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

Timeline

JG to archive the doc before and the after the call
Give them files and send to the minutes thread

** FINISH
* By tmrw at 9 am (Sat), YY, TXL, JG, PE need to do their cuts
Resolve comments
* YY to finish on Sat. at 9 am

** Julie send off Sat. at noon - LS w/ a deadline Mon. at 10 am
Use suggest mode

** TG to go through frin 9 to 10 & thenb TG to send email

** 12 Fri - Alex gdoc (or Sat?)
PE to send Alex to send a cleaned doc - basically the jUlie + figures
PE also send Alan

* after back from Julie
Tue at 6 am finishefs polishing and carefully going through the doc
* JR, MG, PE read the doc for english on Tue & Tue night
Meet on Wed morning

* Check in call on F at 4 pm
PE to do word counts

* 15 Mon. - hard done word doc (they'll do references)
MG is on the call

17 Wed. (late) [meet on Wed. morning, be prepared for doing fixes on Wed. day]
Due at NIH 10/24

Word Count (Fri afternoon before call)

Specific Aims = 635
Intro = 823
Innovation = 157
Aim 1 = 2046
Aim 2 = 3212
Aim 3 = 3188 (?)
Total = (9400) - 7500 is 10pg
Curr ~ 12.5 pg (without figures)

Specific Aims = 635
Intro = 750
Innovation = 157
Aim 1 = 1850 [cutting 200]
Aim 2 = 2637 [cutting 269 from imaging, cutting “Network visualization” to 350]

Aim 3 = 2800+133 [cutting 50 overview, 44, 47, 100, 65, 50]

Total = (8500) - 7500 is 10pg
Curr ~ 12.5 pg (without figures)

{{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. In the first place, 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 (e.g. 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 different data types from the output of the pipelines from aim 1 and determine initial candidate signatures of the transition from acute to chronic pain. 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 types with neuroimaging data. This will enable discovery of brain-wide data-driven markers of chronic pain signatures, which can in turn be merged with other -omics datasets. 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. These tools will be constructed collaboratively with members of the consortium based on specific priorities as directed by the Analysis Working Group (AWG) (see aim 3). Such tools will enable the identification of signatures and potential biomarkers that distinguish acute from chronic pain individuals from the cohorts under investigation. We will also develop tools to integrate -omics data with available electronic health records (EHR) data.

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, both in more basic science (e.g. ENCODE and 1000 Genomes Project) and disease-oriented contexts, particularly in relation to psychiatric diseases (e.g. PsychENCODE). We anticipate that the integrative analyses will involve data integration of the large-scale omics data, imaging data and the HER data from the A2CPS Consortium, as well as connecting the Consortium's data with complementary data types from external sources, including genomic variation data, functional genomic data, and phenotypic characterization. Particularly, we will apply advanced machine learning techniques such as interpretable deep-learning model for acute to chronic pain signatures analysis. We will describe our large-scale experience in integrative analyses 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 to 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, 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), regardless of 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 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 chemokines are upregulated over a time scale of two weeks in peripheral tissues 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 RNAs detected outside the cellular context, especially body fluids,may be useful sources for non-invasive biological signatures. 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 related to nervous lesions \cite{25274330}. Altered miRNA profiles are also detected in the cerebrospinal fluids for patients with fibromyalgia, a disorder related to central sensitization \cite{24205312}.

Some plasma miRNAs also have altered expression levels for patients after treatment with opioids and may serve as prognostic markers \cite{4110167}. Generally, a systematic study of the significance of circulating RNAs in the development of chronic pain is still lacking. Larger-scale analysis of transcriptomics of RNAs from body fluids would enable identification of more novel non-invasive markers.

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}. Accumulation of high spatial and temporal resolution imaging data and incorporation of pattern recognition methods would help to identify neurological signatures\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.

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 by our Aggregation and Correlation Toolbox (ACT), a tool for comparing genomic signal tracks \cite{21349863}. We developed incRNA \cite{21177971} to predict novel ncRNAs

using known ncRNAs of various biotypes. We created FusionSeq to detect transcripts that arise due to trans-splicing or chromosomal translocations \cite{20964841}. We have also constructed IQSeq \cite{22238592}, which calculates the relative and absolute abundance of contributing transcript isoforms to a gene from RNA-seq data. We have developed AlleleSeq tool \cite{21811232} that combines diploid genomic information with RNA-Seq data to identify transcripts showing allele-specific expression. We have further developed Pseudo-seq \cite{22951037} and PseudoPipe \cite{25157146} to addresses the issue of quantification of pseudogene expression.

We recently developed the extracellular RNA processing toolkit, exceRpt (submitted to Cell Systems, available at 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. The exceRpt pipeline has been used for uniform processing of hundreds of RNA-Seq datasets submitted to the exRNA Atlas (http://exrna-atlas.org/) repository.

Consortium experience in transcriptomics analysis

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 \cite{22388286} and Tophat \cite{19289445}. We describe our consortium experience further in aim 3.

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 (UCSD) has recently processed and curated its 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%). The DRCC is actively accepting metabolomics data for small and large studies on cells, tissues and organisms via the Metabolomics Workbench, which 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. Studies are available for browsing, analysis and download (subject to embargo release) in the NIH Data Repository section of the website (http://www.metabolomicsworkbench.org/).

The UCSD Center for Computational Biology & Bioinformatics (CCBB) provides bioinformatics expertise to analyze large molecular datasets in the areas of genomics, systems biology and translational medicine. The 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 \cite{30075744} (**Aim 1 Figure 1**). CCBB has carried out a number of research and clinical studies that establish our expertise in the field of metabolomics, scientific ability as well as the capacity both technical and instrumental to successfully perform accurate and precise metabolomics measurements on a large scale and in high throughput settings.

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 \cite{22070478}. The complexity of the lipidome both in dynamic range and structural diversity represents a major analytical challenge. To address these challenges, the LIPID MAPS Consortium 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 (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 \cite{20671299}. Immunologically-activated macrophages were also profiled and over 500 discrete lipid species were measured and associated pathways were mapped, integrating transcriptomics, proteomics and lipidomics \cite{20923771}. Our laboratory now routinely profiles plasma, urine, bronchoalveolar lavages, cerebral spinal fluid and various other tissues of both human and animal origin for biomarker discovery 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 \cite{25598080}. Our lipidomics platform for monitoring over 200 oxidation and signal transduction consequences is the most developed platform to emerge in the metabolomics area \cite{21689782,25074422,26139350,Quehenberger et al 2018 (in press)}. 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 \cite{22493235,30130298}. 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 \cite{25404585}. Similar approaches were used to identify eicosanoid targets in various bacterial and viral infectious diseases including Lyme disease and influenza \cite{22695969,23827684}.

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.

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 \cite{30150755}. 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 \cite{29039932}.

Lipidomics: Primary Data Analysis

After the initial data acquisition from LC-MS under various acquisition modes, 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: 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.

The UC San Diego Center for Computational Biology & Bioinformatics (CCBB) will leverage the Metabolomics Workbench \cite{26467476} and XCMS \cite{22533540} for the metabolomics data, and the expertise of the LIPID MAPS Lipidomics core for the lipidomics, to develop open source, automated and reproducible primary and secondary analysis pipelines for the A2CPS DIRC. 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.

Aim 2) Building Tools

2.1 Preliminary Results

Tools for identification of transcriptome signatures

We will leverage our extensive experience processing and analyzing transcriptomic data in addressing the aims of the DIRC DIAC. We evaluated several independent computational methods and protocols 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.

Following transcriptomic data processing, 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) for analyzing the “Dynamics of gene expression driven by Regulatory networks, both External and Internal, based on State Space models” \cite{27760135}. DREISS employs dimensionality reduction to help identify canonical temporal dynamics (e.g., degradation, growth and oscillation) representing the regulatory effects emanating from various subsystems. Another such tool, 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 1**). Loregic is able to characterize complex circuits involving both transcription factors (TFs) and miRNAs. 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, and then outputs optimized cross-species modules, either conserved or species-specific. 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. In exRNA studies, deconvolution methods can help us identify fractions in the bulk expression data

and their 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. Given the number of desired components, the bulk tissue gene expression matrix X is decomposed into the product of two matrices: H represents NMF "top components" (NMF-TCs) and V represents the expression level of genes in the NMF-TCs. We found that NMF-TCs recovered the expression patterns of different cell types in bulk RNA-Seq data on brain cell population. We then applied a supervised approach that uses single-cell expression signatures to find the fractions of different cell types. We defined the sample gene expression matrix B, and fraction gene expression matrix iC. We used the non-negative least square method to find a non-negative matrix W as the linear combination coefficients. 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 2**).

Network Analysis and Visualization Tools

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. 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}. It computes a variety of topological parameters given the input network and specified subnetworks and calculates the power-law degree distribution for each sub-network. In addition, we developed TopNet-like Yale Network Analyzer (tYNA), a Web system for managing, comparing and mining multiple networks, which efficiently implements methods that are useful in network analysis \cite[17021160}.

We have also published many papers on construction of hierarchy structures for the regulatory network for both transcriptional and post-transcriptional regulation. We proposed the hierarchical score maximization (HSM) algorithm, which first defines a score to quantify the degree of hierarchy in a network, and then perform simulated annealing procedure to infer a hierarchical structure that maximizes the score \cite{4404648} (**Aim 2 Figure 3**). We applied our

algorithm to determine the hierarchical structure of the phosphorylome in detail. Using genome-wide binding locations of human, worm, and fly transcription-regulatory factors (RFs), we performed simulated annealing to reveal the organization of RFs in three layers of master-regulators, intermediate regulators, and low-level regulators \cite{4336544}. We organized the binding profiles of 119 TFs in 458 ChIP-Seq experiments from ENCODE into a hierarchy and integrated it with other genomic information (e.g. miRNA regulation), forming a dense meta-network. Factors at different levels have different properties: for instance, top-level TFs more strongly influence expression and middle-level ones co-regulate targets to mitigate information-flow bottlenecks \cite{4154057}.

Neuroimaging

We have recently quantified whole-brain global functional connectivity in 300 adults, in relation to behavioral measures of alcohol use and impulsivity, which may relate to chronic pain markers. We identified robust relationships between global connectivity in prefrontal motor planning areas and maximum lifetime drinks, which was fully mediated by self-reported impulsivity (**Figure 2**)[14]. These results replicate existing effects[15] and establish the viability of proposed methods to discover novel relationships between neural functional architecture and altered behavior.

All necessary pipelines for the processing and analysis of anatomical, BOLD functional connectivity, and structural connectivity data have been fully implemented at Yale. These analyses will be greatly facilitated by sophisticated processing protocols developed by the Human Connectome Project (HCP)[23-25]. We have implemented the HCP pre-processing and analysis pipeline on Yale's High Performance Computing Cluster. All analyses relating imaging and clinical measures will be performed in collaboration with the investigators of the DIRC.

We have developed the Multimodal Neuroimaging Analysis Platform (MNAP) suite of tools, which integrates several packages that support an extensible framework for data organization, preprocessing, quality assurance, and various analyses across neuroimaging modalities. The MNAP suite architecture is robust yet flexible and can be readily extended by adding functions developed by its core tools. It provides a high throughput 'batch' engine and seamless analytic integration with other widely-adopted community tools such as FSL, Connectome Workbench, HCP Pipelines, PALM, Octave/Matlab, AFNI, R Statistical Environment, FreeSurfer, and AFNI packages. Overall, the MNAP suite supports full 'turnkey' workflow, from imaging data upload to derived neuro-behavioral phenotypes (https://dev-mnap-tools-yale-edu.pantheonsite.io/)

In addition, 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 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 \cite{Saeys, Y., Abeel, T., de Peer, Y. V., Robust Feature Selection Using Ensemble Feature Selection Techniques, *Machine Learning and Knowledge Discovery in Databases, Part II*, Proceedings, 2008}, Principal Component Analysis \cite{18327243}, Support Vector Machine-Recursive Feature Elimination \cite{Guyon, I., Weston, J., Barnhill, S., Vapnik, V., Gene selection for cancer classification using support vector machines, *Machine Learning*, 46: 389-422}, 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.

We will also develop network analysis tools to 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}. We will construct the regulatory networks for biomarker genes using ENCODE and other publicly available molecular profiles such as ChIP-seq data.

We will conduct out analysis 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.

Metabolomics

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 \cite{22533540}. Dysregulated metabolic pathways will be identified using mummichog \cite{23861661} 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 \cite{29893550}.

Lipidomics

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 \cite{27816901}. To integrate lipidomics with other omics data, we will employ both cluster based and network based approaches \cite{28193460}.

Neuroimaging

Few studies have systematically studied multi-modal alterations in this condition or linked such impairments to its genetic risk, functional outcome, and individual differences in clinical outcome in chronic pain disorders. This neuroimaging component has the potential to map undiscovered neural alterations in pain disorders with unprecedented level of data integration quality. The approach is further strengthened by the use of the following key innovative tools in the analyses of the DIRC DIAC:

Seed-based Analyses Focused on Subcortical Reward Pathways. Seed-based approaches will closely follow our prior studies using subcortical anatomically-defined nuclei \cite{20498341}. The analysis starts using individual-specific, anatomically defined subcortical seeds focused on reward pathways (e.g. accumbens, see **Figure 2C-E**) to test if there is widespread reward-related 'connectomic' signature in chronic pain. Here we will use validated tools \cite{21415225,21193174} to examine brain-wide subcortical coupling. First, we compute a seed-based correlation map by extracting average time-series across all voxels in each subject's bilateral anatomically defined seed through FreeSurfer-based segmentation \cite{11832223,15501102} (or any other subcortical seed of interest). This signal is then covaried with each gray matter voxel. Then, the computed Pearson correlation values are transformed to Fisher Z values (Fz). This yields a subject-level map, where each voxel's value represents connectivity with the anatomically-defined subcortical seed.

***Network-level Analyses Based on Existing Parcellations*.** We will perform regionally constrained analyses based on well-established functional network segmentation, consisting of ~600 brain areas in 12 functional networks, derived from resting-state connectivity. This approach accomplishes a dimensional reduction, while reducing noise. This, however, reduces spatial resolution and could potentially mask disorder-specific alterations. Thus, we use this strategy in parallel with, rather than in place of, voxel-wise analysis. For functional connectivity analyses, we will average fluctuations in BOLD signal within an area for each network and compute co-variation among the resulting regional signals. For structural connectivity, we sum streamlines within each of the cortical areas, generating a 'parcellated connectome' for each subject, resulting in a data-reduced structural connectivity matrix on a standard cortical mesh. This will provide a reduced large-scale functional and structural connectivity matrix for each subject where appropriate data is available.

***Data-driven Analyses*.** Prior evidence implicates specific networks and regions in functional impairment in chronic pain. However, functional/structural dysconnectivity in chronic pain disorders, especially within cortical networks, may be highly variable, given clinical heterogeneity. We thus designed new neuroimaging techniques to identify dysconnectivity in a data-driven fashion, termed global brain connectivity (GBC) \cite{22980587,24314349,21496789}. GBC provides a measure of the average connectivity strength from one voxel/area to all other voxels/areas – thus producing an unbiased approach as to the location of dysconnectivity. Unlike seed-based approaches, GBC is sensitive to brain-wide perturbations (irrespective of target locations) in the functional/structural connectivity of an area. Further, GBC involves one statistical test per voxel rather than one test per voxel-to-voxel pairing, substantially reducing multiple comparisons (e.g., 30k vs. ~450 million). These improvements dramatically increase the chances of identifying group differences in connectivity, or individual differences correlated with symptoms \cite{22980587,21496789,19909818,22745498}. By extension, this approach can be applied to structural connectivity derived from DWI, either at the whole-brain level or within associative networks.

Group-level Analyses. To examine between-group differences or relationships with clinical assessments, Fz maps are entered into an independent samples t-test (or other appropriate 2^{nd}-level tests) (**Figure 2**). Whole-brain type I error correction is accomplished via threshold-free cluster enhancement (TFCE) non-parametric techniques implemented in FSL's *Randomise* tool \cite{18501637}.

Normalize processed datasets and deconvolute multi-omics profiles

The molecular profiles obtained by RNA-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.

Aim 3) Integrative Analysis [currently 3081 words]

3.1 Overview

In aim 3, we will try to do large-scale integrative analyses based on all the data types produced by the A2CPS Consortium. Firstly, we will integrate the large-scale omics data, imaging data, and the EHR data together, using a variety of machine learning techniques, such as deep learning models, to accurately predict the acute to chronic pain signatures. Then, we will also integrate the consortium data with complementary external data sources, including genomic variation data and other functional genomic data, to understand the regulation of these signatures.

We will try to do this large-scale data integration in a consortium framework, through forming an Analysis Working Group (AWG). We will also help lead a futility analysis in this group, to estimate the number of samples needed for detecting the signature effect. For this aim, we will

demonstrate our ability and experience carrying out this aim through our past work in large-scale genomics efforts, both in more basic science (e.g. ENCODE and 1000 Genomes Project) and disease-oriented contexts, particularly in relation to psychiatric diseases (e.g. PsychENCODE) in particular we will mention our past work on integrating large-scale genomic data with genomic variation data.

3.2 Preliminary Results

3.2a Experience leading consortium analyses for general genomics

3.2a1 Integrative analysis of consortium's wide datasets

We served in a variety of leadership roles for several large-scale national collaborations focused on general functional genomics and data science, including (mod)ENCODE Consortium, 1000 Genomes Project, extracellular RNA Consortium, KBase and Northeast Structural Genomics Consortium.

We played a lead role in the integrative analysis of multi-omic datasets from ENCODE Consortium \cite{22955616,22955619,22955620,25164755,25164757} and modENCODE Consortium \cite{25164755,21177976}. By integrating large-scale RNA-seq and ChIP-seq datasets from ENCODE, we have developed statistical models to quantify the relationship between gene expression and transcription factor binding and/or chromatin modification signatures \cite{21926158,22955978}. We have also developed a number of approaches for constructing and studying biological networks that can be applied to analyze ENCODE datasets. We integrated multiple genomic datasets to construct gene regulatory networks consisting of various regulatory factors including transcription factors and micro-RNAs and their target genes \cite{22955619,25164757,22125477}. For constructed gene regulatory networks, we developed methods to construct and analyze human and model organism gene regulatory networks \cite{20439753,21177976,22125477,21430782,22955619} using ENCODE and modENCODE datasets. We also analyzed hierarchical structures of gene regulatory networks and found that hierarchy rather than centrality ("hubiness") better reflects the importance of regulators \cite{22955619,17003135,20122235,20351254,20523742}.

We also helped lead the structural variation (SV) analysis for the 1000 Genomes Project \cite{20981092,26432245,24092746}. We developed an annotation pipeline that maps SNPs, indels and SVs on to protein coding genes \cite{28851873}. We also developed algorithms to identify indels and structural variations based on split-read, read-depth and paired-end mapping methods. Using the datasets from 1000 Genomes Project, we studied the distinct features of SVs originating from different mechanisms \cite{28662076}. We performed SV mechanism annotations for the 1000 Genomes Project Phase 3 deletions using BreakSeq \cite{20037582}, categorizing 29,774 deletions by their creation mechanisms.

We are an integral part of the extracellular RNA (exRNA) Consortium \cite{27112789}{27076901}, a large-scale collaboration project aimed at establishing dada standards, a data portal, and tools and reagents to the scientific community. We performed integrative analysis of consortium data and provide support to the broader consortium \cite{Cell Sys}. 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.

We also participated in the DOE KBase (The United States Department of Energy Systems Biology Knowledgebase) \cite{29979655}, which is an open-source software and data platform that enables data sharing, integration, and analysis of microbes, plants, and their communities; and Northeast Structural Genomics Consortium \cite{18487680}, which employs both X-ray crystallography and NMR spectroscopy to provide novel structural information useful in modeling thousands protein domains.

3.2a2 Integrative analysis of omics datasets with 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 (**Aim 3 Figure 1**). 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. Recently, we developed RADAR by extending the FunSeq variant prioritization framework to the RNA transcript level \cite{Genome Biology in press}. RADAR integrates the ENCODE eCLIP datasets, Bind-n-Seq datasets and RBP KD RNA-seq datasets to reconstruct a comprehensive post-transcriptional network. By combining other genomic information including conservation and motif features, RADAR could pinpoint deleterious variants, such as splicing-disruptive ones, which may be missed by other methods. 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.

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 \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}.

3.2b Experience leading consortium analyses for disease genomics

3.2b1 Brain diseases

It is clear that the sensory, process and modulation of acute and/or chronic pain is involved in a distributed network in the brain, especially for the pain-relevant brain regions. Pain can affect mood, sleep, memory and concentration, and has association with various psychiatric diseases \cite{17087832,28146315}. Considering our extensive experience in neurogenomics and psychiatric diseases, our work is very relevant to the A2CPS Consortium.

We played a lead role in the data analysis for the PsychENCODE Consortium \cite{26605881,29439242}, a project aimed at understanding regulatory variants in the context of their functional connections to psychiatric disorders, with several papers currently in the revision stage \cite{capstone4,capstone1,capstonedevelopment}. In our recent work, we identified functional elements, multiple QTLs and regulatory-network linkages specific to the adult brain by integrating data from the PsychENCODE Consortium together with relevant external data sources from ENCODE, CommonMind, GTEx, and Roadmap \cite{capstone4}. In addition to the adult brain, we also assessed the degree of chromatin differences between developmental stages relative to that between tissues. Furthermore, we used the regulatory network based on Hi-C, QTLs, and activity relationships to connect noncoding GWAS loci to potential psychiatric disease genes including schizophrenia, autism, bipolar and Alzheimer's disease. We also participate in the BrainSpan Consortium, which aims to create a comprehensive map of gene expression and to understand how the human brain changes throughout life. In collaboration with Prof. Nenad Sestan'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 \cite{24695229}. 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.

Particularly, we developed an integrated and interpretable deep-learning model, Deep Structured Phenotype Network (DSPN), that could predict psychiatric disorder phenotypes using genotype and functional genomic elements \cite{capstone4}. The model combines a Deep Boltzmann Machine (DBM) architecture \cite{R. Salakhutdinov, G. Hinton, Deep Boltzmann Machines. Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics (2009)} with conditional and lateral connections derived from a gene regulatory network. Traditional classification methods such as logistic regression predict phenotype directly from genotype, without using intermediates such as the transcriptome. In contrast, the DSPN is constructed via a series of intermediate models that add layers of structure. We included layers for intermediate molecular phenotypes associated with specific genes (i.e., their gene expression and chromatin state) and pre-defined gene groupings (cell-type marker genes and co-expression modules), multiple higher layers for inferred groupings (hidden nodes), and a top

layer for observed traits (psychiatric disorders and other brain phenotypes). Finally, we used sparse inter- and intra-level connectivity to integrate our knowledge of QTLs, regulatory networks, and co-expression modules from the sections above. By using a generative architecture, we ensure that the model is able to impute intermediate phenotypes, as well as provide forward predictions from genotypes to traits.

3.2b2 Cancer

In addition to neurogenomics and psychiatric diseases, we also played a lead role in the data analysis for the Pan-Cancer Analysis Working Group (PCAWG) Consortium \cite{https://www.biorxiv.org/content/early/2017/07/12/162784,https://www.biorxiv.org/content/early/2018/09/07/179705}, and participated in TCGA PRAD (prostate cancer) and KICH (kidney cancer) projects \cite{21307934,28358873,26536169}. PCAWG Consortium represents 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. We leveraged our expertise in non-coding regions in the first whole-genome analysis of TCGA kidney cancer (KIRP) samples, in which we found significant genomic noncoding alterations beyond traditional known drivers of KIRP located within coding exons \cite{28358873}.

We also developed a variety of tools for integrative analysis of cancer genomics data. 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}. 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{ENCODEC 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.

3.3 Proposed Research

3.3a Developing signature from Integrative analysis of the consortium data

3.3a1 Formation of a AWG to Coordinate Integrated Analysis

The DIAC will help organize and coordinate the formation of an analysis working group (AWG) in order to lead the data analysis efforts of the consortium. All members of the consortium will be invited to participate in the regular (frequency to be determined) AWG conference calls. Data analysis efforts by members of the DIAC as well as other members of consortium will be encouraged to be presented on these AWG calls and this will facilitate integrative analysis of consortium wide data as well as provide a forum for discussion of these integrative analysis

efforts. The AWG will be responsive the steering committee of the A2CPS Consortium and will routinely report on the ongoing analysis efforts.

3.3a2i General Support for Consortium Activities

In addition to the integrative analysis efforts the DIAC will also provide more general analysis support for members of the A2CPS Consortium. Members of the DIAC will collaborate with members of the consortium on more focused analyses. The DIAC will support the DCC a in support of the pipelines and tools developed by the DIAC and will support the SOC in outreach efforts to publicize the analysis pipelines, tools and analysis work products developed and performed by the DIAC. In addition to coordinating the AWG, the DIAC will participate in other A2CPS Consortium efforts such as conference calls and in person meetings in order to further support the analysis needs of the consortium.

3.3a3 Using advanced deep learning models for the signatures analysis

To model the complex interactions between genotype and acute to chronic pain signatures, we will develop an interpretable deep-learning framework called the Deep Structured Phenotype Network (DSPN). Briefly, the DSPN model combines a Deep Boltzmann Machine (DBM) architecture with conditional and lateral connections derived from a gene regulatory network (\cite{R. Salakhutdinov, G. Hinton, Deep Boltzmann Machines. Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics (2009)}), which allows prior structure to be embedded within a deep model of the joint distribution of molecular and high-level phenotype signatures (i.e. acute to chronic pain signatures).

Traditional classification methods such as logistic regression could predict phenotype directly from genotype, without using intermediates such as the transcriptome (**Aim 3 Figure 2A**). In contrast, the DSPN model will be constructed via a series of intermediate models that add layers of structure. Diverse data types from the A2CPS Consortium and prior knowledge such as genotype data, transcriptome data, proteome data, metabolome data, cell-type marker genes, co-expression modules, enhancers, as well as imaging data and EHR, will be included in the intermediate models and further integrated in the DSPN model for a top-layer prediction on the observed traits (i.e. phenotypic pain signatures) (**Aim 3 Figure 2B**).

A key advantage of the DSPN model is interpretable. By using this generative architecture, we ensure that the DSPN model is able to impute intermediate phenotypes, as well as provide forward predictions from genotypes to phenotypic pain signatures. The DSPN model is defined to be a conditional DBM, with extra structure added to the visible units to reflect regulatory relationships between various intermediate phenotypes. The model will be trained using persistent Markov Chain Monte Carlo, and prediction is performed by minimizing the model free-energy (intermediate variables) or feed-forward prediction (observed traits) (see \cite{R. Salakhutdinov, G. Hinton, Deep Boltzmann Machines. Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics (2009)}).

3.3a4 Futility analysis on the core analysis

As stated in the RFA, we will help lead a discussion of whether.... comprehensive analysis to check whether the gene signatures that we developed for acute to chronic pain will be sufficiently powered after the study completes at the end of year three. We imagine that this will take place in to be formed analysis group (described above in XXX). It will involve many participants not just those in the DIRC but those in the CCC, which has many statisticians. Also, we anticipate that not just signatures developed by the DIRC will be evaluated but those by other components of the A2CPS and that these signatures will reflect many different data modalities (eg transcriptome, proteome, metabolome....) and that a big part of the assessment will be determining which modality provides the greatest power.

To be concrete about this process , we describe more for what we could to specifically assess the signatures developed in the DIRC above, For these these, We will check if we can develop a signature with enough statistical power or just marginal power but can be reasonably improved with the addition of more data. In addition to power analysis, we will also perform mediation analysis. Identified candidate signatures will be evaluated using mediation analysis to distinguish between true biological association and mediators. For example, one differentially expressed gene between case and control group can be independently associated with the phenotype and other variables, or 2) associated with only with the other variable through its effects the phenotype or vice versa. Mediation will be assessed using the Baron and Kenny criteria. Only in the absence of a significant indirect effect, we will conclude the true association. We will do this analysis in a consortium wide fashion by looking at the signatures that other groups have developed and try to benchmark them in it in terms of their power to unify our scientific discoveries.

We will use different statistical schemes tailored specifically for particular computational methods. For instance, the simplest approach would be for differentially expressed genes. We will calculate a desirable sample size by assuming the negative binomial distribution of the read counts and using a generalized linear model at the gene level that considers the dependence between gene expression level and its variance (dispersion) \cite{29843589}. These assumptions have been proved to properly control the false positive rate and provide an accurate estimation. Finally, we have extensive experience in relating such calculations to PsychENCODE and cancer driver detection in determining cohort size. With a similar sample size, we have successfully found DEX signatures and modules to in various cancer and psychiatric diseases.

3.3b Integrative analysis of the A2CPS Consortium's data with genomic variants

To investigate the genetic basis of pain signatures, we will integrate processed functional genomic data with the genotype information from the same individuals in order to identify

eQTLs, and will extend this strategy to identify metabolomic variants (mQTLs), as well as variants associated with proteomic changes (pQTLs). To further dissect the associations between genomic elements and QTLs, we will compare all of the different types of QTLs with each other and with different genomic annotations. In addition, to characterize the rare variants within the individuals studied, we will perform burden tests with LARVA \cite{26304545}, a statistical method developed by our group, to identify genomic regions that are over or under represented in terms of the number of rare variants. Furthermore, we will perform allelic analysis of available functional genomic data to identify allelic heterozygous variants using AlleleSeq \cite{21811232}. Finally, we will use FunSeq \cite{25273974} 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.

We will also compile these identified eQTLs, mQTLs, pQTLs, as well as their associated pain phenotypic significance into easy-to-use database formats, and collaborate with the SOC to make them accessible to the consortium, as well as external researchers through the A2CPS data portal.

3.3c Integrative analysis of the A2CPS Consortium's data with external data sources

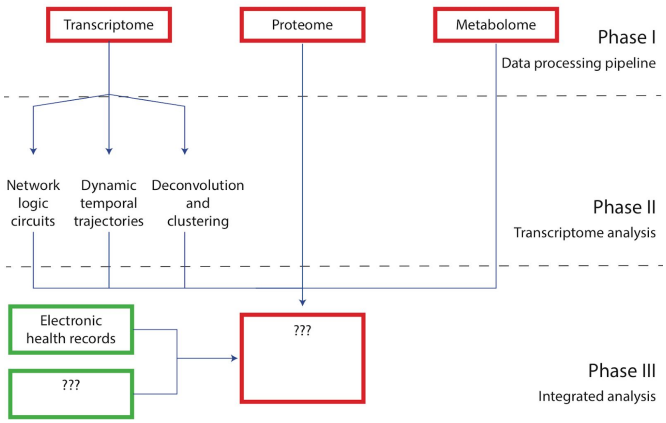
To generate a comprehensive molecular map and regulatory network for the acute to chronic pain signatures, we will integrate complementary large-scale data from external sources, including ENCODE, Roadmap Epigenomics, Blueprint Epigenome, GTEx and exRNA Consortium. Specifically, we will incorporate the ENCODE TF data, the GTEx tissue-specific profiles, and the epigenetic marks of transcriptional regulatory elements from the Roadmap Epigenomics and Blueprint Epigenome. From these datasets we will construct integrative models relating epigenetics and transcriptomics using our previously developed statistical and machine learning approaches \cite{21926158,22955978}. Briefly, combined datasets of genomic features in small bins (e.g. 100bp) will be correlated with expression values over those regions. We will then generate statistical models relating epigenetic marks, TF binding, as well as gene expression.

We will further extend these models to incorporate proteomic and metabolomic data. To build our integrated models, the metabolomic and proteomic data will be combined with pathway information, such as Gene Ontology, KEGG, Reactome and WikiPathways \cite{17923450,22424236,21177976,12140549}. These pathways will be linked to transcriptomic data through their associated genes, using the same machine learning approaches to relate transcriptional activity to metabolite and protein abundances. Thus, we can integrate metabolomic and proteomic data with epigenetic and cis-regulatory data. Finally, the large depth and coverage of transcriptomic experiments will be leveraged to develop integrative models, which will be valuable for the identification of key biomarkers for acute to chronic pain signatures.

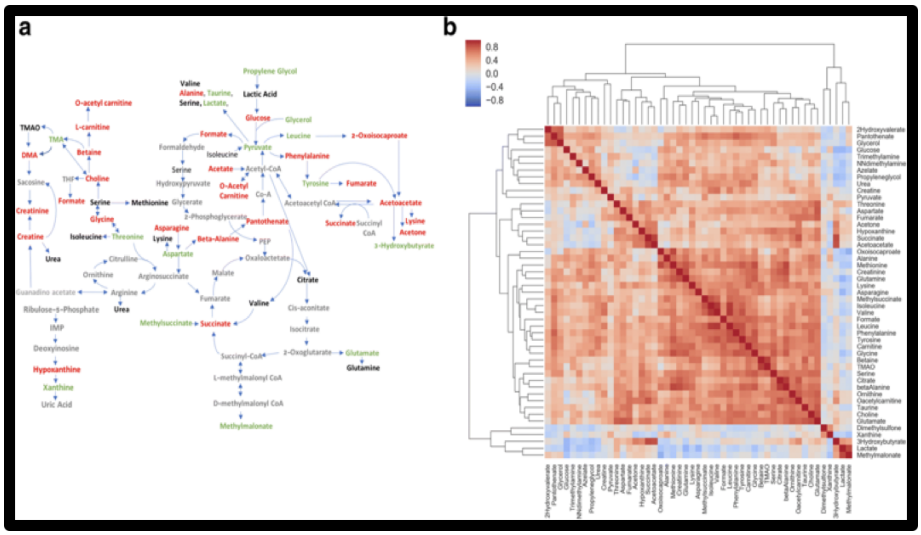
==> Julie END}}

Figures (6 => 1.5pg)

A very primitive version [[TXL]]:



Aim 1 figures



Metabolomics Figure 1. Blood metabolite clustering. **a** Overview of the metabolites identified by 1H-nuclear magnetic resonance (NMR) organized by metabolic pathway. **b** Heat map and hierarchical cluster analysis indicate positive relationships between polar metabolites identified by 1H-NMR in serum from patients with rheumatoid arthritis before treatment with rituximab.

Aim 2 figures

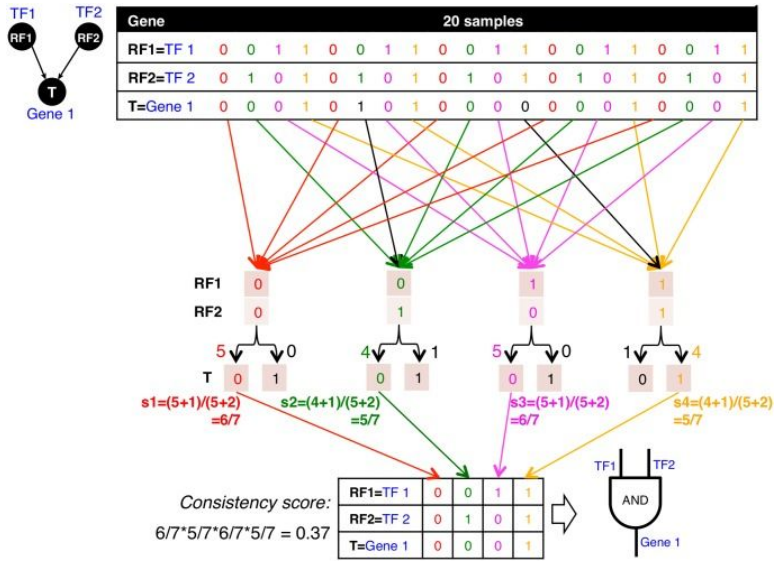


Figure 1 Procedures for mapping logic gates and calculating consistency scores.

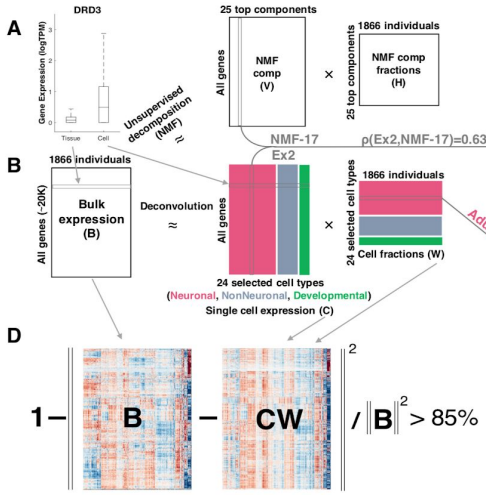


Figure 2 Deconvolution analysis of bulk and single-cell transcriptomics reveals cell fraction changes across the population.

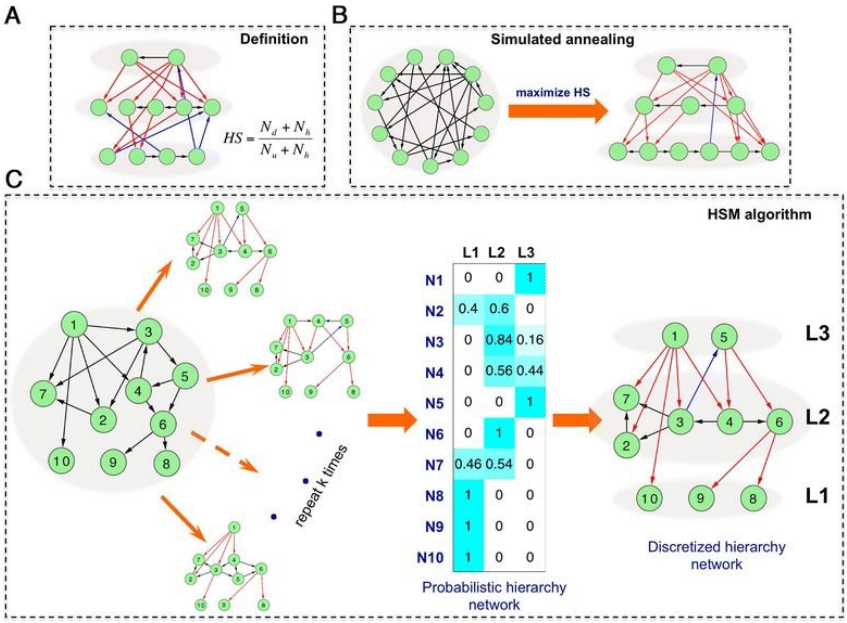


Figure 3 The schematic diagram of the hierarchy score maximization algorithm. In hierarchical networks, the downward, upward, and horizontal edges are shown in red, blue, and black colors, respectively.

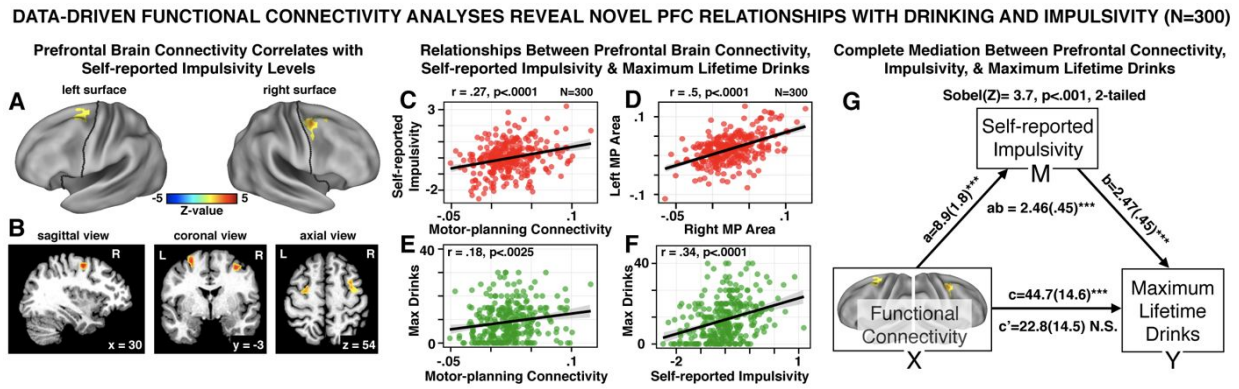


Figure 5 (Imaging). (A-B) Data-driven global whole-brain connectivity reveals premotor PFC area that is related to self-reported impulsivity and in turn drinking. (C) Relationship between bilateral premotor connectivity and impulsivity. (D) Relationship across two bilateral premotor areas. (E) Relationship between maximum lifetime drinks and premotor connectivity (r=.18, p<.0025, 2-tailed) and (F) impulsivity (r=.34, p<.0001, 2-tailed). (G) Direct relationship between premotor connectivity and drinking establishing a possible causal model.

Aim 3 figures

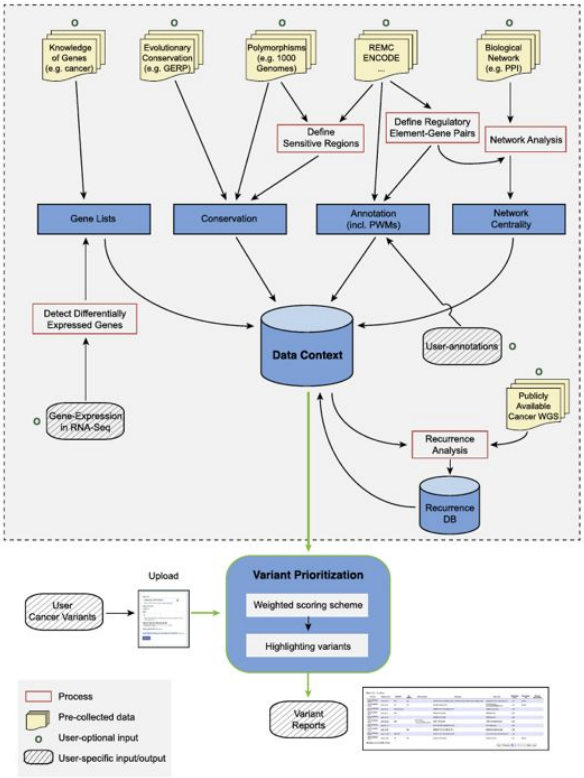


Figure 1. The workflow of FunSeq.

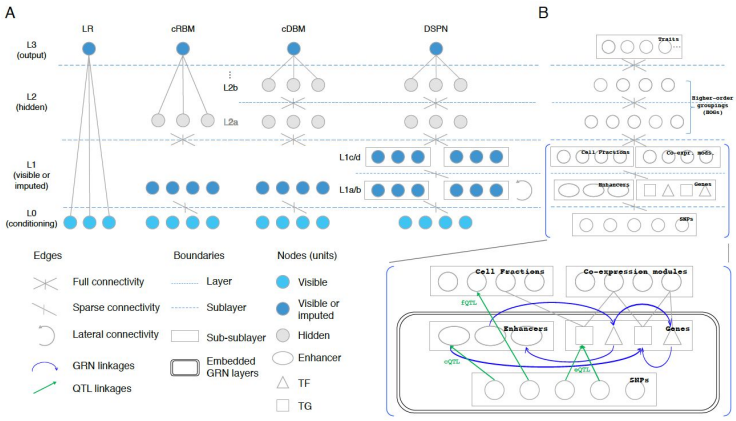


Figure 2. DSPN model (A) The schematic outlines the structure of the following models: Logistic Regression (LR), conditional Restricted Boltzmann Machine (cRBM), conditional Deep Boltzmann Machine (cDBM), and Deep Structured Phenotype Network (DSPN). Nodes are partitioned into four layers and colored according to their status as visible, visible or imputed or hidden. (B) DSPN structure is shown in further detail.

###END OF FIGURES

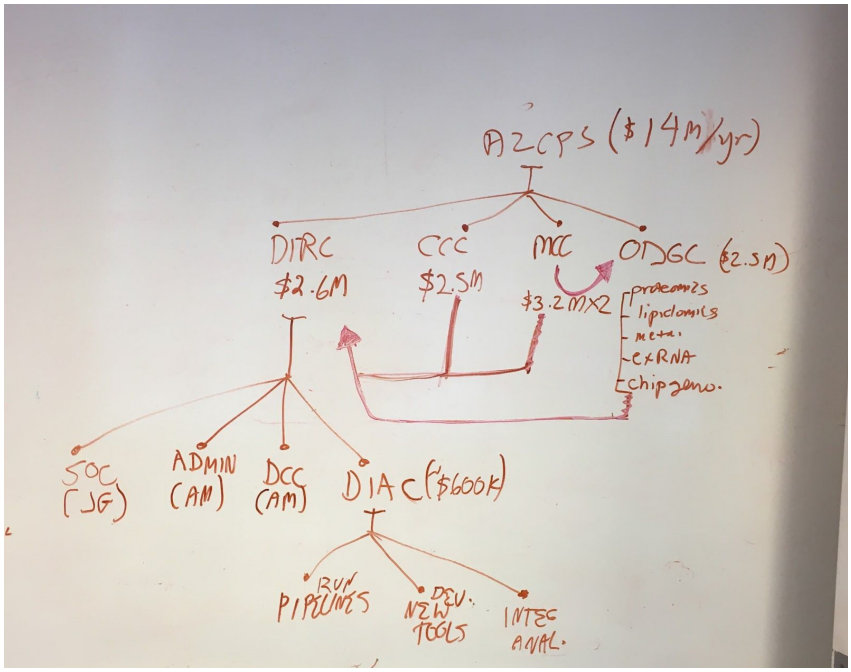
DIRC SOC

%%% 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. PMID:26467476

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. PMID:

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. PMID:30075744

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. PMID: 30150755

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. PMID:23861661

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. PMID: 29893550

Lipidomics Section References

1. Quehenberger, O., and E. A. Dennis. 2011. The human plasma lipidome. *N. Engl. J. Med.* **365**: 1812-1823. PMID:22070478

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. PMID: 20671299

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. PMID: 20923771

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. PMID: 25598080
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. PMID: 21689782
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. PMID: 25074422
7. Dennis, E. A., and P. C. Norris. 2015. Eicosanoid storm in infection and inflammation. *Nat Rev Immunol*. **15**: 511-523. PMID: 26139350
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. 22493235
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*. 30130298
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. 25404585
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. 22695969
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. 23827684
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. 15722563

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. 19098281
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. 27816901
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. 28193460