BCATS 2015

Abstract Book

April 6th, 2015

Welcome

To the fifteenth annual Biomedical Computation at Stanford Symposium (BCATS)!

As a symposium that stands at the intersection of computer science, statistics, biology and medicine, BCATS has enjoyed continued growth over the last decade as the fields of bioinformatics and computational biology gained traction. Today, BCATS draws researchers around the Bay area from fields as diverse as proteomics and population epidemiology.

We would like to thank our keynote speakers, student presenters, sponsors, and all 2015 attendees.

Organizing Committee

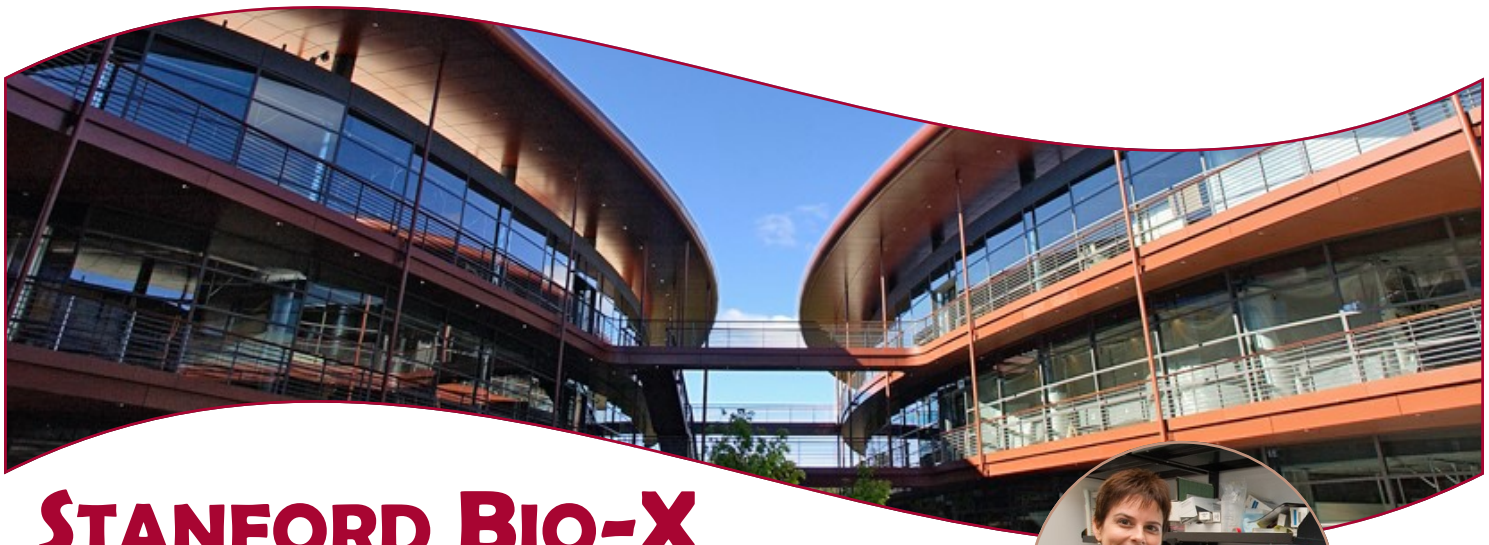
Rachel Goldfeder (PhD candidate in Biomedical Informatics)

Emily Mallory (PhD candidate in Biomedical Informatics)

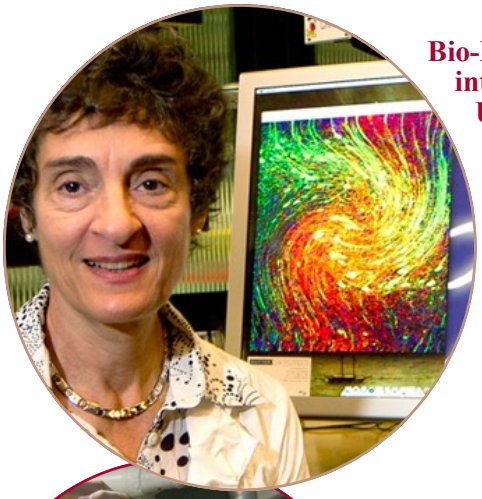
Yifei Men (MS candidate in Computer Science)

Kun-Hsing Yu (PhD candidate in Biomedical Informatics)

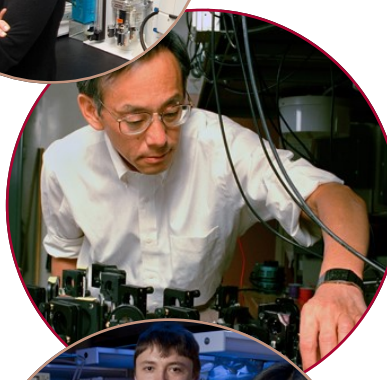
Weizhuang Zhou (PhD candidate in Bioengineering)



STANFORD BIO-X

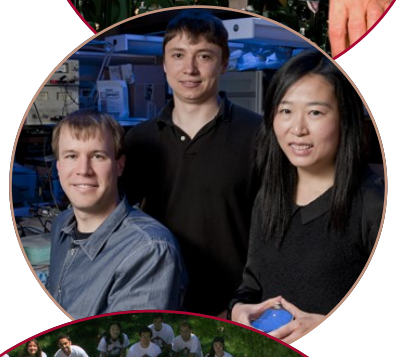


Bio-X is Stanford's pioneering initiative in the interdisciplinary movement at Stanford University. Over the past 16 years, Bio-X has developed into an **impactful institute that brings together biomedical and life science** researchers, clinicians, engineers, physicists, and computational scientists to unlock the secrets of the human body.



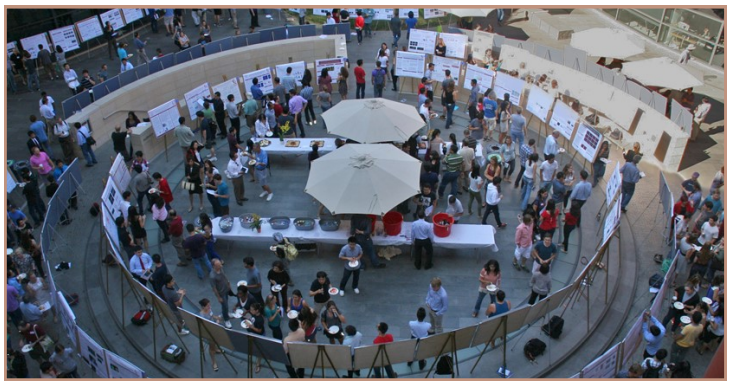
Over 700 Stanford faculty from all 7 Stanford schools and 60+ departments are affiliated with Bio-X, and participating in the following **Core Programs**:

- **Interdisciplinary Initiatives Seed Grant Program (IIP)**
- **Ventures**
- **Graduate Student Fellowships and Bio-X Stanford Interdisciplinary Graduate Fellowships**
- **Undergraduate Research Awards**
- **Travel Awards**



In addition, Bio-X establishes numerous **collaborations between the Stanford scientists and large corporations** by being a networking portal and providing opportunities for companies to join and participate in **symposia/seminars/mixers and customized technical summits, developing faculty liaisons**, and more.

The **James H. Clark Center is the hub of Bio-X**, comprising the equipment, resources, and utilities required to conduct breakthrough research at the cutting edge of engineering, science, and medicine.



To learn more about Bio-X, visit our website:

biox.stanford.edu

STANFORD BIO-X

Stanford Biomedical Informatics Training Program

The Biomedical Informatics Training Program (BMI) is an interdisciplinary graduate and postdoctoral training program, part of the Biosciences Program at Stanford University's School of Medicine. We offer MS and PhD degrees, and other coursework and research options.

Full-Time Graduate Programs:

- PhD in Biomedical Informatics
- Research MS degree (primarily, but not exclusively, for those with PhD and/or MD)

Part-Time Distance Education Programs:

- Professional/Honors Cooperative Program MS degree
- Certificate in Clinical Research Informatics or Bioinformatics
- Non-degree option for individual classes

For Stanford Students:

- Coterminal MS degree in Biomedical Informatics in addition to BA/BS
- PhD Minor in Biomedical Informatics for Stanford graduate students

For Stanford Medical Students:

- Coursework and research opportunities through Scholarly Concentration and Medical Scholars programs

All degree programs require rigorous coursework in Biomedical Informatics core courses, a coherent set of electives from Computer Science, Statistics, Math and/or and Engineering, and training in Social, Legal and Ethical issues.

For more information, contact:

Student Services Officer
Stanford Biomedical Informatics Training Program
Medical School Office Building, Room X-215
1265 Welch Road, Mail Code: 5479
Stanford, CA 94305-5479

Phone: (650) 723-1398
Fax: (650) 725-7944
email: bmi-contact@lists.stanford.edu
http://bmi.stanford.edu

Supported by National Library of Medicine Training Grant LM007033

Stanford Center for Computational, Evolutionary and Human Genomics

Upcoming Talks

Pierre Luisi
Fraser Lab
12pm April 8, Clark S360

CEHG Symposium
Paul Brest Hall
April 13-14

Brian Johnson
UC Davis
12pm April 15, Clark S360

Fernando Racimo
UC Berkeley
12pm April 29, Clark S360

Pleuni Pennings
UC San Francisco
12pm May 6, Clark S360

Sandeep Ventakaram
Petrov Lab
12pm May 20, Clark S360

The Stanford Center for Computational, Evolutionary and Human Genomics (CEHG) was founded in 2012 to foster interdisciplinary research at the University. A collaboration between the School of Humanities and Sciences and the School of Medicine, the Center is the intellectual home of 40 professors and over 200 postdoctoral scholars and graduate students. CEHG funding opportunities include the Fellowship Program, Trainee Research Grants, and Faculty Seed Grants.

CEHG researchers at Why We Can't Wait in San Francisco, CA

Contact Us

cehg.stanford.edu
facebook.com/StanfordCEHG
twitter.com/StanfordCEHG
StanfordCEHG@stanford.edu

Stanford Cancer Institute

Stanford MEDICINE

The **Stanford Cancer Institute (SCI)** is dedicated to efficiently translating scientific discoveries into personalized cancer prevention, identification and treatment. SCI coordinates cancer research, clinical trials and patient care throughout the University's academic departments, hospitals and clinics in order to identify cross-cutting opportunities which may not be apparent to individuals investigators or departments.

Research Programs

Drawing upon the expertise of more than 400 researchers and clinicians from the Schools of Medicine, Engineering and Humanities & Sciences, SCI is at the forefront of today's advances in cancer research and medicine.

- Cancer Biology
- Radiation Biology
- Cancer Stem Cells
- Cancer Imaging and Early Detection
- Translational Oncology
- Lymphoma and Leukemia
- Immunology and Immunotherapy of Cancer
- Population Sciences

Shared Resources

SCI provides expert support services and state-of-the-art equipment to all Institute members through ten subsidized core facilities.

- Animal Tumor Models
- Bioscience Screening
- Cancer Biostatistics
- Cancer Imaging Shared
- Cell Sciences Imaging
- Flow Cytometry
- Genomics
- Human Immune Monitoring
- Proteomics
- Tissue Procurement

Training Programs

Fostering interdisciplinary collaboration from the laboratory to the clinic, the Stanford Cancer Institute offers an exceptional training ground for the next generation of leaders in cancer research and medicine.

- Cancer Biology PhD Program
- Immunology PhD Program
- Research Training Programs
- Comprehensive Cancer Research Training Program
- Continuing Medical Education

cancer.stanford.edu

STANFORD COMPUTER FORUM

A Corporate Affiliates Program

The Stanford Computer Forum is a cooperative venture that encourages collaboration between the Computer Science and the Electrical Engineering Departments at Stanford, and 60+ companies located in Silicon Valley, the rest of the U.S., Asia, and Europe. The Forum provides a mechanism for developing interaction with industrial researchers and their academic counterparts, promoting the exchange of the most advanced technological ideas in fields of computer science and electrical engineering. The Forum also offers industry the opportunity to become familiar with the professional abilities and interests of Stanford students through its active recruiting program.

For further information please visit http://www.forum.stanford.edu or e-mail forumstaff@cs.stanford.edu

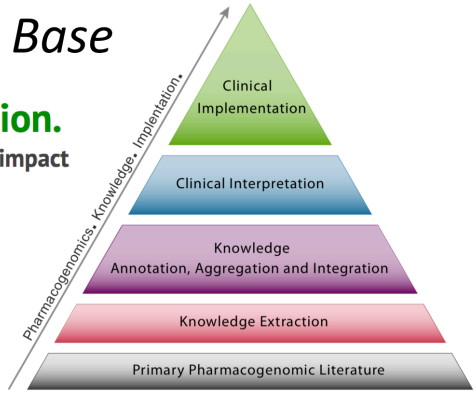
The Pharmacogenomics Knowledge Base

Pharmacogenomics. Knowledge. Implementation.

PharmGKB is a comprehensive resource that curates knowledge about the impact of genetic variation on drug response for clinicians and researchers.

PharmGKB

Visit us at https://www.pharmgkb.org



Additional Sponsors

Genentech
A Member of the Roche Group

STANFORD
CENTER FOR
GENOMICS &
PERSONALIZED
MEDICINE

STANFORD | Biophysics
SCHOOL OF MEDICINE

Stanford | Bioengineering
Schools of Engineering & Medicine

Schedule for the BCATS 2015 Symposium

8:15 am Badge Pickup and Breakfast

8:45 am Opening Remarks

Session 1 (9:00 am – 10:15 am)

9:00 am **Keynote Address:** Michael K. Gilson, M.D., Ph.D. (pg. 10)

9:45 am Scientific Talks

- Nishith Reddy: *Predicted effects of host immune reponse on treatment outcomes with the "shock and kill" HIV eradication framework* (pg. 14)
- Suzanne Tamang: *Improving the Foundation of Population-based Payment Approaches by Predicting Cost Blooms* (pg. 20)

10:15 am Coffee Break

Session 2 (10:30am – 11:30 am)

10:30 am Scientific Talks

- Sarah Allec: *Solving the phase problem from few and noisy data* (pg. 24)
- Linda Szabo: *Statistical algorithms for improved detection of linear and circular splice events from RNA-Seq data* (pg. 16)
- Zachary Zappala: *A statistical framework for evaluating the regulatory impact of rare, population-specific non-coding variation* (pg. 17)
- Sharon Briggs: *Single cell expression analysis in human iPSCs reveals predictive developmental patterns and novel associations*. (pg. 22)

11:30 am Lunch

Session 3 (1:00 pm – 2:30 pm)

1:00 pm **Keynote Address:** Eliezer Van Allen, M.D. (pg. 11)

Scientific Talks

- Alex Pankov: *Cell cycle signature and tumor phylogeny are encoded in the evolutionary dynamics of DNA methylation in glioma* (pg. 23)
- Alicia Schep: *High-resolution nucleosome positioning from ATAC-seq chromatin accessibility data* (pg. 19)
- Graham Heimberg: *Widespread gene expression covariance enables massively parallel, low-coverage mRNA-seq* (pg. 15)

2:30pm Poster Session and Coffee

Session 4 (3:30 pm – 5:00 pm)

3:30 pm Scientific Talks

- Julia Fukuyama: *Interpretable ordinations for microbiome data using sparse DPCoA* (pg. 18)
- Alison Feder: *Less efficient drugs lead to softer sweeps in intra-patient populations of HIV-1* (pg. 21)

4:00 pm **Keynote Address:** Kristen Naegle, Ph.D. (pg. 12)

4:45 pm Closing Remarks

5:00 pm Networking Session on LKSC lawn

Keynote Talks

Keynote Talk

Methods for computer-aided drug design

Michael K. Gilson, M.D., Ph.D.

Computer-aided drug design is a growing field that is increasingly used to develop new therapeutics. I will discuss the two main thrusts of our lab, where the challenge of drug design is tackled by informatics and computer simulations. BindingDB, the first publicly accessible protein-ligand interaction database, now holds over a million data, including unique datasets drawn from the chemical biology literature and US Patents. I will provide an update of the project, including the wide range of methods now available for query, browsing, downloading, and application of this large dataset. Our second project uses simulations to understand the physical basis of molecular recognition, with the ultimate goal of improving the accuracy of tools for computer-aided drug design. Experimentally measured protein-ligand binding thermodynamics, typically from calorimetry, often are quite puzzling. For example, small changes in a compound may lead to correspondingly small changes in affinity, but surprisingly large swings in binding enthalpy and entropy. I will discuss our recent progress in using detailed molecular simulations to suggest explanations for these puzzles.

Keynote Talk

Driving precision cancer medicine through computational oncology

---

Eliezer Van Allen, M.D.

Utilizing increasingly complex somatic and germline genomics in the prospective clinical setting, such as whole exome sequencing (WES), may directly impact the care of cancer patients. Since this approach greatly expands the quantity of genomic data generated for an individual patient, multiple analytical and informatics innovations are needed for successful clinical implementation. These include new methods that enable clinical analysis and interpretation of large scale WES data, effective representation of WES data in a clinical environment for physician consumption, and knowledge-sharing platforms to ensure effective curation of clinically relevant variants. In this talk, we will explore emerging methods for the clinical interpretation of WES data and approaches that may help physicians understand these data for point-of-care clinical use. Finally, we will explore new cancer genomics knowledge bases and emerging platforms for crowdsourcing variant curation that may enable prospective WES-scale genomic profiling. Overall, this field of clinical computational oncology may drive the widespread implementation of precision cancer medicine.

Keynote Talk

Computational systems biology approaches for unraveling tyrosine phosphorylation in cell signaling networks

---

Kristen Naegle, Ph.D.

Tyrosine phosphorylation is a regulated post-translational modification, which is the cornerstone of receptor tyrosine kinase networks. Receptor tyrosine kinases are fundamental to multicellular organism development and homeostasis. Given the advance in recent technologies, the discovery of the extent of tyrosine phosphorylation has exploded, yet specific understanding of the function of these modifications has lagged. In this talk, I will discuss our labs efforts to develop computational tools that when combined with systems- or proteome-level information yield hypotheses about the function of tyrosine phosphorylation within cell signaling networks. I will specifically cover how we have developed a community resource for post-translational modifications and developed ensemble clustering techniques that allow for the discovery of novel, transient macromolecular complex formations.

Scientific Talks

Scientific Talk

Predicted effects of host immune reponse on treatment outcomes with the "shock and kill" HIV eradication framework

---

Nishith Reddy

Current antiretroviral treatments (ART) for HIV have been successful at suppressing new infection events and maintaining patient viral loads at below clinically detectable levels. In spite of these benefits, routine drug use continues to place a hefty burden on both patient health and global resources. As a result, emerging therapies are now focusing on developing a permanent cure. The primary obstacle for complete HIV eradication is a latent reservoir of resting CD4+ T cells, which harbor replication competent HIV DNA. Recent studies have proposed a “shock and kill” strategy, where latency reversal agents are used to reactivate latent infection to productive infection, which can then be cleared by the host immune system under ART. Existing in vitro models for the “shock and kill” framework fail to account for the effects of immune factors, specifically cell-intrinsic immunity, on treatment outcomes. To address these concerns, we developed a stochastic computational model assuming constant rates for uninfected CD4+ T cell production, viral replication, and viral clearance to predict long-term patient treatment outcomes in the context of host immune responses. These treatment outcomes are characterized by patient viral load and reservoir size. A class of latency reversal agents called histone deacetylase (HDAC) inhibitors has been the focus of current research. A particular HDAC inhibitor, vorinostat, has been shown to induce transient increases in cell susceptibility in uninfected cells and proliferation in latently infected cells. This model predicts that temporarily enhanced proliferation in the reservoir due to vorinostat can profoundly limit reservoir depletion and duration until viral recrudescence. A recent vorinostat clinical trial showed that there are significant changes in host restriction factor expression affecting post-integration viral production both on-treatment and post-treatment. We demonstrate that these host restriction factors may help prevent viral rebound after ART cessation. Additionally, the model indicates these factors may explain clinical HIV transcriptional ratios, which show significant changes in viral load, but not HIV DNA. These finding suggest that immune responses are crucial to determining “shock and kill” treatment efficacy.

Scientific Talk

Widespread gene expression covariance enables massively parallel, low-coverage mRNA-seq

---

Graham Heimberg

Motivation: Deep mRNA-sequencing (mRNA-seq) is a powerful technique that sequences millions of short cDNA reads in order to precisely estimate global gene expression profiles. While high read depth requirements often limit the scale of RNA-seq experiments to just dozens of samples, a dramatic reduction in read depth would introduce noise into estimating gene expression profiles. Our goal is to understand how noise introduced by low read depth impacts common downstream analysis and the extraction of biological information.

Results: To understand the impact of measurement noise on RNA-seq, we develop a quantitative model that characterizes the impact of low-coverage sequencing on Principal Component Analysis (PCA). We focus on PCA because it is a common analysis technique that shares a mathematical foundation with clustering methods. Interestingly, the mathematical model reveals that the tolerance of PCA to measurement noise depends on the effective dimensionality of the high-dimensional data. We show that covariance in gene expression data reduces effective dimensionality and provides increased robustness to noise. Using the model and parameters obtained from published data (Shen et al.), we predict 70% recovery of principal components at less than 1% of the original read depth in two published data sets. We confirm this prediction through simulations of shallow sequencing and demonstrate accurate Gene Set Enrichment Analysis and sample classification at very low read depth. These results suggest that for many analyses that are based on PCA or related methods, mRNA-seq throughput can increase by two orders of magnitude. To ask whether low dimensional geometry might enable shallow sequencing more generally, we surveyed almost 2,000 additional gene expression data sets from yeast, mouse, and human, and found that low effective dimensionality due to covariance is ubiquitous. This result demonstrates that high throughput mRNA-seq can be reliably used to the vast majority of gene expression profiling experiments.

Scientific Talk

Statistical algorithms for improved detection of linear and circular splice events from RNA-Seq data

---

Linda Szabo

The pervasive expression of circular RNA from protein coding loci is a recently discovered feature of many eukaryotic gene expression programs. Computational methods to discover and quantify circular RNA from RNA-Seq data are essential to the study of the mechanisms of circular RNA biogenesis and potential functional roles they may play. Here, we present a new statistical algorithm that increases the sensitivity and specificity of circular RNA detection by discovering and quantifying circular and linear RNA splicing events at both annotated exon boundaries and in un-annotated regions of the genome. Unlike previous approaches that rely on heuristics such as read count and exon homology to determine confidence in prediction of circular RNA expression, our algorithm takes a statistical approach and can detect circular RNAs with high confidence even when present at very low levels. Applying our algorithm to RNA-Seq data available from Sequence Read Archive (SRA) for cells treated with RNase-R to enrich for circular RNA, we not only demonstrate improved accuracy, but also show that inferences about the regulation of circular RNA need to be revisited due to biases in the algorithms used for these genome-wide analyses. Additionally, we applied our algorithm to RNA-Seq data generated by the ENCODE consortium from polyA+ and polyA- fractions of 18 cell types and found several striking examples of highly expressed putative circular RNAs known to be false positives. Such results are heuristically post-filtered by other algorithms, but are automatically flagged by our algorithm as artifacts. In work currently under review, we have used this algorithm to analyze a panel of human fetal tissues sampled at several developmental time points. We discovered regions of the fetal brain, such as the frontal cortex, with marked enrichment for genes where the circular isoform is dominant and found general induction of circular RNAs in many tissues, including the heart and lung, during human fetal development. Moving forward, we are building upon the concepts implemented in the original version of our algorithm for circular RNA detection to develop a statistical algorithm that can be used to quantify known and novel circular and linear isoforms in genomes that differ significantly from the reference, such as cancer genomes.

Scientific Talk

A statistical framework for evaluating the regulatory impact of rare, population-specific non-coding variation

---

Zachary Zappala

Large-scale genome sequencing efforts have revealed a great diversity of coding and non-coding genetic variation. Identifying the impact of rare, non-coding genetic variation on phenotypic variation in a population remains an open challenge in population genomics. In the absence of large sample sizes, rare variants do not have an appreciable frequency to power traditional expression quantitative trait loci (eQTL) studies. Using population-scale RNA-Seq data for 606 Sardinian individuals, we have developed a framework for combining population structure, genetic background, gene expression, and allele-specific expression (ASE) to find rare variants that are associated with family-specific patterns of gene regulation. Our analysis of 66 trios in the Sardinian population reveals a number of gene expression traits that co-segregate in a family with significant signals of ASE. These signatures are consistent with the model of a rare, causal cis-regulatory element that also segregates within the family. Within these outlier individuals, we test for rare non-coding variants that overlap with functional annotations such as conservation, DNase hypersensitivity, and transcription factor binding. By utilizing this pipeline, we have identified rare variants in the Sardinian population that contribute to family-specific regulatory traits and hold pathogenic potential.

Scientific Talk

Interpretable ordinations for microbiome data using sparse DPCoA

Julia Fukuyama

Recent advances in sequencing technology have made it possible to obtain a systematic picture of the composition of bacterial species living in various communities. Not much is known about most of the bacteria identified in this way, but it is possible to estimate their phylogenetic relatedness. It is often desirable to take this phylogenetic information into account when making ordinations of the data for exploratory analysis. A particularly popular approach to incorporating phylogenetic information has been to use UniFrac distances (Lozupone and Knight, 2005) in conjunction with multidimensional scaling. Although it is very widely used, UniFrac has the drawback that the ordination axes don't have any intrinsic meaning --- any description of the axes is done in an ad hoc way by the analyst based on external information. There have been other proposals for more interpretable ordinations that take into account phylogeny, in particular double principal coordinates analysis (DPCoA, Pavoine et al. 2004). DPCoA turns out to be very similar to UniFrac but gives a little more information to guide the interpretation of the axes. However, although the axes given by DPCoA have a meaning (an improvement over UniFrac), the interpretation is still unsatisfying since it is in terms of every species present in the analysis and only leads to a very general interpretation (e.g., an axis corresponds to more species in a certain phylum). We have developed a new method, based on DPCoA, that gives an ordination that incorporates phylogenetic information and has meaningful, interpretable axes. We added a sparsity-inducing penalty to the objective in a certain formulation of DPCoA, which leads to axes that are sparse and have their non-zero coordinates grouped together on the tree. This allows us to interpret axes as associated with the abundance of certain lineages. We demonstrate our method on a microbiome dataset taken from the teeth of 14 healthy individuals over the period of a week, and also show how our method compares to other competitors on simulated data.

Scientific Talk

High-resolution nucleosome positioning from ATAC-seq chromatin accessibility data

---

Alicia Schep

Chromatin accessibility and nucleosome positioning are important determinants of transcriptional regulation, as nucleosomes compete with transcription factors for binding of DNA. ATAC-Seq (Assay for transposase-accessible chromatin using sequencing) profiles chromatin accessibility genome wide, and fragmentation patterns of DNA generated by this method provide information about nucleosome positioning in regulatory regions. By applying ATAC-Seq to Saccharomyces cerevisiae, for which a chemical map of nucleosomes at base-pair resolution is available, we observe that ATAC-Seq signal around nucleosomes is highly structured. We have developed a method—NucleoATAC—that uses this highly structured pattern to infer nucleosome positions at high-resolution within accessible regions of the genome. Unlike methods developed for MNase-Seq, NucleoATAC can resolve the 10-bp rotational positioning of nucleosomes. Furthermore, as ATAC-Seq assays accessible chromatin, high-confidence nucleosome position calls are concentrated within regulatory regions of the genome, nucleosome calls can be assigned occupancy levels, and the precise locations of nucleosome-free regions (NFRs) can be inferred. By applying NucleoATAC to Saccharomyces cerevisiae, Schizosaccharomyces pombe, and the human GM12878 lymphoblastoid cell line, we observe different nucleosome sequence preferences and chromatin architecture between the species. We show that for all three species, nucleosome-free regions uniformly span less than one nucleosome. Additionally, we show that in contrast to S. cerevisiae and human transcription start sites the majority of transcription start sites in S. pombe are not characterized by a stable nucleosome free region. Future studies will apply NucleoATAC towards understanding nucleosome positioning and occupancy changes in dynamic systems.

Scientific Talk

Improving the Foundation of Population-based Payment Approaches by Predicting Cost Blooms

---

Suzanne Tamang

Accurately identifying individuals likely to incur the bulk of healthcare spending is crucial for the success of many smaller healthcare organizations under value based payment models. The objective of this study was to improve prediction of health spending for an individual's care in the subsequent year by distinguishing new “cost blooms” from continued high health cost. We modeled longitudinal health spending trends from 2005 through 2011 for over 2.1 million Danish residents and evaluated cost prediction models on unseen data from the subsequent year. We defined “high cost” as an individual in the top decile of annual health expenditures and a “cost bloom” as an individual in the top 10% of health spenders that were not high cost the prior year. Using population healthcare data from 2010 to predict future costs in 2011, our best model captured 60% of the annual costs accrued by high cost individuals. For individuals with cost blooms, the best model captured 49% of the total surge in costs. Compared with established approaches to predicting an individual’s future health cost, we obtain a 21% and 30% improvement in predicted costs for high cost individuals and individuals that experienced cost blooms respectively. Our results demonstrate that enhancing traditional claims data sources to include social relationship as well as temporal indicator variables can lead to substantially better healthcare expenditure prediction models—especially for individuals that experienced a cost bloom. We expect our study to inform payers and providers, who need to effectively operate under value-based payment models. Identifying individuals that will experience spending blooms can enable accurate risk adjustment as well as enable targeted interventions.

Scientific Talk

Less efficient drugs lead to softer sweeps in intra-patient populations of HIV-1

Alison Feder

Viral populations tend to be characterized by exceptionally large size and high mutations rates. These populations can be highly adaptable because they are not limited by the influx of mutations, and multiple positively selected variants can arise simultaneously in a so-called soft sweep. However, under combination antiviral therapy, a declining population size and the low rate at which fully resistant types emerge may bring populations into a mutation-limited regime, where a population must wait for the right variant which will then spread to the exclusion of others in a hard sweep. These two regimes should leave markedly different patterns on the diversity profiles of the populations, which may allow us to test whether or not a population is evolving in a mutation-limited setting. We examine diversity among 7918 HIV-1 Direct-PCR sequences and use ambiguous reads to investigate the prevalence of hard and soft sweeps associated with resistance to antiviral drugs. Using these data, we confirm that drug resistance is associated with lower diversity levels, a common signature of selective sweeps. We further find that resistance to treatments with low success rates generate patterns consistent with soft sweeps, as marked by the relatively small decrease in diversity associated with resistance. Populations receiving high success rate treatments showed patterns more consistent with hard sweeps. This suggests that effective drugs may push HIV-1 populations into a hard sweep regime in which adaptation becomes mutation-limited. Finally, we suggest how this information can be leveraged to test new drug therapies in which information about efficacy is unavailable.

Scientific Talk

Single cell expression analysis in human iPSCs reveals predictive developmental patterns and novel associations.

Sharon Briggs

Background: Single cell gene expression analysis has recently given insights into developmental biology allowing distinct subsets of cells to be identified from heterogeneous populations. In induced pluripotent stem cells (iPSCs) many groups have profiled single cells to understand the mechanism of reprogramming and the similarity to other pluripotent cell types. However, much of this single cell work has been focused on the reprogramming process itself or early time points and an analysis spanning reprogramming to late passage has not yet been performed.

Results: In this study, we profiled over 1300 single cells from two human cell lines across four developmental time points: fibroblasts, iPSC passage 0, passage 3, and passage 20. In addition, cells of both cell line and at all four time points were cultured in different oxygen levels. We show that differences in gene expression are dependent on passage number more than cell line or oxygen tension effects. In addition, we show significant changes in gene expression (including pluripotency and housekeeping genes) over time and are able to use these gene expression patterns to make predictions. Interestingly, this approach has identified a new pluripotency-associated gene in humans, NGLN4X.

Conclusions: Profiling how gene expression changes over different developmental states reveals just how dynamic and variable gene expression is. Even housekeeping genes, assumed to be relatively constant, change appreciably. This work suggests a changing suitability of human iPSCs over time for downstream applications such as differentiation into specific cell types.

Scientific Talk

Cell cycle signature and tumor phylogeny are encoded in the evolutionary dynamics of DNA methylation in glioma

---

Alex Pankov

The clonal evolution of tumor cell populations can be reconstructed from patterns of genetic alterations. In contrast, tumor epigenetic states, including DNA methylation, are reversible and sensitive to the tumor microenvironment, presumably precluding the use of epigenetics to discover tumor phylogeny. Here we examined the spatial and temporal dynamics of DNA methylation in a clinically and genetically characterized cohort of IDH1-mutant low-grade gliomas and their patient-matched recurrences. WHO grade II gliomas are diffuse, infiltrative tumors that frequently recur and may undergo malignant progression to a higher grade with a worse prognosis. The extent to which epigenetic alterations contribute to the evolution of low-grade gliomas, including malignant progression, is unknown. While all gliomas in the cohort exhibited the hypermethylation signature associated with IDH1 mutation, low-grade gliomas that underwent malignant progression to high-grade glioblastoma (GBM) had a unique signature of DNA hypomethylation. Genes with promoter hypomethylation and concordant transcriptional upregulation during evolution to GBM were enriched in cell cycle function, evolving in concert with genetic alterations that deregulate the G1/S cell cycle checkpoint. Additionally, despite the plasticity of tumor epigenetic states, tumor evolution phylogeny reconstructed from DNA methylation was remarkably similar to the evolutionary patterns derived from somatic mutations. These findings highlight widespread co-dependency of mutations and epimutations throughout the clonal evolution of glioma.

Scientific Talk

Solving the phase problem from few and noisy data

Sarah Allec

Additional Authors: Jesse Benavides and Kevin S. Raines

Determining the structure of biomolecules is a fundamental goal of modern biology. To the extent that the phase problem can be solved, progress is generally made in molecular biology, as exemplified by the results of crystallography over the past century. However, due to the difficulty in preparing crystals for most proteins, the domain of applicability of crystallography is limited. Thus the requirement emerges for extending the methods of crystallography to non-crystalline objects. That is, we would like to solve the phase problem for amorphous objects under practical experimental conditions. Progress in several fields over the past three decades has opened the door to realizing this goal through a variety of experimental and computational techniques that together form the field of lensless imaging. For example, advances in accelerator physics, such as the development of large-scale X-ray free electron lasers, has enabled the realization of diffract-and-destroy techniques that have yielded the successful structural determination of virions and nanocrystals. Additionally, algorithmic developments have enabled the routine phasing of diffraction patterns from amorphous objects. However, such methodologies remain limited by the beam flux which remains insufficient to determine the three dimensional structure of biomolecules. To address this limitation, we and others previously developed an algorithm, called Adaptive Phase Retrieval (APR), which requires less diffraction data to solve the phase problem for amorphous objects. This computational framework applies to typical amorphous objects and can be extended to include prior knowledge when it is available. Here we report new results for solving the phase problem with noisy and incomplete data by using APR.

Scientific Posters

Posters

Poster	Presenter	Title	Page
1	Patrick Bradley	Identifying variable and invariable gene families in the human gut microbiome	28
2	Ryan Eshleman	An extensive empirical evaluation of four Named Entity Recognition (NER) systems using textual clinical narratives and full journal biomedical articles.	29
3	Yang Ha	Application of DFT calculations of metal active site in bio-inorganic chemistry	30
4	Daniel Himmelstein	Bringing the compendium of drug indications into the public domain	31
5	Zeinab Jahed	A disulfide bond is required for the transmission of forces through SUN-KASH	32
6	Hardik Kabaria	Modeling current density using a flexible trans-cranial stimulator	33
7	Irene Kaplow	A pooling-based approach to mapping genetic variants associated with DNA methylation	34
8	Aaron Kravitz	A Text-Driven Approach for Constructing a Drug Metabolite Database	35
9	Shruti Marwaha	Drug Repurposing for Gastric Cancer using Genomics Data	36
10	Shaila Musharoff	The Inference of Sex-Biased Demography is Affected by Population Size Changes	37
11	Santrupti Nerli	Genome-wide Prediction of Splice Sites using Maximal Dependence Decomposition	38
12	Chloe O'Connell	Pathways analysis of rare, high-impact variants in mothers of children with autism	39
13	Aleksandra Olow	Phospho Atlas – the resource to explore human phospho-reactome	40
14	Kris Sankaran	structSSI: Simultaneous and Selective Inference for Grouped or Hierarchically Structured Data	41
15	Alicia Schep	High-resolution nucleosome positioning from ATAC-seq chromatin accessibility data	42
16	Christof Seiler	Bayesian Inference of Geometric Deformations Associated to Lower Back Pain	43
17	Hengameh Shams	Regulation of vinculin activation through α-actinin binding	44
18	Craig Smail	Comparing chronic kidney disease profiles identified using diagnostic histories and clustered clinical data with an established genetic risk model	45
19	Mohammad Soheilypour	Characterization of mRNA quality control mechanism through an agent-based modeling framework	46
20	Kelly Speth	Comparison of Bioconductor Packages edgeR and DESeq for Analyzing Differential Expression from RNA-Sequencing Experiments	47

21	Timothy Sweeney	COmbined Mapping of Multiple clUsteriNg ALgorithms (COMMUNAL): A Robust Method for Selection of Cluster Number, K	48
22	Linda Szabo	Statistical algorithms for improved detection of linear and circular splice events from RNA-Seq data	49
23	Zhiying Wang	smallWig: Compression on RNA-seq Expression Data	50
24	Michael Wong	FEATURE-Viz: A Web-Based Visualization System for Comparing Pairs of Protein Pocket-Ligand Interactions	51
25	Junaid Zaman	Is Clinical Atrial Fibrillation Caused by Multiwavelet Reentry? Numerical Analysis of Data from Patients with Atrial Fibrillation	52
26	Tina Baykaner	The Dynamic Interplay Between Organized and Disorganized Zones During Human Atrial fibrillation determines the Likelihood of Termination During Ablation	53

*Please note that these posters will not longer be presenting: Mohammad Soheilypour (poster #20)

Poster No. 1

Identifying variable and invariable gene families in the human gut microbiome

---

Patrick Bradley

The human gut microbiome is a diverse ecosystem, harboring microbes capable of performing a wide array of biochemical functions. Previous studies have shown that the taxonomic composition of the gut microbiota can vary dramatically between healthy individuals. It has also been observed that the overall repertoire of gene families encoded by these microbiota tends to be much less variable across the same samples. This suggests that there may be a "core" set of functions that promote survival in the human gut. Additionally, gene families that are involved in adaptation to more specific niches within the gut might break this overall trend and show more variable abundances. Identifying gene families that are invariable or variable across individuals could therefore help us understand strategies of adaptation to the human gut. To detect gene families in both of these categories, we develop a resampling-based statistical test that identifies gene families that are either more or less variable across individuals than expected. This method controls for between-study variation, as well as the overall mean-variance relationship in the data. We then apply this test to shotgun sequencing samples from healthy individuals, drawn from four different studies of the human gut microbiome. Finally, we assess whether the variable and invariable gene families that we identify can be explained by their estimated phylogenetic distribution. We find that several two-component signaling pathways, despite being predicted to have a restricted phylogenetic distribution, appear to have low variability across individuals. We also find that transporters, particularly oligosaccharide transporters, have abundances that are highly variable despite having broad phylogenetic distributions. Interestingly, many of the protein families that we identify as variable also co-vary across individuals, but do not appear to correlate with measured clinical variables, indicating a potential role for unmeasured variables in driving the functional composition of the microbiome.

Poster No. 2

An extensive empirical evaluation of four Named Entity Recognition (NER) systems using textual clinical narratives and full journal biomedical articles.

Ryan Eshleman

In this project, we perform an extensive empirical evaluation of four Named Entity Recognition (NER) systems using textual clinical narratives and full journal biomedical articles. The four NER systems under evaluation are the National Library of Medicine's MetaMap , Apache cTAKES, University of Michigan's MGrep, and Arizona State University's BANNER. Both MetaMap and cTAKES employ Maximum Entropy based chunking followed by a dictionary lookup, while BANNER uses a pure Conditional Random Fields implementation. MGrep is closed source, and its mechanism is not explicitly known. Furthermore, we have also extensively studied different ensemble approaches built upon the above four NER systems to exploit their collaborative strengths. Evaluations are performed using the hand annotated patient discharge summaries provided by the Informatics for Integrating Biology and the Bedside group (I2B2). We also evaluate these NER systems using the CRAFT dataset. Evaluation measurements include the standard Precision, Recall and F1 based on three span-based matching criteria (exact match, single boundary, any overlap), with or without requiring matching the semantic type of an entity. The main results include (1) BANNER significantly outperforms the other three systems on the I2B2 dataset with F1 values in the range of .73-.89, in contrast to .28 - .60 of other systems; (2) cTAKES, MetaMap and MGrep demonstrate different strengths and weaknesses at identifying different types of medical entities/concepts. An ensemble approach of these three systems in general outperforms each individual system on the least restrictive quality measurements by .03- .07 in F1; (3) Surprisingly, an ensemble approach of BANNER with any combinations of the other three approaches tends to degrade the performance by .08 - .11 in F1 when being evaluated on the I2B2 dataset; and finally, (4) our evaluation on the CRAFT dataset indicates that BANNER might not be effective at dealing with full-length journal articles due to its highly demanding training phase.

Poster No. 3

Application of DFT calculations of metal active site in bio-inorganic chemistry

---

Yang Ha

Bio-inorganic chemistry focus on the metal active site in metalloproteins. Density Functional Theory (DFT) calculations allow us to study the geometric structure and electronic structure around the metal active site, and provide insightful explanations on experimental data. What is more, using experimental calibrated parameters, DFT calculations can be used to simulate the whole reaction coordinate to find rate determine steps and key transition states, while perturbations around the metal site could guide mutations in future studies. All such calculations, coupled with experiments, can provide us a better understand on protein reaction mechanism. In this presentation, two examples will be demonstrated. Molybdenum enzymes are a family of enzymes that have a mononuclear Mo site and catalyze oxo transfer reactions. In DMSO reductase (DMSOr) family, oxo is transferred from substrate to Mo, while in sulfite oxidase (SO), oxo is transferred from Mo to the substrate. By comparing the two mechanisms, we found the fundamental difference in the oxo transfer reactions in general. Oxo transfer to substrate is initiated by electron transfer, while oxo transfer from substrate leads to electron transfer. This difference in reactivity is related to the difference in frontier molecular orbitals (FMO) of the metal–oxo and substrate–oxo bonds. [J. Am. Chem. Soc. 136: 9094, 2014] My next focus will be on interactions between iron-sulfur proteins and DNA. Both MutY and EndoIII are base excision repair proteins that contain a [Fe4S4] sites. They bind to mismatched DNA and electrochemical experiments showed that upon binding to DNA, the reduction potential of the [Fe4S4] couple decreases by 200 mV [J. Am. Chem. Soc. 132: 891, 2010]. Crystal structure show very little distortion of the cluster itself, and the ligand environment around the cluster remain similar upon DNA binding. Our DFT calculations show that it is the large amount of negative charge from the DNA that tunes the reduction potention.

Poster No. 4

Bringing the compendium of drug indications into the public domain

Daniel Himmelstein

The compendium of therapeutic drug indications is the outcome of many years of not only science and medicine, but also of trial and error. Despite its fundamental and universal relevance to public health, a standardized compendium of drug indications does not yet exist in the public domain. Here, we discuss our efforts to build this first-of-a-kind resource and its usefulness for the future of biomedical computation. Whether a drug treats a disease is a statement that can be judged taking into account several levels of evidence: pre-clinical tests, efficacy in a double-blind trial, and ultimately FDA approval. We have developed a framework that predicts therapeutic indications at a user-defined level of certainty. In this framework, predictions are initialized by a pre-assembled set of approved-indications (gold standard). Our machine learning method builds off the gold standard to predict an exhaustive set of indications. Specifically, predictions are accomplished by integrating information from drug labels (SIDER 2 database), literature curation (Comparative Toxicogenomics Database), literature co-occurrence (MEDLINE abstracts), text mining (DrugBank), and web scraping. Data integration occurs via controlled vocabularies to allow standardization of indications and enable reuse. Training the learning algorithm requires considerable manual curation, which we will facilitate through ThinkLab, an open science platform that rewards community participation. Biomedical computation, specifically systems pharmacology, requires standardized knowledge and is home to a plethora of applications that would benefit from our database. Once completed, we will release our resulting indication resource into the public domain. These indications will form the foundation of our attempt to use heterogeneous network edge prediction (http://het.io) for drug repurposing.

Poster No. 5

A disulfide bond is required for the transmission of forces through SUN-KASH complexes

---

Zeinab Jahed

Numerous cell functions including polarization, differentiation, division and migration rely on the ability of cells to endure mechanical forces generated by the cytoskeleton on the nucleus. The coupling of the cytoskeleton and nucleoskeleton is ultimately mediated by LINC complexes that are formed by a strong interaction between SUN domain and KASH domain proteins in the nuclear envelope. These complexes are mechanosensitive and essential for the transmission of forces between the cytoskeleton and nucleoskeleton and the progression of cellular mechanotransduction. In this study, we examine the effect of tension on the human SUN2-KASH2 complex, and show that it is remarkably stable under physiologically relevant tensile forces and large strains. However, a covalent disulfide bond between two highly conserved cysteine residues of SUN2 and KASH2 is crucial for the stability of this interaction and the transmission of forces through the complex.

Poster No. 6

Modeling current density using a flexible trans-cranial stimulator

---

Hardik Kabaria

Noninvasive brain stimulation techniques using low amplitude currents applied to the scalp via surface electrodes have been shown to reliably modulate cortical excitability and cognitive function. Transcranial direct current stimulation (tDCS) is under investigation as a treatment modality for multiple neurological conditions including depression, dementia and aphasia. It is gaining traction in cognitive enhancement as a way to improve working memory, the acquisition of motor skills and attention. Where the electrodes should be placed and what functional anatomy should be modulated to achieve the desired clinical effect remains unclear. Clinical effect is thought to be dependent on the current density at the desired anatomical location. It is influenced by voltage, electrode surface area and intervening anatomical structure. There have been significant advances towards computational modeling and simulations of tDCS. Computational methods, such as Finite Element Analysis, can be used to predict current density for a specific electrode montage. This amounts to solving a Laplacian of the scalar voltage field. In this talk we present the recent advances in the field of transcranial stimulation. We couple tDCS simulations with multiview geometry reconstruction method of computer vision to build personalized montages for patient specific geometry. We present preliminary results and simulations that are an important step towards building a personalized tDCS device.

Poster No. 7

A pooling-based approach to mapping genetic variants associated with DNA methylation

Irene Kaplow

DNA methylation is an epigenetic modification that plays a key role in gene regulation and is known to be influenced by nearby genetic variants. Previous studies have mapped genetic variants that are associated with DNA methylation at specific sites, but these have been limited to microarrays that cover less than 2% of the genome and cannot measure allele-specific methylation (ASM) [1], [2]. Other studies have performed whole-genome bisulfite sequencing on a few individuals, but these lack the statistical power needed to identify variants associated with DNA methylation [3]–[5]. We developed a novel approach in which bisulfite-treated DNA from many individuals is sequenced together in a single pool, resulting in a truly genome-wide map of DNA methylation generated by a single experiment (Figure 1). We first tested our approach in simulations and then applied it to bisulfite-sequencing data from a pool of 60 human lymphoblastoid cell lines (LCLs). Compared to methods that do not account for ASM, our approach increases statistical power to detect associations while sharply reducing cost, effort, and experimental variability. We tested almost twice as many CpGs as the largest microarray studies and identified over 2,000 genetic variants associated with DNA methylation. We found that these variants are highly enriched for associations with chromatin accessibility and CTCF binding but are less likely to be associated with traits indirectly linked to DNA, such as mRNA levels and disease phenotypes. In summary, our approach allows genome-wide mapping of genetic variants associated with DNA methylation in any tissue of any species, without the need for individual-level genotype or methylation data. We anticipate that this approach will enable us to identify many more variants that affect gene expression through methylation, providing novel insights into the role of methylation in transcriptional regulation.

Poster No. 8

A Text-Driven Approach for Constructing a Drug Metabolite Database

---

Aaron Kravitz

The active metabolites of drugs play a major role in how drugs interact with genes and other small molecules within a cell. Much of the information surrounding a drug's metabolic pathways resides in the literature. Unfortunately, the literature is growing exponentially, rendering a manual search for these metabolites infeasible. Deepdive is a trained system for extracting information from a variety of sources, including text. We aimed to build a tool for automated text extraction of metabolites using Deepdive. We applied the extractor to 600,000 PubMed abstracts and 1,000 full publications that had previously been curated for drug-gene relationships by PharmGKB. While we are currently tuning the system, the results are promising, with an estimated precision of 79% and an estimated recall of 79.7%. When refined, we will apply the system to a larger corpus of full text documents in order to construct a publicly accessible database of metabolites published in the literature. This database will facilitate the discovery of gene-drug relationships by enabling direct comparisons between the active drug metabolites, rather than the drug's original form.

Poster No. 9

Drug Repurposing for Gastric Cancer using Genomics Data

Shruti Marwaha

Background: Gastric cancer is the fourth most common malignancy and second leading cause of cancer-related mortality worldwide. There is an urgent need for development of better drugs for gastric cancer. However, a new drug development pipeline could take 10 years and cost upwards of several billion dollars. One approach to reduce the high cost and time associated with drug discovery is through repurposing of existing drugs approved for other diseases. Here, we use genomics data and Broad Institute's Connectivity Map to identify FDA approved drugs, or nondrug bioactive compounds, that can be potentially used for treatment of gastric cancer.

Method: Disease signature was derived by analyzing microarray data from three independent gastric cancer datasets obtained from the GEO database. The derived signature from each dataset was matched against Connectivity Map to identify candidate therapeutic compounds. Similarly, microarray data from gastric tumors resistant to cisplatin-fluorouracil (CF) therapy were compared with tumors sensitive to CF to derive drug resistance signature and to identify compounds which can potentially reverse CF resistance in these tumors.

Results and Conclusion: Our bioinformatics approach predicted vorinostat and trichostatin-A, HDAC (Histone Deactylase) inhibitors as lead potential compounds for treatment of gastric cancer. Our analysis suggests that a number of small molecules, including trichostatin-A and alpha-estradiol can potentially reverse the sensitivity of CF resistant cells. The approach we outline is sensitive to the expression profile of the disease. Analysis of gene expression profiles, combined with the connectivity map can help to identify and prioritize existing drugs or compounds with potential therapeutic value for the disease of interest.

Poster No. 10

The Inference of Sex-Biased Demography is Affected by Population Size Changes

Shaila Musharoff

Sex-biased demography ("sex-bias"), such as a male-biased migration or a female-biased effective population size, can be inferred by comparing X chromosomal to autosomal genetic variation. The conventional sex-bias estimator Q is a function of current genetic variation and assumes a constant population size. However, due to differences in copy number and effective size, the X chromosome and autosomes recover genetic variation at different rates following a population size change. As a result, inference based on Q for a population which has changed in size can lead to incorrect conclusions about sex-bias and the female fraction of the effective population size (pF). We developed a novel nested likelihood ratio test for sex-bias which explicitly models population size changes and derived an estimator of pF accounting for these changes. Our method estimates demographic model parameters from autosomal genomic data and tests a series of nested sex-bias models on X chromosome data. When applied to simulated data, our test has more power to detect sex-bias from unlinked or partially linked sites than the conventional test based on Q for several demographic scenarios relevant to human history. For example, for simulated recent growth, our test has 15% more power to detect a mild bias and up to 40% more power to detect an extreme bias, and our estimator of pF recovers the true parameter whereas the conventional statistic does not. We applied our methodology to human genome-wide sequencing data from the Thousand Genomes Project and found evidence for a male bias in Africans and Europeans. Using the naive Q-based estimator which assumes a constant population size for an African population (YRI) gives an estimate of 0.359 for the proportion of breeding females while our method using a more realistic demographic model including an old growth event gives an estimate of 0.465. For a population of European descent (CEU), the naive estimator gives an unrealistic estimate of 0.080 for the proportion of females, while our method using an appropriate demographic model gives an estimate of 0.435. These estimates are consistent with a slight male bias in YRI and CEU; the excess bias in CEU could be due to a male-biased migration out of Africa. Our findings underscore the importance of modeling size changes when estimating sex-bias in populations and highlight the pervasive role of sex-bias in human history.

Poster No. 11

Genome-wide Prediction of Splice Sites using Maximal Dependence Decomposition

---

Santrupti Nerli

Eukaryotic mRNA splicing is a key process involved in gene expression. It plays a vital role in producing functional proteins. It is carried out by a complex called spliceosome that ensures introns are removed and exons are joined correctly prior to translation. Spliceosome spots specific sequence motifs called splice sites within mRNA sequence, two of which are the 5’ and 3’ sites that flank all introns. Mutations within splice sites can disrupt normal splicing process leading to malignant diseases. The aim of this work is to predict splice sites in a human genome using higher order position weight matrix (PWM). PWM is a popular computational method used to represent splice sites or any sequence motif. It is constructed by capturing the occurrences of nucleotides at each position of aligned sequences. Simple PWMs performed fairly well in classifying known 5' and 3' splice sites and predicting cryptic splice sites in human genes. However, they make a strong assumption of independency between adjacent and non-adjacent nucleotide positions. Therefore, we developed a method that will identify statistically significant splice sites and incorporated maximal dependence decomposition algorithm (MDD) to achieve it. MDD operates by constructing a classification tree on aligned sequences. It takes into account the interdependency between nucleotide positions unlike simple PWMs and degree of dependencies is evaluated by chi-squared tests. The leaves of the classification tree represent distinct groups with their own PWM and are used to cluster unknown test sequences and classify unknown splice sites. We performed 10-fold cross validation of the MDD algorithm for 5' and 3' authentic human splice sites from the HS3D database and observed area under the Receiver Operating Characteristic curve (ROC) to be 0.97 and 0.93, respectively. Similarly, we performed classification of putative 5' and 3' cryptic splice sites in the breast cancer type 1 susceptibility protein (BRCA1) and observed area under ROC to be 0.93 and 0.81. Our algorithm is fast; scoring each human chromosome takes on average 5 minutes. We are currently working on implementing a graphical-user interface for our method to facilitate data input and output as well as visualize the predicted splice sites in whole genome sequences. The tool can easily be extended to rapidly predict any DNA sequence motif.

Poster No. 12

Pathways analysis of rare, high-impact variants in mothers of children with autism

Chloe O'Connell

Background: Events during pregnancy, such as maternal infection, stressful life events, and even exposure to severe storms and maternal residential proximity to areas of agricultural pesticide application, have been linked to increased risk of giving birth to a child with autism. However, the effect of maternal genotype and its interaction with these events is unknown. The few existing studies on the topic have searched for single alleles that impact autism risk in the fetus, rather than the pathways to which these alleles belong. The goal of this study was to determine whether genes involved in oxidative stress pathways are enriched for high-impact variants in the mothers of children with autism.

Methods: Filtering variants to only those found in <1% of the samples in the 1000 genomes project, we developed a list of rare alleles in 239 unique mothers of children with autism, within 206 exomes and 33 whole genomes. We further filtered the rare variants via SnpEff predicted impact, selecting only the variants that were predicted to be "high-impact". We then compiled a list of 591 genes containing rare, high-impact variants in mothers of children with autism, and tested for pathway enrichment using DAVID.

Results: We discovered 4 pathways enriched for variants in the mothers' genomes in the Kegg and Reactome pathway databases. Of note, the Kegg pathway "Metabolism of xenobiotics by cytochrome P450" was significantly enriched for variants in the mothers of children with autism (p = .0216*), but not the fathers (p = 1.00*) or the probands themselves (p = 1.00)*. The most significantly enriched Reactome pathway in the mothers, "Biological Oxidations" (p = 3.36E-04*), also was not enriched in the fathers of trios (p = 1.00*).

Discussion: Given their roles in protection from oxidative stress, particularly xenobiotic-induced stress, the genes within the pathways found to be enriched within autism mothers are worth further investigation. Furthermore, our results are consistent with the hypothesis that toxic and oxidative stress may impact autism risk during fetal development. More investigation is required to determine the nature of these potentially teratogenic variants, as well as the nature of the gene-environment interactions they may influence.

Poster No. 13

Phospho Atlas – the resource to explore human phospho-reactome

Aleksandra Olow

Introduction: Phosphorylation cascades intimately regulate cells' behavior. Understanding how kinase networks can circumvent therapeutic interventions is the challenge researchers and clinicians face. We created a resource to gather the current knowledge of known phosphorylation events called the "Phospho Atlas". This resource was curated from 38 public databases using computational methods that integrate non-prediction-based, experimentally validated data. We were able to not only catalogue the kinase-substrate interactions but also elucidate the 2,492 human biological kinase target peptide sequences, and extract information on the known variants affecting these target sites.

Methods: An identification of exact, verified phosphorylatable residues, and the related biological peptide target sequences is achieved through the extraction, harmonization and classification of public protein records. The resulting database informs of all known kinase – target site – substrate phosphorylation events. Furthermore, the extracted target peptide sequences are mapped back to their genomic location (GENCODE GRCh37/HG19) providing peptide locations for all possible substrate isoforms. The variants within these locations are then found by overlapping the variant GRCh37 genomic locations acquired from a complete COSMIC database with our peptide genomic locations. Peptide targets with highest number of reported variants are identified and mapped back to pinpoint most affected kinase-substrate network.

Results: The Phospho Atlas contains 3,641 complete entries, with known both kinase substrates and their target sites. Additional 1,117 entries represent kinases with known substrate targets but not the target peptide sequences. Moreover, there are 277 kinases without known substrates. In total, the database represents 776 unique kinases, 1,276 substrate proteins and 2,492 peptide target sequences. The most represented kinase protein families are tyrosine kinases, CAMK and AGC families. Within the identified peptide target sequences, there are 1,031 peptides without any reported variants. The coding variants reported in the remaining peptides create an additional 2896 variant peptides, and they are most commonly associated with adenocarcinomas (19%). Peptides with highest number of reported coding variants belong to known proteins famously implicated in cancers, including TP53, BRAF, EGFR and PTEN.

Conclusion: The Phospho Atlas is a novel resource encompassing the current functional knowledge of phospho-circuits; and it provides valuable insight into areas requiring further investigation.

Poster No. 14

structSSI: Simultaneous and Selective Inference for Grouped or Hierarchically Structured Data

---

Kris Sankaran

The R package structSSI provides an accessible implementation of two recently developed simultaneous and selective inference techniques: the group Benjamini-Hochberg and hierarchical false discovery rate procedures. Unlike many multiple testing schemes, these methods specifically incorporate existing information about the grouped or hierarchical dependence between hypotheses under consideration while controlling the false discovery rate. Doing so increases statistical power and interpretability. Furthermore, these procedures provide novel approaches to the central problem of encoding complex dependency between hypotheses. We briefly describe the group Benjamini-Hochberg and hierarchical false discovery rate procedures and then illustrate them using two examples, one a measure of ecological microbial abundances and the other a global temperature time series. For both procedures, we detail the steps associated with the analysis of these particular data sets, including establishing the dependence structures, performing the test, and interpreting the results. These steps are encapsulated by R functions, and we explain their applicability to general data sets.

Poster No. 15

High-resolution nucleosome positioning from ATAC-seq chromatin accessibility data

Alicia Schep

Chromatin accessibility and nucleosome positioning are important determinants of transcriptional regulation, as nucleosomes compete with transcription factors for binding of DNA. ATAC-Seq (Assay for transposase-accessible chromatin using sequencing) profiles chromatin accessibility genome wide, and fragmentation patterns of DNA generated by this method provide information about nucleosome positioning in regulatory regions. By applying ATAC-Seq to Saccharomyces cerevisiae, for which a chemical map of nucleosomes at base-pair resolution is available, we observe that ATAC-Seq signal around nucleosomes is highly structured. We have developed a method—NucleoATAC—that uses this highly structured pattern to infer nucleosome positions at high-resolution within accessible regions of the genome. Unlike methods developed for MNase-Seq, NucleoATAC can resolve the 10-bp rotational positioning of nucleosomes. Furthermore, as ATAC-Seq assays accessible chromatin, high-confidence nucleosome position calls are concentrated within regulatory regions of the genome, nucleosome calls can be assigned occupancy levels, and the precise locations of nucleosome-free regions (NFRs) can be inferred. By applying NucleoATAC to Saccharomyces cerevisiae, Schizosaccharomyces pombe, and the human GM12878 lymphoblastoid cell line, we observe different nucleosome sequence preferences and chromatin architecture between the species. We show that for all three species, nucleosome-free regions uniformly span less than one nucleosome. Additionally, we show that in contrast to S. cerevisiae and human transcription start sites the majority of transcription start sites in S. pombe are not characterized by a stable nucleosome free region. Future studies will apply NucleoATAC towards understanding nucleosome positioning and occupancy changes in dynamic systems.

Poster No. 16

Bayesian Inference of Geometric Deformations Associated to Lower Back Pain

---

Christof Seiler

Latest studies in developed countries show that about 80% of the population suffer from acute and about 5-10% from permanent lower back pain. An early diagnosis is crucial to reduce the individual suffering of patients and to lower the economic burden on the society as a whole. We use Computational Anatomy (CA) to address this important problem. The aim of CA is to describe human anatomy not by what it looks like but by how it deforms. The task of finding deformations between patient images can be understood as a form of spatial normalization. In this talk/poster, we approach CA from a Bayesian perspective. First, we setup a statistical model of the anatomy, we construct a prior on admissible deformations, link it to the data through a likelihood term, and do inference by sampling from the posterior of deformations. We sample from this posterior using Hamiltonian Monte Carlo and then use Bayesian logistic regression to find geometric features that distinguish between back and abdominal pain in 400 patients. This is joint work with Susan Holmes, Simon Rubinstein-Salzedo, Xavier Pennec, and Nicolas Bronsard.

Poster No. 17

Regulation of vinculin activation through α-actinin binding

Hengameh Shams

Cells associate with their microenvironment through integrin-mediated adhesions that are highly engaged with the actin cytoskeleton, in order to control cell shape and motility. Powerful intracellular machinery underlies transducing mechanical stimuli to a biochemical cascade inside the cell. The vinculin molecule is known to be a critical factor in mediating force transmission in cell-substrate adhesion and giving rise to force-dependent adhesion maturation in many cell types. Although vinculin is suggested to contribute to cytoskeletal integrity in cells, its specific role in adhesion formation is not yet clear. One hypothesis is that α-actinin, an actin cross-linker, plays an important in regulating vinculin function. We investigated the underlying molecular mechanisms of the interplay between α-actinin and vinculin.

Poster No. 18

Comparing chronic kidney disease profiles identified using diagnostic histories and clustered clinical data with an established genetic risk model

---

Craig Smail

There are around 26 million Americans currently diagnosed with chronic kidney disease, with associated annual health care costs estimated at 42.5 billion dollars. Amongst those affected, Blacks are approximately twice as likely as Whites to progress from chronic kidney disease to end-stage renal disease – irreversible decline in kidney function below that which is necessary for normal life. The results of recent genome-wide association studies suggest that Blacks possessing two variants in exon 5 of the apolipoprotein L, 1 (ApoL1) gene (known as G1 and G2 and conferring protection against the sleeping sickness parasite Trypanosoma brucei rhodesiense) are at an increased risk for a subset of hypertension-related chronic kidney disease and a faster rate of decline in kidney function. These variants are found almost exclusively in those of recent West African ancestry. It would be helpful to identify putative ApoL1 high-risk patients in the clinic; such an approach could be used to prioritize genotyping/sequencing test orders and intensive treatment interventions. It was hypothesized that patients possessing ApoL1 risk variants would share distinct clinical profiles, as defined by laboratory data extracted from the Electronic Medical Record of study sites. Furthermore, diagnosis lists often contain ICD-9 codes specific to more-severe chronic kidney disease – often added as a clinical diagnosis of exclusion when suspected by clinicians - which could be leveraged to help identify putative ApoL1 high-risk patients. The method presented here (1.) uses hierarchical clustering of a standardized matrix of laboratory data to identify clinically similar chronic kidney disease patients, (2.) ranks resulting clusters by the frequency of patients with ICD-9 codes indicative of more-severe chronic kidney disease, and (3.) constructs a control cohort matched on several clinical and demographic variables of patients appearing in the top-ranked cluster from step (2.). Similar differences in rates of decline in kidney function were observed in putative ApoL1 high- and low-risk Black patients from primary and validation sites, but were equivalent when applying the method to a control cohort of White patients. The results were not explained by baseline differences in kidney function, time since diagnosis of chronic kidney disease, blood pressure, age, gender, or diagnostic history of hypertension. Similar differences in rate of annual decline were observed in a large-scale study of Black patients with known ApoL1 genotype.

Poster No. 19

Characterization of mRNA quality control mechanism through an agent-based modeling framework

---

Mohammad Soheilypour

Following transcription, nascent messenger ribonucleic acids (mRNAs) will be transported to the cytoplasm to transfer genetic information and direct synthesis of functional proteins. These transcripts require successfully passing several checkpoints to be considered as an export-competent messenger ribonucleoprotein (mRNP). Multiple co-transcriptionally occurring processes, applied on precursor mRNA (pre-mRNA), are followed by loading of several key proteins, i.e. RNA-binding proteins (RBPs), along the length of pre-mRNA, eventually forming a mature mRNA prepared for efficient export through the nuclear pore complex (NPC). This process is quality controlled by an evolutionary conserved, highly efficient mechanism in eukaryotic cells, called mRNA quality control mechanism. Major proteins involved in the quality control of mRNAs can be categorized into two groups with respect to their nuclear localization, namely RBPs that are bound to mRNA itself, and nuclear basket-associated proteins. We have recently studied dynamics of mRNA export through NPC (Azimi et al., 2014). Here, we extend our previous study by accounting for the quality control mechanism in the model. In this study, we seek to explore the role of cross-interactions between three major role-players in mRNA export and quality control, namely RBPs, export factor molecules, and quality control molecules in the nuclear basket. We are primarily interested to understand how unspliced mRNAs are distinguished and retained inside the nucleus while normal mRNAs are efficiently exported. Method: Agent-based modeling (ABM) has emerged as a highly promising computational biology approach to efficiently simulate the spatiotemporal interactions between multiple independent entities, i.e. agents, with the objective of assessing their individual effect on the overall system and confirming/predicting subsequent emergent phenomena. ABMs consist of a collection of agents with governing rules that dictate local behavior and interactions with adjacent agents, resulting in a complex emergent behavior that may not be obvious from the individual rules. Following the pre-defined environmental conditions, the agents move and locally interact with adjacent agents at each time step. In this study, we build upon our recently developed ABM platform (Azimi et al., 2013 - Azimi et al., 2014) to shed light on dynamics of mRNA export and quality control.

References:

Azimi, M. et al. (2013). Higher nucleoporin-Importinβ affinity at the nuclear basket increases nucleocytoplasmic import. PloS One 8, e81741.

Azimi, M. et al. (2014). An agent-based model for mRNA export through the nuclear pore complex. Mol. Biol. Cell.

Poster No. 20

Comparison of Bioconductor Packages edgeR and DESeq for Analyzing Differential Expression from RNA-Sequencing Experiments

Kelly Speth

Next-generation sequencing (NGS) technologies have greatly improved the landscape of genomic research and have, within the past decade, served to reduce dramatically the cost and increase the data output potential. One of the specific applications that utilizes NGS methods is RNA-sequencing (RNA-Seq), which has a number of uses, one of which is the determination of differential (gene) expression (DE) among two or more RNA samples. Although the cost of NGS technology continues to decrease and open the field of genomic research to a broader scope of scientists, the cost of large-scale experiments in which sufficient biological and technical replicates are used and for which standard, large-sample statistical methods are appropriate are still cost prohibitive to all but the most well-funded industry or government enterprises. Thus, academic scientists, while prolifically conducting genomic research on DE, rely on experimental designs that include few biological and/or technical replicates. Additionally, several other experimental conditions, for example, differences among sample library sizes and the conduct of multiple hypothesis tests, confound the analysis. As a consequence, new analytic tools were needed to account for these adjusted model assumptions. R/Bioconductor supports a multitude of statistical analysis packages designed for the analysis of DE. These packages include edgeR, DESeq, DESeq2, NBPSeq, baySeq, and ShrinkSeq, among others—each of which differs in one or more aspects, including the way it handles normalization, parameter estimation, hypothesis testing, and correction for multiple testing. Whereas different analysis methods are not at all unusual, a complicating factor for the analysis of DE is the fact that, given the same assumptions, the analysis packages produce different results—often significantly. This project explores the analytic differences and performance among several R/Bioconductor packages, including edgeR and DESeq. To conduct this analysis, a count dataset was simulated using model parameters estimated from a real RNA-seq dataset. Count data for the treatment condition were then simulated to reflect DE (i.e., either up- or down-regulated) for a pre-specified number of genes. The simulated dataset—for which DE status was known—was then challenged using each of the analysis packages. The results and performance, including Type I and Type II error rates, were compared across packages.

Poster No. 21

COmbined Mapping of Multiple clUsteriNg ALgorithms (COMMUNAL): A Robust Method for Selection of Cluster Number, K

Timothy Sweeney

Multiple unsupervised clustering techniques exist, with multiple metrics by which the optimal number of clusters in a dataset (K) may be judged. Here we report a new method, COmbined Mapping of Multiple clUsteriNg ALgorithms (COMMUNAL), that combines information from multiple clustering algorithms and multiple cluster validity metrics to enhance the robustness of a given assessment of K. We further show that the stability of a dataset's optimal K can be mapped by sequentially adding variables to the clustering algorithms. We first determined an optimized set of clustering algorithms and validity metrics using gene expression datasets and simulated data. We then ran the optimized set of algorithms and metrics over input datasets multiple times with sequentially more variables added in, and plotted K versus number of variables included versus the mean normalized validity metric results to produce a 3D map of cluster optimality. This process was repeated for simulated data with known K, as well as all cancers in the TCGA with microarray gene expression data. The simulated data were called correctly, even in the presence of 1:1 informative:noisy variables. The COMMUNAL optimal K matched TCGA optimal K for NMF consensus clustering for three cancer types, and revealed new insights about stability of cluster number assignment over a range of both number of included genes and K for all cancers tested. COMMUNAL combines the signals from multiple clustering algorithms and metrics to improve the signal-to-noise ratio and increase the robustness of the selection of cluster number K. Furthermore, by mapping these calls over a gradually increasing number of variables for a given dataset, the stability of cluster optimality to additional information can be assessed. COMMUNAL may be a useful tool for determining K in complex datasets, and has been submitted as an R package on CRAN.

Poster No. 22

Statistical algorithms for improved detection of linear and circular splice events from RNA-Seq data

---

Linda Szabo

The pervasive expression of circular RNA from protein coding loci is a recently discovered feature of many eukaryotic gene expression programs. Computational methods to discover and quantify circular RNA from RNA-Seq data are essential to the study of the mechanisms of circular RNA biogenesis and potential functional roles they may play. Here, we present a new statistical algorithm that increases the sensitivity and specificity of circular RNA detection by discovering and quantifying circular and linear RNA splicing events at both annotated exon boundaries and in un-annotated regions of the genome. Unlike previous approaches that rely on heuristics such as read count and exon homology to determine confidence in prediction of circular RNA expression, our algorithm takes a statistical approach and can detect circular RNAs with high confidence even when present at very low levels. Applying our algorithm to RNA-Seq data available from Sequence Read Archive (SRA) for cells treated with RNase-R to enrich for circular RNA, we not only demonstrate improved accuracy, but also show that inferences about the regulation of circular RNA need to be revisited due to biases in the algorithms used for these genome-wide analyses. Additionally, we applied our algorithm to RNA-Seq data generated by the ENCODE consortium from polyA+ and polyA- fractions of 18 cell types and found several striking examples of highly expressed putative circular RNAs known to be false positives. Such results are heuristically post-filtered by other algorithms, but are automatically flagged by our algorithm as artifacts. In work currently under review, we have used this algorithm to analyze a panel of human fetal tissues sampled at several developmental time points. We discovered regions of the fetal brain, such as the frontal cortex, with marked enrichment for genes where the circular isoform is dominant and found general induction of circular RNAs in many tissues, including the heart and lung, during human fetal development. Moving forward, we are building upon the concepts implemented in the original version of our algorithm for circular RNA detection to develop a statistical algorithm that can be used to quantify known and novel circular and linear isoforms in genomes that differ significantly from the reference, such as cancer genomes.

Poster No. 23

smallWig: Compression on RNA-seq Expression Data

Zhiying Wang

The start of the 21st century has witnessed a dramatic decrease in the cost of DNA/RNA sequencing. As a result, peta bytes of genomic data are generated per year. However, current storage solutions and prices do not scale appropriately with such massive data surges. It is therefore of great importance to develop efficient and task-oriented compression methods. One such genomic information is expression data, representing the amount of RNA molecules transcribed from every nucleotide location in the genome. Except for the classical BigWig compression method, based on gzip, and the very recent work termed cWig, no highly efficient compression methods are currently known. We developed a specialized lossless approach to compression of RNA-seq expression and whole transcriptome data, combining various source coding methods so as to achieve fast encoding and compact data representation. We also provided for flexible random access that enables fast visualization of subsequences of expression data, where the access time and the compression rate can be tuned by the application. For archiving purposes, more elaborate context-based compression techniques were implemented as well, which further reduced the compressed file sizes.

Poster No. 24

FEATURE-Viz: A Web-Based Visualization System for Comparing Pairs of Protein Pocket-Ligand Interactions

---

Michael Wong

Protein pocket-ligand interactions are still yet sufficiently characterized for fully-automated identification and comparison. Consequently, human expertise is still necessary for evaluation, demonstrating need for frontend applications for human visualization and interpretation. We present FEATUREViz, a pocket alignment visualization system developed in collaboration with the Stanford Helix group, as a front-end to PocketFEATURE, an algorithm developed by the Helix group that identifies novel proteindrug interactions by analyzing the similarity between pockets in proteins. We demonstrate the utility, including automated visual pocket alignment, publication-ready image generation, and workbooks, alignment states that are configurable, downloadable, and shareable.

Poster No. 25

Is Clinical Atrial Fibrillation Caused by Multiwavelet Reentry? Numerical Analysis of Data from Patients with Atrial Fibrillation

Junaid Zaman

Introduction: Multiwavelet reentry (MWR) is considered the mechanism of atrial fibrillation (AF), and is treated by lines to prevent wavelet meandering. We set out to test this mechanistic endpoint by using clinical measurements of action potential duration (APD) & conduction velocity (CV).

Methods: Clinical data were obtained in 41 AF patients (63±10y, LA 48±7 mm. Monophasic APD catheter pacing defined atrial APD rateresponse. Activation time to each of 64poles of a basket catheter (Boston Scientific) defined biatrial CV rateresponse. Computational studies used wavelength (=APD x CV) to determine the minimum size of reentry circuits at each rate in imaged atrial surface areas.

Results: Standardized LA ablation was analyzed, producing 4 compartments: left WACA, right WACA, posterior & anterior LA. A LA of diameter 5 cm has area > 65 cm2, for ≈16.25 cm2 compartment. The largest circuit sustainable in this compartment has diameter 4 cm, or wavepath 12.6 cm. In 41 patients, minimum APD was 110±21 ms which, for this wavepath, gives CV 113 cm/second. However, actual CV measured on electroanatomic shells produced slower maximum CV.

Conclusions. Physiological analysis of observed APD and CV in AF patients shows that often successful ablation lesion sets are unlikely to eliminate AF from MWR. Future studies should examine how commonly used ablation sets are successful, such as by interacting with localized sources, ganglionic plexi or other mechanisms as well as examine the effect of spiral wave dynamics through computational studies.

Poster No. 26

The Dynamic Interplay Between Organized and Disorganized Zones During Human Atrial fibrillation determines the Likelihood of Termination During Ablation

Tina Baykaner

Introduction: It is difficult to predict whether atrial fibrillation (AF) will terminate during ablation. We tested the hypothesis that the likelihood of AF termination is inversely related to the extent of electrical remodeling, by quantifying the ratio of organized ("drivers") to disordered ("fibrillatory") zones in patients undergoing ablation.

Methods: We recruited 28 patients with AF (age 60.1±9.3 years, 18% persistent). AF was recorded from a 64-pole basket in each atrium (Constellation). Maps were analyzed for organized sources, and fibrillatory disorder was quantified as the width of the Gaussian distribution of cycle lengths (full width half maximum, FWHM) during ablation.

Results: All patients showed sources (3.1±1.3 each, LA 61%) and fibrillatory disorder. Sources had smaller diameter in patients with persistent vs paroxysmal AF (0.20 vs 0.07, p=0.036), and were inversely related to fibrillatory disorder (p=0.003) that in turn reflected AF duration (p=0.033) and other comorbidities. During ablation, AF termination reflected FWHM (p=0.033) rather than number or location of sources. In the figure, A shows AF with 5 sources before ablation, B. Wide FWHM (disordered). C. AF organized by ablation just before termination; D. Narrow FWHM (organized).

Conclusions: Clinical AF shows organized sources, yet it is the extent of fibrillatory disorder that determines if AF will terminate by ablation. Future studies should define the influence of comorbidities on these dynamics and on long-term ablation outcomes.

BCATS 2015

Thank you!

We would like to express our appreciation in particular to Patrick More for being our administrative support.

Previous Organizers

Zoe June Assaf
Amy Goldberg
Kimberly McManus
Shaila Musharoff
Chris VanLang

Platinum Sponsors

Bio-X
Biomedical Informatics Training Program (BMI)
Stanford Center for Genomics and Personalized Medicine (SCGPM)
Stanford Center for Computational, Evolutionary and Human Genetics (CEHG)
Stanford Cancer Institute (SCI)

Gold Sponsors

Genentech
Stanford Computer Forum
Department of Bioengineering

Silver Sponsors

PharmGKB
Stanford Biophysics Program