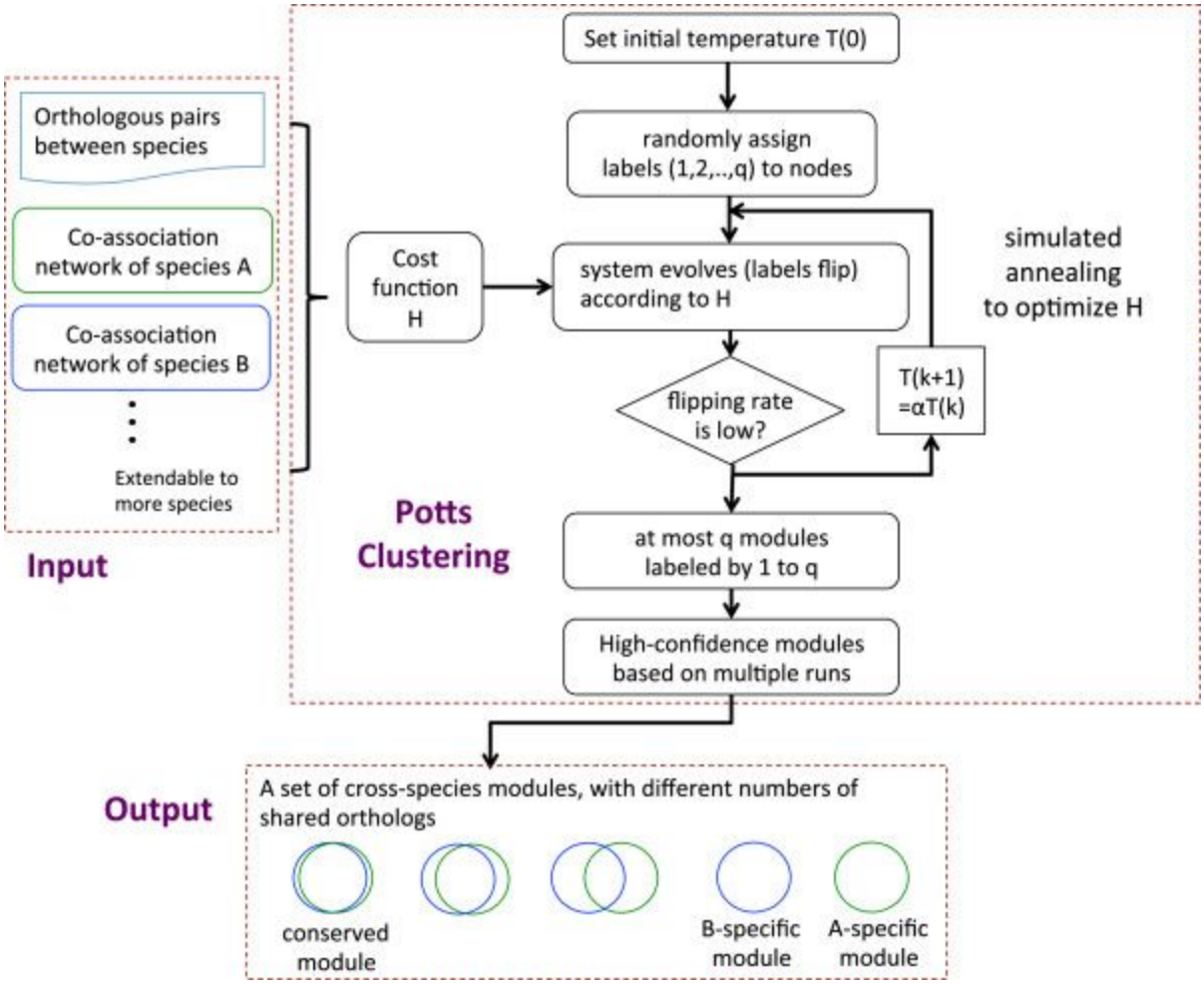


Figure 3 Outline of OrthoClust.

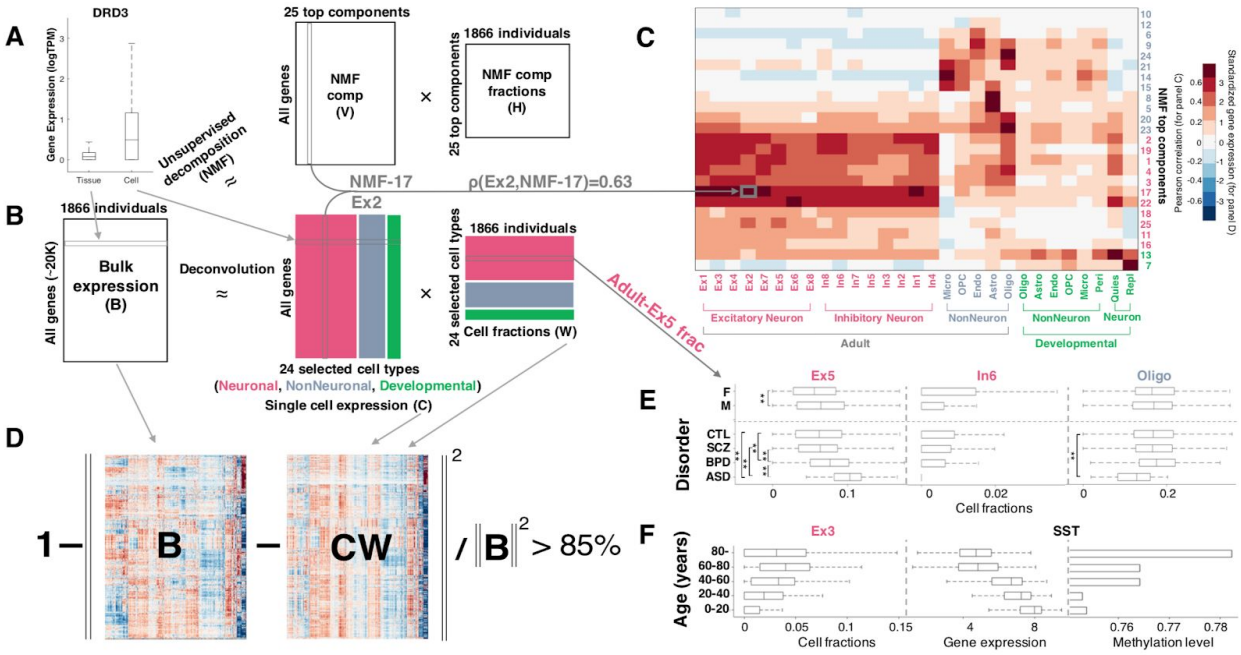
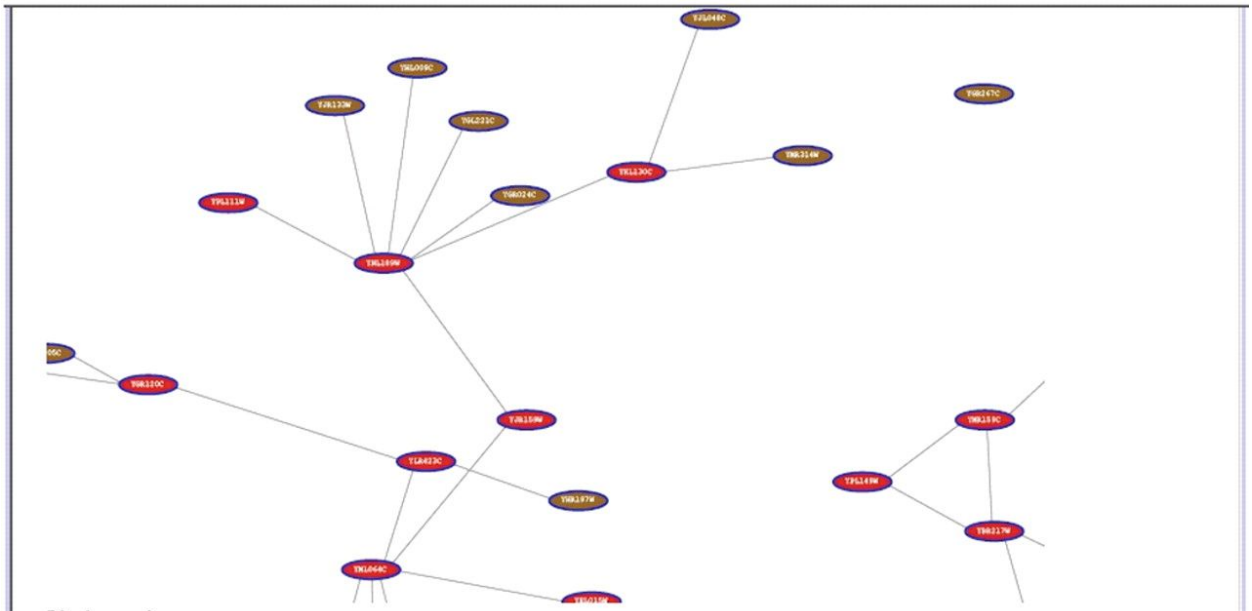


Figure 4

Deconvolution analysis of bulk and single-cell transcriptomics reveals cell fraction changes across the population.



Display options:

Default colors:

Node: blue Edge: lightgrey Text: white

Special coloring:

None

Color gradient: Degree of Original network from green to red

Color class: Class name: white

Redraw

Statistics:

Category	Node Counts	Edge Count	Connected Components	Degrees				Clustering Coefficients				Eccentricities				Betweenness			
				Avg.	S.D.	Min.	Max.	Avg.	S.D.	Min.	Max.	Avg.	S.D.	Min.	Max.	Avg.	S.D.	Min.	Max.
Whole network	276	187	109	1.30	0.74	1	7	0.04	0.19	0.00	1.00	2.51	1.57	1	9	3.60	20.22	0.00	200.00

Figure 5 Network visualization by tYNA

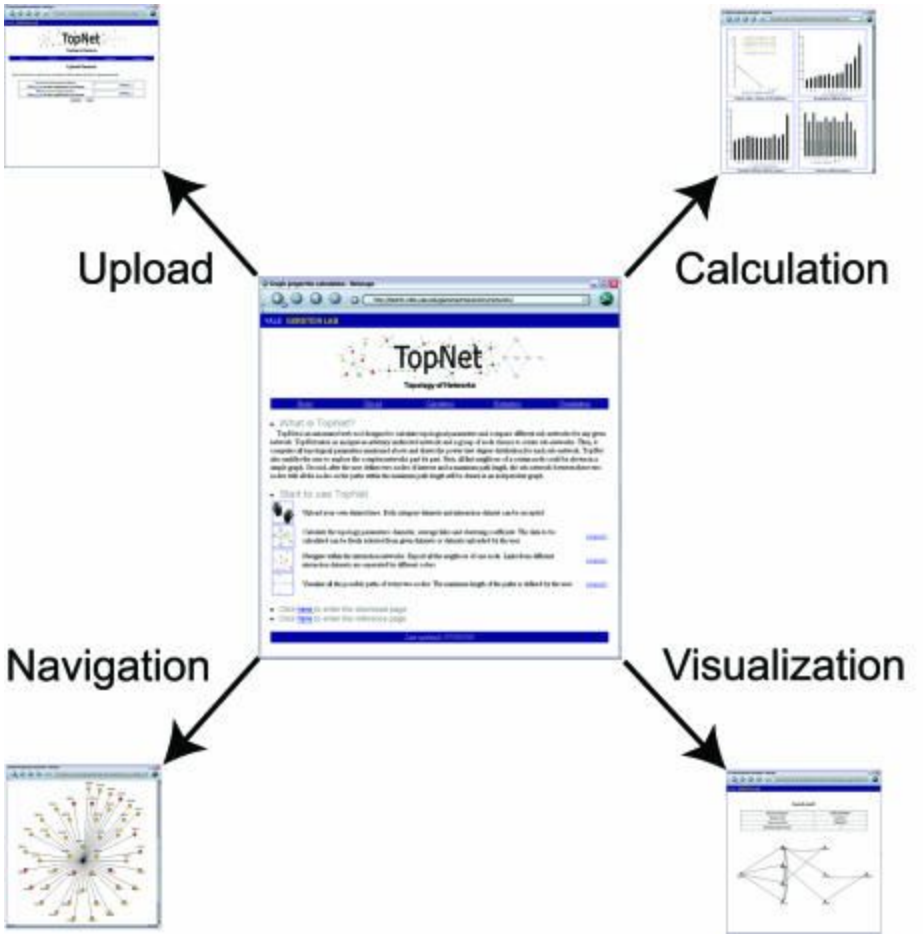


Figure 6 Overview of TopNet.

Aim 3 figures

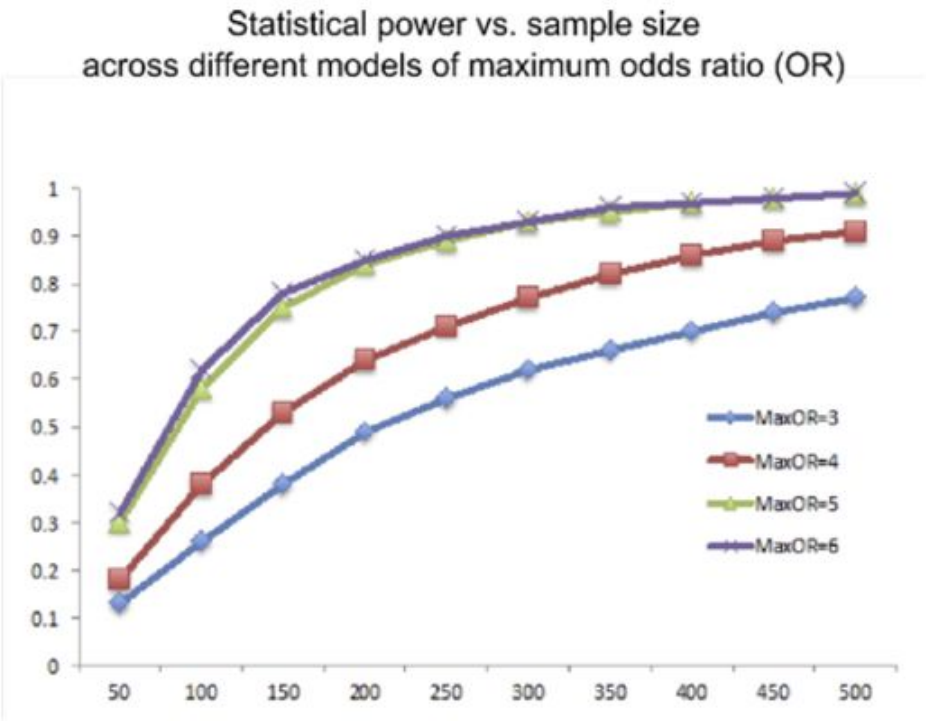


Figure 1. Using the default haplotype information in the SKAT haplotypes dataset, we randomly selected subregions of size = 5k and ran 100 simutations. We show the statistical power obtained across the different models of maximum Odds Ratio.

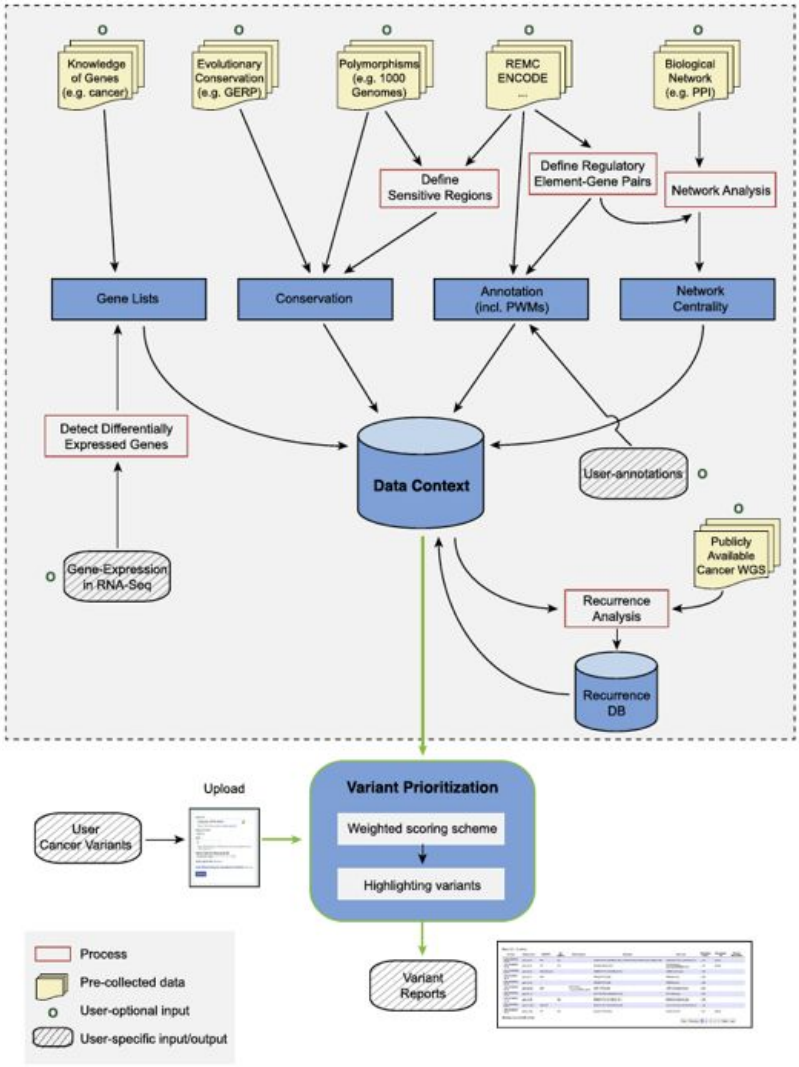


Figure 2. The workflow of FunSeq.

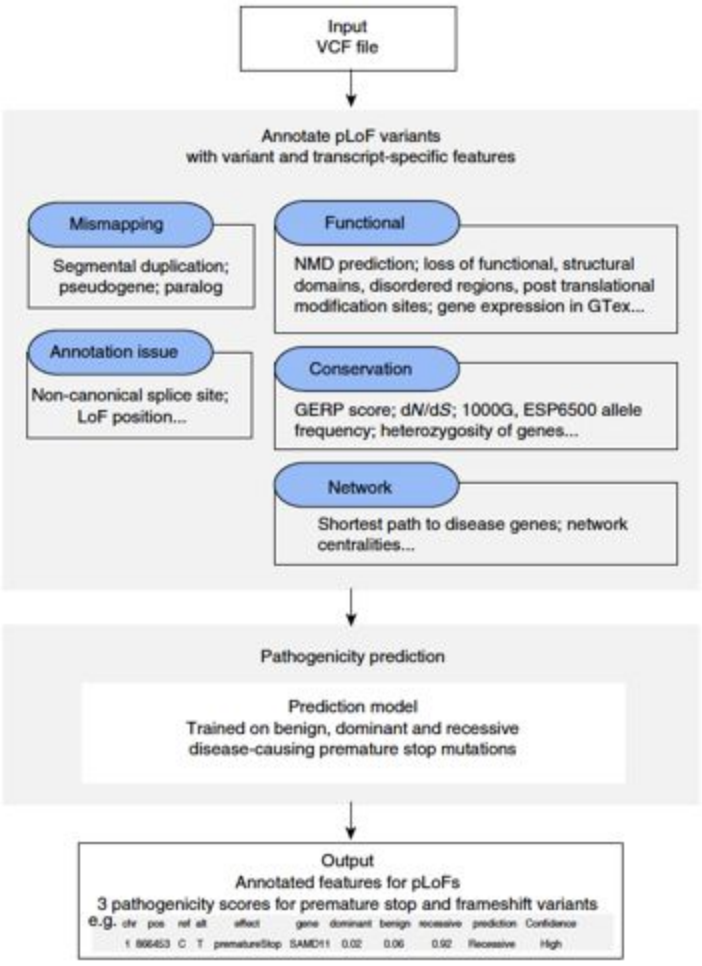


Figure 3. The workflow of ALoFT.

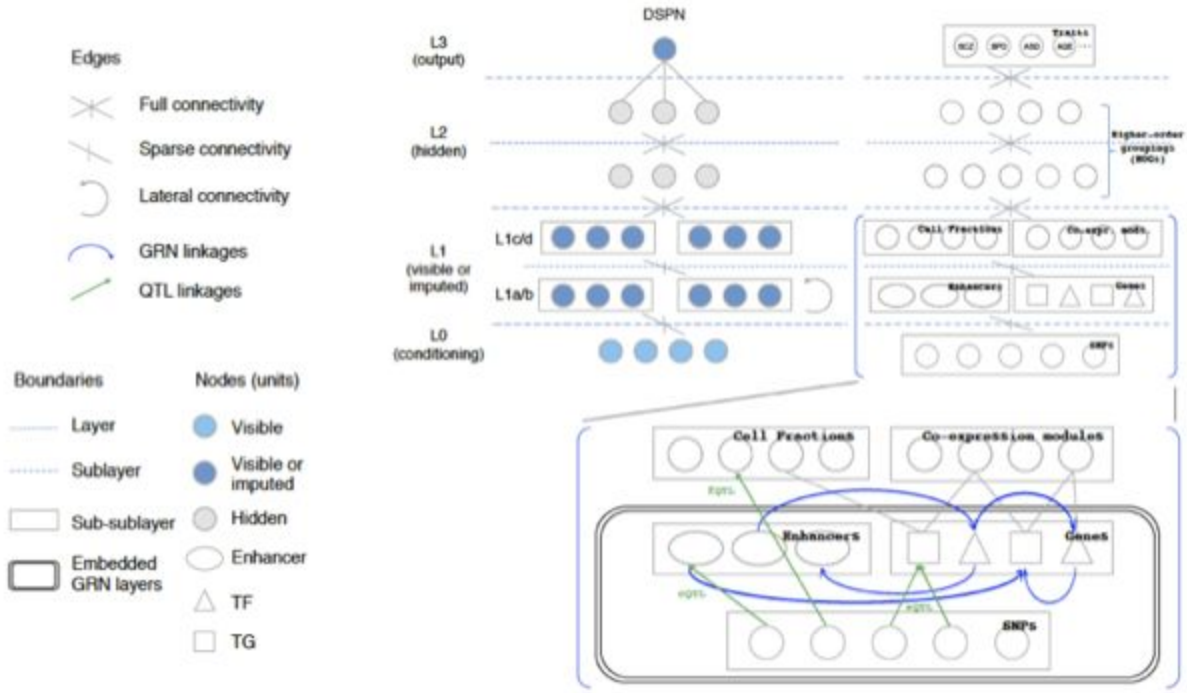


Figure 4. DSPN (Deep Structured Phenotype Network) deep-learning model links genetic variation to psychiatric disorders and other traits.

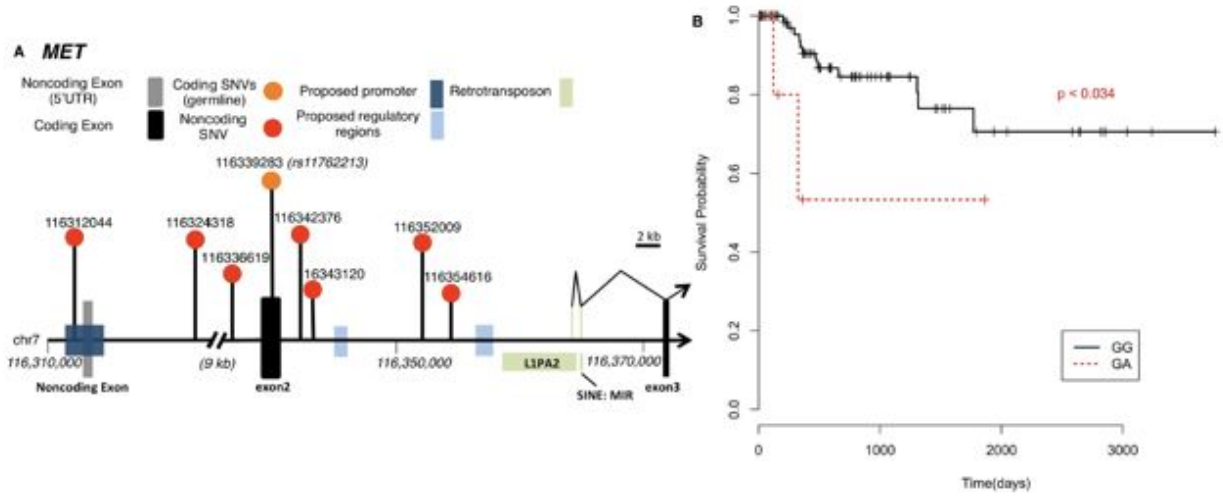


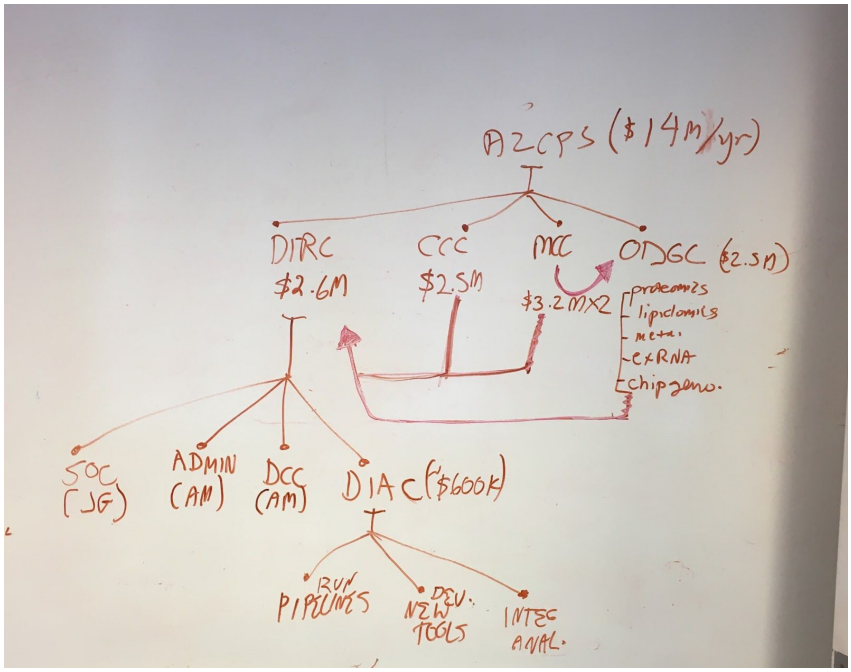
Figure 5. A. Whole genome analysis of 35 pRCC samples finds significant non-coding mutations in MET. B. A germline SNP (rs11762213) predicts survival in type 2 pRCC patients.

DIRC SOC

%%% Contribute 1 page to the SOC on impact analysis [BL, double recycle, by Sun.]

Assess the impact of the consortium and facilitate the dissemination of exRNA knowledge through text mining

The DIAC will build upon the success we have had with our evaluations of the impact of large scientific consortia and will continue to analyze the patterns of dissemination of knowledge about exRNA within the consortium and across the consortium into external communities. We will construct co-authorship networks from temporal data available using PubNet and use the diffusion base model we developed in Yan et al. \cite{21603617} to measure how quickly information about ex-RNA diffuses out of the consortium. As there will be many different ways for a scientific discovery to be exposed to the community, the impact of a paper would not be able to merely quantified by the number of citations. In addition to number of citations, we will also collect and analyze statistics such as the number of HTML views, the number of PDF and XML downloads, blog coverage and social bookmarking about papers authored by the consortium. In particular, we will look at article Altmetrics data, such as attention score, number of times each consortium publication is mentioned by twitter users, the geographic breakdown and demographic breakdown of the readers of consortium publications. The distributions of readers by professional status (e.g. Bachelor, Master, Doctor, etc.) and by discipline (e.g. Biology, Genetics, Computer Science, etc.). We will also use text mining to identify high-frequency terminologies about exRNA and collaborate with the SOC to standardize the semantics of those terminologies to facilitate better scientific communications within the consortium as well as external communities. Importantly, we will use text mining to construct a database about exRNA-disease relationships and collaborate with the SOC to make such knowledge easily accessible to the consortium participants as well as external researchers.



Ref

Guyon, I., Weston, J., Barnhill, S., and Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. Machine Learning *46*, 389-422.

Saeys, Y., Abeel, T., and de Peer, Y.V. (2008). Robust Feature Selection Using Ensemble Feature Selection Techniques. Machine Learning and Knowledge Discovery in Databases, Part Ii, Proceedings *5212*, 313-+.

Metabolomics Section References

1. Sud M, Fahy E, Cotter D, Azam K, Vadivelu I, Burant C, et al. Metabolomics Workbench: An international repository for metabolomics data and metadata, metabolite standards, protocols, tutorials and training, and analysis tools. Nucleic Acids Res. 2016;44: D463–70.

2. Tautenhahn R, Patti GJ, Rinehart D, Siuzdak G. XCMS Online: A Web-Based Platform to Process Untargeted Metabolomic Data. Anal Chem. 2012;84: 5035–5039.

3. Narasimhan R, Coras R, Rosenthal SB, Sweeney SR, Lodi A, Tiziani S, et al. Serum metabolomic profiling predicts synovial gene expression in rheumatoid arthritis. Arthritis Res Ther. 2018;20: 164.

4. Domingo-Almenara X, Montenegro-Burke JR, Ivanisevic J, Thomas A, Sidibé J, Teav T, et al. XCMS-MRM and METLIN-MRM: a cloud library and public resource for targeted analysis of small molecules. Nat Methods. 2018;15: 681–684.

5. Domingo-Almenara X, Montenegro-Burke JR, Benton HP, Siuzdak G. Annotation: A Computational Solution for Streamlining Metabolomics Analysis. Anal Chem. 2018;90: 480–489.

6. Li S, Park Y, Duraisingham S, Strobel FH, Khan N, Soltow QA, et al. Predicting network activity from high throughput metabolomics. PLoS Comput Biol. 2013;9: e1003123.

7. Huan T, Palermo A, Ivanisevic J, Rinehart D, Edler D, Phommavongsay T, et al. Autonomous Multimodal Metabolomics Data Integration for Comprehensive Pathway Analysis and Systems Biology. Anal Chem. 2018;90: 8396–8403.

Lipidomics Section References

1. Quehenberger, O., and E. A. Dennis. 2011. The human plasma lipidome. *N. Engl. J. Med.* **365**: 1812-1823.
2. Quehenberger, O., A. M. Armando, A. H. Brown, S. B. Milne, D. S. Myers, A. H. Merrill, S. Bandyopadhyay, K. N. Jones, S. Kelly, R. L. Shaner, C. M. Sullards, E. Wang, R. C. Murphy, R. M. Barkley, T. J. Leiker, C. R. Raetz, Z. Guan, G. M. Laird, D. A. Six, D. W. Russell, J. G. McDonald, S. Subramaniam, E. Fahy, and E. A. Dennis. 2010. Lipidomics reveals a remarkable diversity of lipids in human plasma. *J. Lipid Res.* **51**: 3299-3305.
3. Dennis, E. A., R. A. Deems, R. Harkewicz, O. Quehenberger, H. A. Brown, S. B. Milne, D. S. Myers, C. K. Glass, G. T. Hardiman, D. Reichart, A. H. Merrill, M. C. Sullards, E. Wang, R. C. Murphy, C. R. Raetz, T. Garrett, Z. Guan, A. C. Ryan, D. W. Russell, J. G. McDonald, B. M. Thompson, W. A. Shaw, M. Sud, Y. Zhao, S. Gupta, M. R. Maurya, E. Fahy, and S. Subramaniam. 2010. A Mouse Macrophage Lipidome. *J. Biol. Chem.* **285**: 39976-39985.
4. Gorden, D. L., D. S. Myers, P. T. Ivanova, E. Fahy, M. R. Maurya, S. Gupta, J. Min, N. J. Spann, J. G. McDonald, S. L. Kelly, J. Duan, M. C. Sullards, T. J. Leiker, R. M. Barkley, O. Quehenberger, A. M. Armando, S. B. Milne, T. P. Mathews, M. D. Armstrong, C. Li, W. V. Melvin, R. H. Clements, M. K. Washington, A. M. Mendonsa, J. L. Witztum, Z. Guan, C. K. Glass, R. C. Murphy, E. A. Dennis, A. H. Merrill, Jr., D. W.

Russell, S. Subramaniam, and H. A. Brown. 2015. Biomarkers of NAFLD progression: a lipidomics approach to an epidemic. *J. Lipid Res.* **56**: 722-736.

5. Dumlao, D. S., M. W. Buczynski, P. C. Norris, R. Harkewicz, and E. A. Dennis. 2011. High-throughput lipidomic analysis of fatty acid derived eicosanoids and N-acylethanolamines. *Biochim. Biophys. Acta*. **1811**: 724-736.
6. Wang, Y., A. M. Armando, O. Quehenberger, C. Yan, and E. A. Dennis. 2014. Comprehensive ultra-performance liquid chromatographic separation and mass spectrometric analysis of eicosanoid metabolites in human samples. *J. Chromatogr. A*. **1359**: 60-69.
7. Dennis, E. A., and P. C. Norris. 2015. Eicosanoid storm in infection and inflammation. *Nat Rev Immunol*. **15**: 511-523.
8. Quehenberger, O., S. Dahlberg-Wright, J. Jiang, A. M. Armando, and E. A. Dennis. 2018. Quantitative determination of esterified eicosanoids and related oxygenated metabolites after base hydrolysis. *J. Lipid Res.*: in press.
9. Gregus, A. M., S. Doolen, D. S. Dumlao, M. W. Buczynski, T. Takasusuki, B. L. Fitzsimmons, X. Y. Hua, B. K. Taylor, E. A. Dennis, and T. L. Yaksh. 2012. Spinal 12-lipoxygenase-derived hepoxilin A3 contributes to inflammatory hyperalgesia via activation of TRPV1 and TRPA1 receptors. *Proc. Natl. Acad. Sci. U.S.A.* **109**: 6721-6726.
10. Gregus, A. M., M. W. Buczynski, D. S. Dumlao, P. C. Norris, G. Rai, A. Simeonov, D. J. Maloney, A. Jadhav, Q. Xu, S. C. Wei, B. L. Fitzsimmons, E. A. Dennis, and T. L. Yaksh. 2018. Inhibition of Spinal 15-LOX-1 Attenuates TLR4-Dependent, NSAID-Unresponsive Hyperalgesia in Male Rats. *Pain*.
11. Loomba, R., O. Quehenberger, A. Armando, and E. A. Dennis. 2015. Polyunsaturated fatty acid metabolites as novel lipidomic biomarkers for noninvasive diagnosis of nonalcoholic steatohepatitis. *J. Lipid Res.* **56**: 185-192.
12. Dumlao, D. S., A. M. Cunningham, L. E. Wax, P. C. Norris, J. H. Hanks, R. Halpin, K. M. Lett, V. A. Blaho, W. J. Mitchell, K. L. Fritsche, E. A. Dennis, and C. R. Brown. 2012. Dietary Fish Oil Substitution Alters the Eicosanoid Profile in Ankle Joints of Mice during Lyme Infection. *J. Nutr.* **142**: 1582-1589.
13. Tam, V. C., O. Quehenberger, C. M. Oshansky, R. Suen, A. M. Armando, P. M. Treuting, P. G. Thomas, E. A. Dennis, and A. Aderem. 2013. Lipidomic profiling of influenza infection identifies mediators that induce and resolve inflammation. *Cell*. **154**: 213-227.
14. Fahy, E., S. Subramaniam, H. A. Brown, C. K. Glass, A. H. Merrill, Jr., R. C. Murphy, C. R. Raetz, D. W. Russell, Y. Seyama, W. Shaw, T. Shimizu, F. Spener, G. van Meer, M. S. VanNieuwenhze, S. H. White, J. L. Witztum, and E. A. Dennis. 2005. A comprehensive classification system for lipids. *J. Lipid Res.* **46**: 839-861.
15. Fahy, E., S. Subramaniam, R. C. Murphy, M. Nishijima, C. R. Raetz, T. Shimizu, F. Spener, G. van Meer, M. J. Wakelam, and E. A. Dennis. 2009. Update of the LIPID MAPS comprehensive classification system for lipids. *J. Lipid Res.* **50**: S9-14.

16. Nguyen, A., Rudge, S., Zhang, Q., and MJO Wakelam. 2017. Using lipidomics analysis to determine signaling and metabolic changes in cells. *Current Opinion in Biotechnology.* **43**z; 96-103.
17. Kopczynski, D., Coman, C., Zahedi, R., Lorenz, K., Sickmann, A., and R. Ahrends. 2017. Multi-OMICS: a critical technical perspective on integrative lipidomics approaches. *Biochimica et Biophysica Acta – Molecular and Cell Biology of Lipids.* **1862(8)**: 808-811.