2ND Bioinformatics and Genomics Retreat

September 16th and 17th, 2011
100 Life Sciences (Berg Auditorium)

Bioinformatics and Genomics (BG) Retreat

BG Retreat is an annual Networking Conference for Students, Post-Docs, Faculty, and Professionals in the Bioinformatics and Genomics fields. At the retreat, you will have a chance to find out various exciting BG-related research projects at Penn State. We also want to introduce new students to the BG programs.

The visionary talk, panel discussion, four sessions and keynote talk will offer time for us to share ideas, successes, and challenges with each other on current and future BG research. At the poster sessions, students are given ample opportunities to present their research project and discuss their data with peer students and faculty members. Over breaks, everyone is highly encouraged to network and interact with each other in the BG related fields.

Organized by:

BG Retreat Committee
Nathaniel Cannon
Arkarachai Fungtammasan
Jihye Park

Special thanks to:

All the student volunteers!

BG Retreat 2011 Schedule

September 16th, Friday

Time	Events	Where
4:00-4:30	**Registration**	1st Floor, LSB
4:30-4:35	**Opening Remarks**	Berg Auditorium
4:35-5:35	**What will it take for genomics to do good?** *Can genomics and bioinformatics really help human health?* (Dr. Ross Hardison) *Comparative and functional genomics of plants to enhance food security* (Dr. Claude dePamphilis) *Bioinformatics skills of the present and future* (Dr. Istvan Albert)	Berg Auditorium
5:35-5:50	**Break / Collect questions**	1st Floor, LSB
5:50-6:35	**Panel Discussion** Dr. Ross Hardison, Dr. Clause dePamphilis, and Dr. Istvan Albert	Berg Auditorium
6:35-8:05	**Dinner and Poster Session I**	Ground Floor, LSB

BG Retreat 2011 Schedule
September 17th, Saturday

Time	Events	Where
9:00-10:30	**Session I** *Curiosity, serendipity, genomics, and insight into biology and human history* (Dr. Keith Cheng) *Polygenic pleiotropy in craniofacial variation: association mapping of 15 craniofacial traits in an F34 mouse population* (Nathaniel Cannon) *A comparison study of methods to detect complex, multilocus genetic and GxE interactions* (Dr. David Miller) *A biologically informed method for detecting associations with rare variants* (Carrie Buchanan) *Systems mapping: how to design crop ideotypes through plant economy* (Dr. Rongling Wu)	Berg Auditorium
10:30-11:00	**Coffee Break**	1st Floor, LSB
11:00-12:00	**Session II** *A look at the Earth microbiome project* (Dr. Mary Ann Bruns) *Combinatorial use of poly-A/T tracts in organizing genes, nucleosomes and the transcription machinery in Dictyostelium* (Gue Su Chang) *Assessing vaccination sentiments with online social media: implications for infectious disease dynamics and control* (Dr. Marcel Salathe)	Berg Auditorium
12:00-1:00	**Keynote Speaker** *Human evolution revealed by extinct hominin genomes* (Dr. Richard Edward Green)	Berg Auditorium
1:00-2:30	**Lunch and Poster Session II**	Ground Floor, LSB
2:30-4:00	**Session III** *Challenges and opportunities for eco-evolutionary genomics with trees* (Dr. John Carlson) *GCD: a new method for identifying gene families from genomic sequence data* (Dr. Zhenguo Zhang) *Evolution of genomic rearrangements in Drosophila* (Dr. Stephen Schaeffer) *A fine line between friend and foe: comparative genomics of the Fungi* (Josh Herr) *Standing genotypic variation is basis for coral survival of temperature stress* (Dr. Iliana Baums)	Berg Auditorium
4:00-4:30	**Coffee Break**	1st Floor, LSB
4:30-6:00	**Session IV** *Translational microbial pathogenomics* (Dr. Vivek Kapur) *Host microflora and respiratory pathogen interactions during infection: a multidisciplinary approach* (Olivier Rolin) *Characterizing the ciprofloxacin-inducible bacteriophage populations of Escherichia coli O157:H7 by 454 sequencing* (Dr. Edward Dudley) *The genomic structure and transcriptome of the male-specific region in the bovine Y chromosome* (Ti-Cheng Chang) *Cancer epigenetics* (Dr. Sagarika Kanjilal)	Berg Auditorium
6:00-6:05	**Concluding Remarks**	Berg Auditorium
6:05-	**Students Evening Social**	Ground Floor, LSB

Abstracts

"What will it take for genomics to do good?"

Can genomics and bioinformatics really help human health?

Dr. Ross Hardison

Since the inception of biotechnology in the 1970's and continuing through the growth of genomics to the current interest in personal genomes, many claims have been made about the utility of these technologies to improve human health. Despite some overly-hyped advances and even some ill-advised, premature applications of the technologies, much excitement surrounds the prospects for "genomic medicine". I plan an anecdotal survey of successes and failures, with some thoughts for the future.

Comparative and functional genomics of plants to enhance food security

Dr. Claude dePamphilis

An estimated 2 billion people worldwide suffer from some degree of food insecurity ranging from intermittent hunger to starvation. While the underlying causes for this situation are numerous, biotic and abiotic stresses are major contributors. One of the main biotic sources of crop loss in marginal habitats in Africa, the Middle East, and other regions of the world, is parasitic plants. Parasitic plants have the ability to grow into a host plant to extract water, minerals, and carbon compounds, and can have devastating effects on agricultural productivity. Traditional means of breeding resistant crops have almost completely failed due to the complexity of host-parasite interactions and substantial polymorphism in the parasite population. I will focus on a massive transcriptome sequencing project - the Parasitic Plant Genome Project - in which we are using laser capture microscopy, next generation transcriptome sequencing, and bioinformatic analysis to identify genes that are critical to parasitic plant function, how this genomic information can be applied to better understand how parasites function, and experimental approaches to apply this knowledge to break the parasite-host cycle.

Bioinformatics skills of the present and future

Dr. Istvan Albert

Bioinformatics is a nascent field with increasingly great contributions towards improving the quality of human life and society. Yet even seasoned practitioners may find it difficult to identify what makes a good bioinformatician and have even less advice on how to become a successful one. In this talk I will provide my own perspective on what I believe to be the proper mentality, essential skills and vitally important knowledge that successful and productive scientists of this field must possess.

Abstracts

Session I

Curiosity, serendipity, genomics, and insight into biology and human history

Dr. Keith Cheng

The purpose of graduate school is to optimize one's ability to think about science. Since one can work harder when having fun, and I am having a blast, I will share a story about my science so that you can have more fun with yours. Curiosity about biology and cancer, combined with a fascination with the remarkable insights into biology provided by microscopy and genetics, led to the use of zebrafish for the study of cancer. Persistence yielded insights into cancer whose keys still remain mysteries, but exciting outcomes have come from indulging in a passion for gadgets and computers, and a purposeful appreciation of the gifts of serendipity. I offer an abbreviated survey of those insights, with most yet to come. I have progress on three fun projects to share. First, an *A111T c*oding mutation at rs1426654 in *SLC24A5* characterizes humans of European ancestry. Based on a 150kb region of diminished variation in Europeans, we determined the haplotypes in this region across the globe, and found many precursor haplotypes, including two that recombined to result in the precursor chromosome upon which the mutation occurred. We have been able to ascertain that this mutation occurred in the Middle East between 40-50kya, after the migrations that populated East Asia vs. Western Eurasia. Second, we have shown that we can use the zebrafish to test candidate human polymorphisms from Genome Wide Association Studies for human traits and disease susceptibilities. Finally, since zebrafish is the only well-developed genetic vertebrate small enough for imaging the whole animal at cell resolutions, we are using it to develop the field of computational phenomics - in which quantitative morphometric and other quantitative analyses are linked to genes and chemicals in the form of phenotypic signatures, to gain insight into gene functions as they affect the morphology of *all* tissues, and to facilitate drug development and urgent issues in environmental toxicology.

Polygenic pleiotropy in craniofacial variation: association mapping of 15 craniofacial traits in an F34 mouse population

Nathaniel Cannon

The phenotype is generally thought to be the effect of genotypic causes in environmental contexts, with variation in the genotype being inherently linked to variation in the trait. Phenotypic traits are often viewed as modular, being integrally connected through the developmental history of the individual. Studies seeking to identify the underlying genetic causes of these developmentally complex phenotypes traditionally look for signals of single genes with statistically significant associations with the variation in a single quantitative measure.

Here we present preliminary results from a study in an intercrossed mouse population that estimates the association of genotypes with several measures between a subset of landmarks representative of craniofacial morphology. We have identified several significant marker-trait associations, some of which are present in multiple measures. Pair-wise comparison of the different traits' mapping results identifies several regions of the genome that suggest the presence of many genes affecting multiple measures in small ways. These results are indicative

of a polygenic pleiotropic genetic architecture underlying the development of craniofacial anatomy and its variation in the adult mouse.

A Comparison Study of Methods to Detect Complex, Multilocus Genetic and GxE Interactions

Dr. David Miller

Interactions among genetic loci are believed to play an important role in disease risk. While many methods have been proposed for detecting such interactions, their relative performance remains largely unclear, mainly because different data sources, detection performance criteria, and experimental protocols were used in the papers introducing these methods and in subsequent studies. Moreover, there have been very few studies strictly focused on comparison of existing methods. Given the importance of detecting gene-gene and gene-environment interactions, a rigorous, comprehensive comparison of performance and limitations of available interaction detection methods is warranted.

We report a comparison of seven representative methods, of which six were specifically designed to detect interactions among single nucleotide polymorphisms (SNPs), with the last a popular main-effect testing method used as a baseline for performance evaluation. The selected methods were compared on a large number of simulated datasets, each, consistent with complex disease models, embedding *multiple* sets of interacting SNPs, under different interaction models. The assessment criteria included several relevant detection power measures, family-wise type I error rate, and computational complexity. The experimental results show that while some SNPs in interactions with strong effects are successfully detected, most of the methods miss many interacting SNPs at an acceptable rate of false positives. Moreover, the P-value significance assessment, used by some of the methods to fix the type I error, is quite conservative, which further limits their detection power. As expected, power varies for different models and as a function of penetrance, MAF, and marginal effects. Analytical relationships between power and these factors are derived, which support and help explain the experimental results.

A Biologically Informed Method for Detecting Associations with Rare Variants

Carrie Buchanan

With the recent flood of genome sequence data, there has been increasing interest in rare variants and methods to detect their association to disease. Many of these methods are collapsing strategies, which bin based on allele frequency and functional association; but at this point, most have been limited to candidate gene studies. We propose a novel method to collapse rare variants based on incorporating biological information. This is a useful collapsing strategy because it can be expanded to whole-genome data and can be used as a framework to identify gene-gene (GxG) or gene-environment (GxE) interactions. There is also potential to integrate large complex data sets (i.e. expression profiles, metabolomics, etc) with sequence data into future analyses. This paper introduces the functionality of BioBin, a biologically informed method to collapse rare variants and detect associations with a particular phenotype. We tested BioBin using low coverage data from the 1000 Genomes Project and discovered appropriate binning characteristics based on what one might expect given the size of the gene. We also tested BioBin using the pilot targeted exome data from 1000 Genomes Project. We used biologically informed binning and differences in minor allele frequencies as a means to distinguish between two populations. Although BioBin is still in developmental stages, it will be a useful tool in analyzing sequence data and uncovering novel associations with complex disease.

Systems Mapping: How to Design Crop Ideotypes through Plant Economy

Dr. Rongling Wu

The recent availability of high-throughput genetic and genomic data allows the genetic architecture of complex traits to be systematically mapped. The application of these genetic results to design and breed new crop types can be made possible through systems mapping. Systems mapping is a computational model that dissects a complex phenotype into its underlying components, coordinates different components in terms of biological laws through mathematical equations, and maps specific genes that mediate each component and its connection with other components. In this talk, I will present a new direction of systems mapping by integrating this tool with plant economy. Through synthesizing various plant traits in an optimal way, plants are capable of maximizing whole-plant growth and competitive ability under limited availability of resources. I argue that such an economic strategy for plant growth and development, once integrated with genetic mapping, will not only provide mechanistic insights into plant biology, but also help to spark a renaissance of interest in ideotype breeding in crops and trees.

Session II

A Look at the Earth Microbiome Project

Dr. Mary Ann Bruns

The extraordinary diversity and variation of soil microbial communities have made bioinformatics an indispensable tool in terrestrial microbial ecology. One gram of topsoil contains up to ten billion individual cells comprising bacteria, archaea, fungi, and protists. Although growing microbes in the laboratory is a prerequisite to understanding their metabolisms, only 1-2% of microbial species have been cultured. More comprehensive information about soil microbial composition can be gained through metagenomics approaches using high-throughput sequencing of amplicons from16S and 18S rRNA genes. To learn about microbial function, however, shotgun sequencing of total community DNA (or RNA) must be used to tap into the rest of the soil "metagenome." The Earth Microbiome Project (EMP) is an international initiative that will build upon data from a few current soil metagenome projects and ultimately analyze 200,000 soil samples from biomes around the world (www.earthmicrobiome.org). The EMP thus proposes to construct a "Microbial Biomap for Planet Earth" based on well characterized "environmental parameter spaces." Other proposed products include a global Gene Atlas, assembled genomes, visualization portal, and metabolic reconstructions. The EMP so far involves scientists at several universities, private companies, the Joint Genome Institute and national research labs of the U.S. Dept. of Energy. The First International EMP Conference, organized by the Beijing Genomics Institute (BGI) in Shenzhen China in June of this year, was an outgrowth of BGI's collaboration with the European Union on metagenomics of bacteria in human intestinal flora. BGI's pilot contribution to EMP will be 30-40 trillion base pairs of sequence data from 10,000 soil samples. Improved sample processing procedures, algorithms and tools will be needed to assemble, interpret, and compare massive amounts of multiple genome data. A key question is whether such "snapshot" views of terrestrial microbial communities will capture sufficient information to answer fundamental questions about microbial roles in global biogeochemical cycles.

Combinatorial use of poly-A/T tracts in organizing genes, nucleosomes and the transcription machinery in *Dictyostelium*

Gue Su Chang

Dictyostelium discoideum is a member of the amoebozoa that exists in both a free-living unicellular and a multi-cellular form. It is situated in a deep branch in the evolutionary tree, and is particularly noteworthy in having a very A/T-rich genome. This organism provides an ideal system to examine the extreme to which nucleotide bias may be employed in organizing promoters, genes, and nucleosomes across a genome. We find that Dictyostelium genes are demarcated precisely at their 5' ends by poly-T tracts and precisely at their 3' ends by poly-A tracts. These tracts are also associated with nucleosome-free regions, and are embedded with precisely positioned TATA boxes and AATAAA cleavage and polyadenylation sites. Homo- and heteropolymeric tracts of A and T demarcate nucleosome border regions. Together these findings reveal the presence of a variety of functionally distinct polymeric A/T elements. Strikingly, Dictyostelium chromatin is organized in di-nucleosome units, where in contrast to the rather uniform spacing of genic nucleosomes seen in other organisms, pairs of nucleosomes have close spacing. Dictyostelium nucleosomes are organized similar as in animals, having a largely cell state-independent canonical organization, including a +1 nucleosome in position of a paused RNA polymerase. Indeed, we find a strong phylogenetic relationship between the presence of the NELF pausing factor and positioning of the +1 nucleosome. Pausing and nucleosome positioning may have co-evolved with animal multi-cellularity.

Assessing Vaccination Sentiments with Online Social Media: Implications for Infectious Disease Dynamics and Control

Dr. Marcel Salathe

There is great interest in the dynamics of health behaviors in social networks and how they affect collective public health outcomes, but measuring population health behaviors over time and space requires substantial resources. We've used publicly available data from 101,853 users of online social media collected over a time period of almost six months to measure the spatio-temporal sentiment towards a new vaccine. We validated our approach by identifying a strong correlation between sentiments expressed online and CDC- estimated vaccination rates by region. Analysis of the network of opinionated users showed that information flows more often between users who share the same sentiments - and less often between users who do not share th! e same sentiments - than expected by chance alone. We also found that most communities are dominated by either positive or negative sentiments towards the novel vaccine. Simulations of infectious disease transmission show that if clusters of negative vaccine sentiments lead to clusters of unprotected individuals, the likelihood of disease outbreaks are greatly increased. Online social media provide unprecedented access to data allowing for inexpensive and efficient tools to identify target areas for intervention efforts and to evaluate their effectiveness

Keynote Speaker

Human evolution revealed by extinct hominin genomes

Dr. Richard Edward Green

Neandertals are our closest extinct relatives. The show up in the fossil record about 200,000 years ago and disappear about 30,000 years ago. We have recently completed an effort to recover genome-scale DNA sequence information from Neandertals using DNA recovered from 38,000 year old bones. These data that derive from short, chemically damaged DNA fragments present unique challenges for assembly and analysis. These data show that human ancestors admixed with Neandertals early in their range expansion. Further, we show how variation within currently living humans can be contrasted with that of the Neandertal to identify regions of recent positive selection in humans. In this way, Neandertal genetic information can be used to define the basis of human uniqueness.

Session III

Challenges and opportunities for eco-evolutionary genomics with trees

Dr. John Carlson

Genomics and bioinformatics play fundamental roles in plant research. This is as true for agricultural crop plants as for model species such as Arabidopsis. But what about trees? Trees are economically important sources of food and fiber, and play keystone roles in many ecosystems. However trees are certainly not typical research systems, being mostly un-domesticated and having long generation times. However, many tree species live in large, natural populations that are the direct result of many thousands of years of major evolutionary forces of mutation, natural selection, genetic drift, migration, and hybridization. Thus natural populations of forest trees represent long-term experiments in adaptation which are rich, largely untapped resources for evolutionary and ecological genomics. However, the sustainability of many of tree species and populations are now threatened by introduced insects, exotic diseases, invasive plants, climate change and forest fragmentation. The complete genome sequences for grapevine, poplar, cacao, Eucalyptus, apple, papaya, and peach are enabling interesting new comparative and population genomics research efforts focused on woody plants. However a broader, phylogenetically-based set of genomic resources and bioinformatics tools are required to capture all of the lessons which the evolution of woody plants can teach, and to address the challenges which forest ecosystems face.

GCD: A new method for identifying gene families from genomic sequence data

Dr. Zhenguo Zhang

Identification and classification of gene families from genomic sequence data are crucial for the study of evolution of genes, genomes and phenotypic characters. However, because of the complexity of genomic evolution, the accurate identification of gene families has been a challenging problem. Here we develope a new computational method, termed the generalized conserved domains (GCD), which can cluster genes into families based on the conserved domains. Briefly, an initial sequence similarity network is first constructed by using the homolog search results with BLAST. Then the conserved domains in each sequence are determined by checking its neighborhood in the network. After that, only the homologous relationships

generated by the conserved domains are retained and the gene families are extracted by using the single-linkage algorithm. We applied this new method to the proteomes of 11 animal species, containing 225145 genes in total. The results showed that our method can significantly remove false homologous relationships generated by 'promiscuous' domains. Compared to MCL (Markov Cluster Algorithm), the gene families identified by this new method show significantly better functional consistency in terms of domain architecture, enzyme classification and KOG annotations. This method is supposed to perform better when the ensemble of domains is more complete and accurate. In summary, our method is an effective tool in identifying gene families from genomic sequence data.

Evolution of Genomic Rearrangements in *Drosophila*

Dr. Stephen Schaeffer

The gene composition of chromosomes is conserved among *Drosophila* species, however, gene order varies widely among species. An important mechanism for rearranging genes is chromosomal inversions that result from two double strand DNA breaks that are rejoined in reverse order. Four general classes of models are used to explain how chromosomal inversions arise and are established in populations. *Drosophila pseudoobscura* is a model for the study of mechanisms for the origin and spread of inversions. This species has over 30 different gene arrangements that were generated by a series of overlapping inversions. The frequencies of the different arrangements vary among populations and the major shifts in frequency are correlated with changes in the physiographic provinces in the southwestern United States. My laboratory uses a combination of theoretical and experimental approaches to understand the genetic forces that modulate the frequencies of gene arrangements among populations. This talk will summarize data from numerical and molecular population genetic analyses that evaluate hypotheses about the origin and maintenance of different gene arrangements.

A fine line between friend and foe: comparative genomics of the Fungi

Josh Herr

Fungi are responsible for nutrient cycling in ecosystems and are found as beneficial endosymbionts and potent pathogens in both plants and animals. Yet, one could consider the Fungi as the Rodney Dangerfields of the biological world: in the genomic age they still don't get any respect. Although yeasts and easily culturable Ascomycetes have long been used for genetic studies, we're only just starting to make progress in understanding gene function and genome evolution in the Fungi. For the eukaryotes, Fungi have the smallest and most streamlined genomes – which benefits sequencing, assembly, and annotation – but a heavy reliance on easily culturable species has left us with a patchy collection of genomes. One of the most surprising discoveries of fungal comparative genomics is how species with very different lifestyles have similar fungal genomes and vice versa. I will outline our current state of knowledge of fungal comparative genomics and discuss my involvement in on-going fungal genome sequencing projects. The goal of these sequencing initiatives is to identify and compare determinants shaping evolution across the disparate fungal taxa associated with plants and animals. Sequencing fungal genomes contribute to – including but is not limited to – the better understanding of parasite, symbiont, and saprobic interactions with host organisms, the basis for understanding phenotypic structure and development, the description and understanding of organic secondary metabolites typical of antibiotics and volatile compounds, the development of

molecular markers to be used for tractability in various environments, and the breakdown of cellulosic and other organic material in ecosystems.

Standing genotypic variation is basis for coral survival of temperature stress

Dr. Iliana Baums

Dispersal of larvae via ocean currents is a common strategy among sessile marine organisms whereby spatially discontinuous populations maintain a shared gene pool. Long-lived reef building corals employ such strategies to connect their shallow-water populations across stretches of inhospitable open-ocean. Increasing seawater temperatures are predicted to accelerate larval development so that average dispersal distances are expected to decrease. At the same time, species currently limited by minimum annual seawater temperatures may extend their ranges pole-ward in the coming years. The objectives of this project are to measure the effect of high and low seawater temperatures on early life stages and ultimately predict changes in connectivity patterns as a result of global warming. We have measured adaptive trait variation among individuals and populations in the reef-building corals Montastraea annularis and Acropora palmata. Symbiont-free A. palmata larvae were exposed to a range of seawater temperatures and their development patterns were documented. In larval cultures (batches) that consisted of eggs and sperm from four parents (4 batches, 4 replicates per time and temperature) the contribution of each parent to the cultures was monitored over time through genotyping. With time, some parents contributed larger than expected numbers of larvae to the batch cultures while others contributed close to zero larvae indicating selection. Experimental temperatures 1C above environmental seawater temperatures resulted in quicker development so that larvae became competent to settle about 1 day earlier but they also experienced markedly higher mortality rates. Further, parents produced offspring that differed in swimming speed, sperm tail length, lipid content and fertilization rates. A comprehensive microarray experiment (138K features, based on a high quality 454 transcriptome) showed clear differences among larval batches in a suite of differentially expressed genes responding to thermal treatments. Together, these data indicate that large standing genotypic variation might provide the basis for future survival of these threatened species.

Session IV

Translational Microbial Pathogenomics

Dr. Vivek Kapur

Louis Pasteur, the father of modern microbiology, epitomizes the scientist who conducted "use-inspired research" – basic research that was motivated by practical applications and utility. This two-centuries old approach is helping set the modern day research agenda at the intersection of basic and applied research. I will share specific case studies with major animal and human bacterial pathogens, and describe how microbial genomics and its related translational activities lead to a better understanding of basic life processes, as well as help enhance animal and human health through improved diagnostics and vaccines.

Host Microflora and Respiratory Pathogen Interactions during Infection: A Multidisciplinary Approach

Olivier Rolin

As we begin to understand more about the human microbiome, we can also appreciate that we know relatively little about how host microflora and pathogens interact *in vivo* and the mechanisms underlying these interactions, especially in the respiratory tract. Using a robust mouse model of the respiratory pathogen *Bordetella bronchiseptica*, we monitored nasal cavity microflora prior to and after infection. Initially, displacement of culturable host microflora was observed, and 16S RNA sequencing of these microorganisms verified that *Staphylococcus*, *Kytococcus*, *Enterobacter,* and *Bacillus* species were all eliminated within three days post-inoculation. A 454 metagenomic sequencing approach was employed to observe changes in unculturable microbes, identifying over different 80 bacterial genera inhabiting the murine nasal cavity. *B. bronchiseptica m*utants defective in the Type III Secretion System, the Type VI Secretion System, or the master virulence regulator BvgAS were not able to clear microflora, suggesting these bacterial virulence factors work actively and synergistically to cause displacement. *In vivo* microflora displacement was also found to be dependent on T-cells, interleukin 10, and interleukin 17, suggesting this observation in immune mediated. Interestingly, closely related human pathogens, *Bordetella pertussis* and *Bordetella parapertussis,* do not cause clearance of culturable nasal cavity flora, suggesting that host microflora displacement may be a pathogen specific event. These data suggest that the immune mediated mechanism behind host microflora displacement is an active pathogenesis process that affects a wide range of microbial species.

Characterizing the ciprofloxacin-inducible bacteriophage populations of *Escherichia coli* O157:H7 by 454 sequencing

Dr. Edward Dudley

E. coli O157:H7 is a foodborne pathogen most commonly associated with outbreaks linked to undercooked ground beef and fresh produce. One of the best studied virulence factors in this organism is Shiga toxin, and a specific variant designated Stx2 is highly associated with severe clinical disease. Stx2 is encoded on a lysogenic bacteriophage of the lambda family, and this phage can be induced by DNA damaging agents such as ciprofloxacin. As the genes encoding Stx2 are under the control of the phage late promoter, induction dramatically increases production of the toxin. We have been characterizing a set of clinical isolates from the Pennsylvania Department of Health, and have noted a large variability in ciprofloxacin-induction profiles, and in Shiga toxin production. We are using 454 sequencing to determine if these and other phenotypes can be explained by differential genomic content of the Stx2-encoding phage. Consistent with previous publications, we observed that these phage have mosaic genomes, but we also determined that ciprofloxacin induces a number of other phage found within the *E. coli* O157:H7 genome including those previously predicted to be non-functional. The repertoire of phage identified varied between strains, and the quantity of Stx2-bacteriophage induced was not an accurate indicator of Stx2 production. We currently hypothesize that lysogenic bacteriophage that lack known or putative virulence genes may still play a role in modulating the virulence of *E. coli* O157:H7.

The genomic structure and transcriptome of the male-specific region in the bovine Y chromosome

Ti-Cheng Chang

The male-specific region on the mammalian Y chromosome (MSY) was derived during the progressive differentiation of sex chromosomes and provides a particular genomic niche to harbor genes essential for male-related function. The MSY contains two major sequence classes, X-degenerate and ampliconic. Combining a direct testis cDNA selection and a Next-gen sequencing approach, the bovine MSY transcriptome was analyzed in this study. We identified 200 single-copy and 743 ampliconic transcriptional units (TUs). Eleven single-copy TUs matched known Y-genes, including 9 reported genes (*GPR143Y*, *EIF1AY*, *OFD1Y*, *USP9Y*, *UTY*, *DDX3Y*, *ZFY*, *UBE1Y*, *RBMY*) and 2 newly identified genes (*ZRSR2Y*, and *RPL23AY*). The single-copy TUs were distributed in three distinct regions and shared a similarity of ~80-99% with their X-linked counterparts. These regions represent the X-degenerate region in the bovine MSY, whereas the remainders belong to the ampliconic region. The ampliconic region is composed of a tandem array of a basic repetitive unit that is ~ 420 kb in size and was duplicated 80 times during the bovine Y chromosome evolution. Each unit contains 19 novel TUs and 4 known gene families (*ZNF280BY, ZNF280AY, TSPY,* and *HSFY*). The analysis of the novel TUs revealed that ~80% of them may belong to non-coding transcripts. 88 novel TUs contain a motif related to transposable elements, which may contribute to the extensive duplications in the ampliconic region. In conclusion, our results revealed that a large-scale transcriptional activity is present in the bovine MSY. The amplification and accumulation of MSY genes are lineage-dependent.

Cancer Epigenetics

Dr. Sagarika Kanjilal

Epigenetics is the study of heritable changes in phenotype, and the underlying alterations in gene and protein expression, caused by mechanisms other than changes in the genomic sequence. It encompasses diverse mechanisms from those affecting the packaging, binding, and accessibility of DNA and the transcription of RNA isoforms, down to structural and functional alterations in the encoded proteins. Such alterations are increasingly being implicated in the development of a number of diseases including cancer and the nascent field of Cancer Epigenetics is expanding very rapidly.

While the growing body of literature on epigenetic mechanisms that drive cancer development and progression highlight the overwhelming complexity of the disease process, they also bring great hope with respect to cancer prevention and control since many epigenetic modifications can be completely reversed by environmental factors. Studies in our laboratory are presently directed at elucidating the role and epigenetic regulation of the inflammatory cytokine leptin in cancer development as well as the identification of food items capable of reversing these cancer associated changes.

Poster Session

Combinatorial use of poly-A/T tracts in organizing genes, nucleosomes and the transcription machinery in *Dictyostelium*

Gue Su Chang

Genome-wide mapping of nucleosomes has significantly expanded our understanding of chromatin structure and its function in transcriptional initiation and regulation of eukaryotic genes. Here we present high-resolution maps of *in vivo* nucleosome locations of the social amoeba *Dictyostelium discoideum*, whose genome is of unique (A+T)-richness, surpassed only by *P. falciparum*. The nucleosome maps from its two alternative life forms, as of unicellular and multicellular cells, led to advances in our understanding of nucleosome organization around eukaryotic genes. *D. discoideum* showed a canonical nucleosome organization around the 5' end of its genes, maintained regardless of two different developmental states. Strikingly, the *Dictyostelium* TSSs were found upstream of the edge of the +1 nucleosome, as seen for multicellular eukaryotes (but not for fungi). Given the conservation of overall chromatin structure across eukaryotes, this nucleosome architecture was conserved mainly in multicellular organisms such as animals and plants. Nucleosome-free regions (NFR) were evidently detected at the 5' and 3' end of *D. discoideum* genes. In particular, the (A+T)-richness of the *Dictyostelium* genome provided its NFR with the unique features, for example notable directional enrichment of poly(dA:dT) tracts in 5' and 3' NFR. This discovery may indicate a distinct evolutionary constraint imposed on the evolution in the (A+T)-rich *D. discoideum* genome. Our study further demonstrate a strong linkage between transcriptional pausing and the canonical location of the +1 nucleosome, and this may be an essential feature of transcriptional regulation in multicellular eukaryotes (but not fungi)

The *Amborella* Genome Sequencing Project: Generating An Evolutionary Reference Sequence

Joshua Der

The origin and early diversification of flowering plants (angiosperms) had profound impacts on Earth's biota, providing the raw genetic material from which most crops and economically important plants were derived. The diversification of genes, genomes, and important traits cannot be adequately interpreted without a comparative framework firmly rooted with genome sequences from basal angiosperms. As the sister species to all other extant flowering plants, *Amborella trichopoda* holds a singular position in the flowering tree of life for establishing this comparative genomics framework. The *Amborella* Genome Sequencing Project is an NSF funded collaborative project that seeks to produce a high quality finished genome sequence, complete with an accurate chromosome-scale physical map and evidence based gene and transposable element annotations. We are combining the latest multi-platform sequencing technologies with traditional and cutting-edge physical mapping techniques and emerging multi-faceted assembly algorithms to produce a reference sequence ideally suited for comparative evolutionary analyses to shed light on genomic characteristics of the last common ancestor of extant angiosperms and the evolution of gene content and genome structure throughout angiosperm history. To leverage this genome sequence, we are also developing new bioinformatic tools and a public access website to make this important resource available to the scientific community.

Microsatellite evolution in the absence of recombination

Jill Demers

Microsatellites, also called simple sequence repeats, are tandem repeating units of 1 to 6 base pairs of DNA. Two mutation mechanisms have been proposed to explain how microsatellites gain or lose repeat units: (1) slippage of DNA strands during replication and (2) recombination at microsatellite loci, especially unequal crossing-over. Evidence suggests that replication slippage is the main mutational force, but recombination has not been ruled out as a contributing factor. In this study, the behavior of microsatellites in the absence of recombination was investigated by sequencing nine microsatellite loci for 30 isolates of the haploid asexual fungus *Fusarium oxysporum*. Mutational patterns over evolutionary time were studied by mapping changes in repeat number to the species phylogeny. The data were also fit to common models of microsatellite evolution to estimate mutation rates and to estimate the proportion of mutations involving multiple repeat units. Preliminary results indicate that these microsatellites are evolving in a similar manner as microsatellites in eukaryotes that undergo recombination. Estimated mutation rates are similar to those previously reported for fungi that undergo a diploid sexual stage, supporting the hypothesis that replication slippage is the primary mutational mechanism. Some mutations appear to have involved multiple repeat units, suggesting that recombination is not needed to cause large mutations and that replication slippage can act on multiple repeat units.

Complete genome sequence of a *Mycobacterium avium* subspecies *paratuberculosis* isolate from a patient with Crohn's disease

Lingling Li

Mycobacterium avium subspecies *paratuberculosis* (*Map*) has been identified in some human patients with Crohn's disease. In order to identify genetic differences between MAP isolates recovered from humans and those associated with bovine Johne's disease, we characterized the complete genome sequence of strain MAP4 recovered from the breast milk of a Crohn's disease patient. Massively parallel sequencing approaches were used to generate a total of 88.5 million base pairs from a randomly sheared MAP4 genomic DNA library, which where assembled into contiguous sequence fragments with an estimated 60 large (~2kb) and 350 small (<0.5kb) gaps. A primer walking approach with Sanger based sequencing was applied to close all remaining gaps in a iterative manner and areas with low quality sequence re-sequenced in order to obtain an assembled single high quality genome sequence. Compared with the previously described bovine MAP K10 genome, the size of MAP4 genome is about 3.0kb smaller as a result of several sequence deletions, including in one copy of the insertion sequence element, IS*900*. Importantly, the analysis revealed no large genome scale re-arrangements in MAP4 as compared with strain K10, and ~3kb of deletions and ~300 bp of insertions were distributed across the genome. The results also confirmed the presence a total of 233 SNPs between these two isolates. Interestingly, more than half of the newly identified SNPs were located in two genes (MAP1432 and MAP2495), both of which contain repetitive sequences and are orthologs of *Mycobacterium tuberculosis* Rv1128c that encodes a cell wall protein. Taken together, our analysis of the MAP4 and K10 genome sequences confirmed the high similarity between strains from these two different mammalian hosts, and suggest a relative paucity of genetic variation amongst strains recovered from humans and cows.

A gene potentially associated to the onset of puberty in Bighorn sheep

Oscar Bedoya

Northern bighorn sheeps (*Ovis canadensis*) are grouped into three subspecies. They differ in horn and skull characters, breeding and birth time, and geographical distribution. Individuals of the desert subspecies (*O. c. nelsoni*) have earlier testicular descent and sexual maturity than their Rocky Mountain counterparts (*O. c. canadensis*). As a prototype for affordable application of high-throughput sequencing we generated full-genome sequence data for desert and mountain bighorn sheep populations, each a pooled mixture of DNA from several individuals. By analyzing the polymorphisms between these populations, we discovered that the relaxin/insulin-like family peptide receptor 2 gene (*rxfp2*) has different variants fixed in each subspecies. Previous studies have correlated the expression of *rxfp2* with testicular descent and sexual maturation in mammal models. We speculate that the physiological effects of the fixed polymorphisms in *rxfp2* might be responsible for the sexual development differences between bighorn sheep subspecies."

The complete genome sequence of *Bifidobacterium animalis* subsp. *animalis* ATCC 25527 and an investigation of growth in milk

Joe Loquasto

Bifidobacteria are putative probiotic organisms commonly added to fermented dairy products.
The number of complete bifidobacterial genomes has is increasing and analysis of these genomes has provided important insight into the physiology of these organisms. The objective of this work was to sequence the genome of *B. animalis* subsp. *animalis* ATCC 25527^T with the aim of providing insight into the genetic diversity responsible for phenotypic differences reported between *B. animalis* subsp. *animalis* (Baa) and *B. animalis subsp. lactis* (Bal). The genome of ATCC 25527^T was shotgun sequenced using 454 technology. After contig assembly, alignment and several rounds of gap closing, the complete 1,932,963 bp genome was determined and verified by comparison to a *Kpn*I optical map. The genome was annotated using Rapid Annotation using Subsystems Technology (RAST) and at NCBI. Comparative analysis of the Baa ATCC 25527^T and Bal DSM 10140^T genomes revealed high degrees of both synteny and homology. Comparison of the Baa and Bal genomes for differential content revealed 108 and 121 genes that were unique to and absent in, the BAA genome, respectively. Unique genes were identified as having less than 10% amino acid identity between protein sequences of both genomes, as detected by RAST. Among the differential gene content are a set of unique CRISPR-associated genes and a novel CRISPR locus containing 31 spacers in the genome of Baa. Although previous research has suggested one of the defining phenotypic differences between Baa and Bal is the ability of Bal strains to grow in milk and milk-based medium, no obvious differences in gene content responsible for this phenotype were identified between the two genomes. Furthermore, growth and acid production in milk and milk-based medium did not differ significantly in experiments examining Bal (DSM 10140^T and Bl04) and Baa (ATCC 25527^T). These data suggest that this widely accepted defining phenotypic trait may not distinguish the subspecies.

SOLiD RNA-Seq Reveals Differences in Midgut Expression Profiles Between Larvae that Ate Resistant of Susceptible Corn Foliage

Howard W. Fescemyer

Fall armyworm is comprised of two sympatric and morphologically identical strains that are genetically distinct and defined by their nutritional ecology. Populations of the corn strain are almost exclusively found feeding on large grasses (e.g., corn), while populations of the rice strain are primarily found feeding on small grasses (e.g., rice, bermudagrass). Larvae of both strains were fed throughout development on bermudagrass, caterpillar resistant Mp708 corn, or susceptible Tx601 corn. Illumina and SOLiD RNA-Seq were used to sequence and measure expression of transcripts in several tissues (midgut, salivary gland, fat body, central nervous system) from last instars. Although data analysis is ongoing, it is most clear that Mp708 corn induces up-regulation of several genes in both strains. Foliage of Mp708 corn expresses Mir1-CP cysteine protease that attacks the peritrophic membrane (PM). Larvae that ate Mp708 foliage have a transcription expression profile, especially in midguts, that suggests they are trying to counteract the action of Mir1-CP. In contrast to the susceptible Tx601 treatment, these midguts upregulated genes probably involved in repair of the PM, protein digestion by PM bound enzymes, wounding and antimicrobial reactions, protein translation, and hydrolytic defense targeting Mir1-CP. ArmywormBase is being developed as a WEB-accessible transcriptome database for public access to all sequence data generated by this project. These results ascertain the efficacy of RNA-Seq in transcriptome expression profiling to discover with limited genomic resources how a broad group of genes in a caterpillar respond to nutritional differences among their host plants.

Relating Gene Expression Profile and Epigenetic Landscape during Erythropoiesis

Tyler Malys

Erythropoiesis is the process by which red blood cells mature. During Erythropoiesis, changes can be observed in the gene expression profile as well as in the epigenetic landscape. The expression level change of most individual mouse genes was categorized into four classes: significant increase, significant decrease, no change, or ambiguous change. Moreover, genes which undergo significant increase or significant decrease in expression level were considered responsive genes whereas genes which exhibit no change are considered non responsive genes. We ask two questions. How well can we predict the category of a gene based on observed epigenetic features? Which epigenetic features are most important for identifying genes of a given gene category? In general, we found that epigenetic features are useful in gene response prediction. Also, we found that larger epigenetic differences exist between responsive and non responsive genes than amongst responsive genes. This suggests that whether a gene responds or not is under more prominent epigenetic control than how a gene responds. When considering all four gene categories, we find the most powerful epigenetic factors for successful prediction of gene category to be certain Histone marks. Conversely, specific transcription factors become most powerful when considering only responsive gene categories. We hypothesize that this is because these Histone marks control the openness of chromatin and thus, the ability of genes to significantly change expression level during Erythropoiesis. In conclusion, our approach provides a novel method for investigation of the relationship between gene expression profiles

and epigenetic landscapes by quantifying the power of individual epigenetic factors to predict gene response category.

Predicting common fragile sites in the human genome using regression analysis

Arkarachai Fungtammasan

Common Fragile Sites (CFSs) are unstable genomic regions that frequently break under replication stress conditions. Although they are involved in chromosomal rearrangements and breakages in cancer, their characteristics remain unclear making computational predictions of their locations a challenge. Here we predicted CFSs and their breakage frequency of CFSs based on local genomic contexts. We performed the analysis at the whole-genome level utilizing all known 76 CFSs and compared them with non-fragile regions. Our logistic and multiple regression models can predict 83% of CFSs and 43% of variability in their breakage frequency, respectively. Moreover, our models can predict fragile regions in the mouse genome with 70% success rate. The significant predictors we identified in these models allowed us to make conclusions about CFS biology and evolution. CFSs are mostly present in G-negative chromosomal bands. They are enriched in mononucleotide microsatellites and depleted in CpG islands. Also, CFSs tend to locate away from centromere. Interestingly, highly fragile CFSs tend to be co-located with evolutionary break points. This information is valuable for predicting location and studying molecular mechanisms of fragility for CFSs.

Comparison of erythroid and megakaryocytic transcriptomes reveals shared expression patterns during erythromegakaryocitic differentiation

Tejaswini Mishra

Hematopoiesis is the process by which pluripotential hematopoietic stem cells develop into different mature blood cell types. Hematopoietic stem cells (HSCs) can continuously self-renew but are also able to commit to any of the differentiation pathways that lead to specific types of blood cells. They and their progeny are subject to lineage commitment decisions at progressive levels, from pluripotential progenitors to multipotential cells to bipotential progenitors and finally commitment to one lineage of mature blood cells. For example, erythromegakaryocytic differentiation proceeds from HSCs to the common myeloid progenitor (CMP) cells, which in turn differentiate to form either of two bipotential progenitors, megakaryocyte-erythroid progenitors (MEPs) or the granulocyte-macrophage progenitors (GMPs). Hematopoiesis, like other developmental and differentiation processes, is driven by lineage-specific changes in gene expression. Chromatin accessibility, occupancy or co-occupancy by lineage-specific and general transcription factors, and presence of specific histone marks are some of the ways in which gene expression is regulated. We have profiled gene expression changes during erythromegakaryopoiesis and are currently mapping patterns of transcription factor occupancy and specific histone modifications in order to understand the mechanisms of gene regulation that drive decisions on lineage commitment and terminal differentiation during hematopoiesis. We isolated highly purified populations of primary HSCs, erythroblasts and megakaryocytes from mouse fetal liver by a series of purifications based on distinctive cell surface markers and, for megakaryocytes, expansion in thrombopoietin. We measured gene expression levels in the three populations using the Affymetrix Mouse Gene 1.0ST Array, and we are using chromatin immunoprecipitation followed by massively parallel sequencing (ChIP-seq) to monitor informative histone modifications and to determine the DNA segments occupied by Gata1 and Tal1 in the differentiated erythroblasts and megakaryocytes. With these data, we can use

parsimony to assign changes in expression and associated epigenetic features to the lineages from HSC to MEP and to contrast these results to those in the committed lineages. This will not only provide a precise map of key steps during differentiation for thousands of genes, but it will allow us to address long-standing issues central to developmental biology, such as whether differentiation proceeds primarily through inductive or a repressive processes, and the stages at which permissive chromatin states are established.

Transcriptional Regulation of Erythropoiesis via High Resolution ChiP-exo

Garam (Celine) Han

ChIP-exo is a novel DNA sequencing technology that maps the precise binding locations of transcription factors with high resolution and low background (Rhee and Pugh, 2011). During the conventional ChIP-exo analysis, sequencing tags that accumulate about a specific DNA sequence on opposite strands are defined as a peak pair and used to demarcate the border of factor binding locations. However, a more complex pattern of multiple peak pairs has been observed in many transcription factor binding locations, including the erythroid development factor GATA1. Therefore, a new analysis method that incorporates and reconciles the patterns of multiple peaks is being developed to achieve more comprehensive genome-wide transcription factor binding maps. As a means to develop this new data analysis method, we chose to study the precise binding locations of various transcription factors involved in erythropoiesis to shed light on the transcriptional mechanisms at work during hematopoietic stem cell differentiation. This analysis method is being applied to study the binding locations of GATA1 and TAL1 at various time points of erythropoiesis to understand their interaction and role during erythropoiesis. The study will be expanded to map the binding sites of various factors in multiple developmental stages to gain a comprehensive understanding of the dynamic transcriptional regulation during erythropoiesis.

Comparative genomic and phylogenomic analysis of the classical *Bordetella* species

Jihye Park

Despite infecting diverse host populations and having various virulence phenotypes, the three classical bordetellae, *Bordetella bronchiseptica*, *B. parapertussis*, and *B. pertussis*, share an extremely high genomic similarity and thus have been reclassified as sub-species. This study expands upon the 3 previously sequenced genomes by employing comparative genomic and phylogenomic analyses of 10 classical *Bordetella* genomes to understand the genetic and evolutionary basis of host adaptation. Comparative analyses of these 10 genomes show that the classical bordetellae have an open pan-genome (5,497 gene families) of 2,861 core and 2,636 non-core gene families with limited introduction of the new genetic material. This study presents many absence, presence, or sequence variations in the virulence factor genes in the classical bordetellae and more robust phylogenetic trees on genome-wide data. There may be the pertussis toxin locus (*ptx*) and its secretion system locus (*ptl*) horizontal gene transfer events within the classical bordetellae, in addition to the horizontal gene transfer from other bacteria. This analysis unveils the variability and evolutionary history of *Bordetlla* more comprehensively and will serve as the basis for large-scale comparative genomic analysis of the genus *Bordetella*.

Analysis of *Arabidopsis* genome-wide variations before and after meiosis and meiotic recombination by re-sequencing *Landsberg erecta* and all four products of single meiosis

Xinwei Han

Meiotic recombination, including crossovers (COs) and non-crossovers (NCOs), impacts natural variations and is an important evolutionary force for sexually reproducing organisms. COs increase genetic diversity by redistributing existing variations, whereas gene conversions (GCs) associated with COs and NCOs can homogenize variants. Here we sequenced *Arabidopsis* Landsberg *erecta* (L*er*) and two sets of all four meiotic products from a Columbia (Col)/L*er* hybrid, to investigate genome-wide variations and meiotic recombination events at nucleotide resolution. Comparing L*er* sequences with the Col reference uncovered 349,171 Single Nucleotide Polymorphisms (SNPs), 58,085 small and 2,315 large insertions/deletions (indels), with highly correlated genome-wide distributions of SNPs and small indels. 443 genes have at least 10 nonsynonymous substitutions in protein coding regions, with enrichment for disease resistance genes. Another 316 genes are affected by large indels, including 130 genes whose coding regions were completely deleted in L*er*. Using the *Arabidopsis qrt* mutant, two sets of four meiotic products were generated and analyzed by sequencing for meiotic recombination, representing the first tetrad analysis in a nonfungal species. Our results detected 18 COs, 6 of which with observed GC tracts, and 4 NCOs, and revealed that *Arabidopsis* COs and NCOs tracts are fewer and shorter than those in yeast. Meiotic recombination and chromosome assortment events dramatically re-distribute genome variation in meiotic products, sometimes in a non-random fashion, contributing to population diversity. In particular, meiosis provides a rapid mechanism to generate copy number variations (CNVs) for sequences that have different chromosomal positions in the Col and L*er* genomes.

Building Metabolic Reconstructions of *Cyanobacteria* to Determine Methods for Targeted Metabolite Overproduction

Thomas Mueller

This project focused on the development of genome scale metabolic reconstructions of *Synechocystis PCC 6803* and *Cyanothece ATCC 51142*, cyanobacterial strains which have recently garnered much attention in the research community. The biomass equation was developed using experimental data acquired specifically from the cyanobacterial strains being studied. The computational model was implemented in GAMS where flux balance analysis was applied to test the model. Completion of these reconstructions and verification of correct biomass production will allow for development and initial testing of genetic modifications which lead to the overproduction of certain targeted cell metabolites, specifically biofuels.

Ancestral polyploidy in seed plants and angiosperms

Yuannian Jiao

Whole-genome duplication (WGD, polyploidy) followed by gene loss and diploidization has long been recognized as an important evolutionary force in animals, fungi, and other organisms, especially plants. The success of angiosperms has been attributed, in part, to innovations associated with gene or whole-genome duplications, but evidence for hypothesized ancient genome duplications predating the divergence of monocots and eudicots remains equivocal in analyses of conserved gene order. Here we use comprehensive phylogenomic analyses of sequenced plant genomes and more than 12.6 million new EST sequences from phylogenetically

pivotal lineages to elucidate two groups of ancient gene duplications – one in the common ancestor of extant seed plants and another in the common ancestor of extant angiosperms. Gene duplication events were intensely concentrated at around 319 mya and 192 mya, implicating two WGDs in ancestral lineages shortly before the diversification of extant seed plants and extant angiosperms, respectively. Significantly, these ancestral WGDs resulted in the diversification of regulatory genes important to seed and flower development, suggesting that they contributed to major innovations that ultimately contributed to the rise and eventual dominance of seed plants and angiosperms.

Functional Metagenomic Profiling of Asian Longhorned Beetle (Cerambycidae: *Anoplophora glabripennis*) Microbiota Reveals Important Contributions to Digestive Physiology

Erin Scully

The Asian longhorned beetle (ALB) is a destructive, wood-boring pest with a broad host range that thrives in the heartwood of healthy deciduous trees. Despite its ability to enjoy a broad host range, ALB must overcome a number of challenges to subsist in heartwood, which is devoid of easily accessible nutrients. For example, glucose in wood is present as cellulose and hemicellulose, which are inherently difficult to digest, requiring a suite of enzymes for efficient degradation. Access to these polysaccharides is further limited by the presence of a recalcitrant lignin barrier, a structural biopolymer that contains twelve types of chemical bonds and is especially resistant to degradation. Wood is also lacking in essential amino acids, vitamins, nitrogen-containing compounds, sterols, and fatty acids.

The ALB gut harbors a phylogenetically diverse microbial community that likely contributes to digestion of intractable compounds and augments the nutritional content of the beetle's diet. To assess the metabolic potential of these microbiota, we performed shotgun metagenomic sequencing of gut microbial DNA and detected an abundance of genes associated with polysaccharide digestion, including complete suites of microbial-derived cellulase and hemicellulase enzyme complexes. Furthermore, a substantial percentage of genes were associated with amino acid synthesis, indicating that gut microbes may be provisioning ALB with essential amino acids. In addition, we found genes associated with oxidative degradation of small lignin subunits, vitamin biosynthesis, sterol and fatty acid production, and nitrogen fixation, thus demonstrating that the gut microbiota are physiologically capable of helping ALB overcome its digestive and nutritional challenges.

MetRxn: Reaction/Metabolite Standardization and Congruency Across Databases and Genome-Scale Metabolic Models

Akhil Kumar

The ever accelerating pace of DNA sequencing and annotation information generation is spearheading the global inventorying of metabolic functions across all kingdoms of life. Increasingly, metabolite and reaction information is organized in the form of community, organism, or even tissue-specific genome-scale metabolic reconstructions. These reconstructions account for reaction stoichiometry and directionality, gene to protein to reaction associations, organelle reaction localization, transporter information, transcriptional regulation and biomass composition. A key barrier to the pace of extraction of metabolic knowledge from data is our inability to directly make use of metabolite/reaction information from databases (e.g., BRENDA, KEGG, BioCyc, UM-BBD, PubChem, ChEBI, Reactome.org, Rhea, etc.) or other metabolic

models due to incompatibilities of representation, duplications and errors. Therefore, the inadvertent inclusion of multiple replicates of the same metabolite, stoichiometrically inconsistent and/or elementally/charge unbalanced reactions can lead to erroneous model predictions and missed opportunities to reveal (synthetic) lethal gene deletions, repair network gaps and quantify metabolic flows. There have already been a number of efforts aimed at addressing some of these limitations. The Rhea database aggregates reaction data primarily from IntEnz and ENZYME whereas Reactome.org is a collection of reactions primarily focused on human metabolism. Motivated by this challenge we recently carried out an initial construction of the web-based resource MetRxn that integrates, using internally consistent descriptions, metabolite and reaction information from 6 databases and 34 metabolic models. The MetRxn content generation follows the general steps outlined in Figure 1. Metabolite and reaction data was first downloaded from BRENDA , KEGG and BioCyc using a variety of methods based on protocols such as SOAP, FTP and HTTP. We subsequently pre-processed the data into flat files that were imported into MetRxn. For all 34 genome-scale models ancillary information culled form the corresponding publications was also imported. The "raw data" from both databases and models was unified using standard SQL scripts on a MySQL server. Metabolites were also annotated with Canonical SMILES using the OpenBabel Interface from Chemspider. Metabolites with missing structural information were re-visited during the reaction reconciliation step. Using the corrected metabolite elemental composition and protonation states, reactions are evaluated for charge and elementally balance. We used a linear optimization program to charge and elementally balance all reactions.

Deep sequencing reveals persistence of intra- and inter-host genetic diversity in natural and greenhouse populations of *Zucchini yellow mosaic virus*

Heather Simmons

The genetic diversity in populations of RNA viruses is likely to be strongly modulated by their life-history, including mode of transmission. However, how transmission mode shapes patterns of intra- and inter-host genetic diversity, particularly when acting in combination with *de novo* mutation, population bottlenecks, and the selection of advantageous mutations is still poorly understood. To address these issues, we performed in-depth next generation sequencing of Zucchini yellow mosaic virus (ZYMV) in a wild gourd, *Cucurbita pepo* ssp *texana*, under two conditions: aphid-vectored and mechanically inoculated, achieving an average coverage of ~9000X. We show that mutations persist during inter-host transmission events in both the aphid vectored and mechanically inoculated populations, suggesting that the vector-imposed transmission bottleneck is not as extreme as previously supposed. Similarly, mutations were found to persist within individual hosts, arguing against strong systemic bottlenecks. Strikingly, mutations were seen to go to fixation in the aphid vectored plants, suggestive of a major fitness advantage, but remained at low frequency in the mechanically inoculated plants.

A data pipeline for analyzing large-scale comparative ChiP-Seq data to study binding patterns of transcription factors and histone modifications in human genome

Xiaokang Pan

With the generation of large-scale comparative ChIP-Seq datasets in studying the binding patterns of transcriptional factors and histone modifications in human genome associated with the treatment of leukemia disease in our laboratory, the analysis of these comprehensive datasets

is a challenge task. We have developed a data pipeline to deal with these large-scale comparative datasets. This data pipeline consists of three modules: peak calling, visualization and analysis. The peak calling module is to detect peaks consisting of a set of statistically significant binding reads in the genome. It currently includes two publicly available programs, CisGenome and SISSRS. The visualization module was designed to visualize peak and tag distributions for both experimental treatment and control over the whole genome and in more detail relating to specific genomic regions or genes. This module consists of a MySQL database and a Web-based graphical interface, being coded in Perl. The analysis module was built by a set of Perl scripts integrating several computational approaches to analyze the binding sites. This model is also aided by an existing program MEME for motif search. In this poster, we describe the design and development of this data pipeline and then show a set of example output results including peak call, visualization, distribution and motif analysis. We also discuss the future enhancements of this data pipeline.

Evolution of complex gene structures in the 12 *Drosophila* genomes

Luyi Wo

Drosophila genome has complex gene structures, including overlapping, embedded and interdigitated genes. It is important to understand the structure of genes when annotating higher order changes to the genome such as chromosomal rearrangements. The completion of 12 *Drosophila* genomes and the well supported annotation of gene models provided a great opportunity to examine complex gene structures in *Drosophila*. This study targets complex gene structures in these species to investigate the degree of conservation and the underlying evolutionary mechanisms for gene reorganization. A total of 4696 overlapping genes were discovered in the *D. melanogaster* genome, representing more than 30 % of coding genes in the genome. Transcripts with overlapping ends and genes embedded in the introns of other genes were the most common cases observed while genes where the peptide coding sequence overlapped, polycistronic messages, and interdigitated gene clusters were detected less frequently. We used gene order information from CAF1 annotation and supplemented this information with new analyses that define 1:1 orthologs. A conserved call was made where multi-gene structures are the same among all species as opposed to a non-conserved case where genes undergo rearrangements destroying the structure. Overall, the overlapping genes are not strictly conserved among 12 *Drosophila* species. Conservation percentage decreases with increasing divergence time (highest 90% in *D. sechellia* and lowest 70% in *D. grimshawi*). This study provides quantitative measure of multi-gene structure conservation in *Drosophila*, may elucidate the overlapping gene evolution, and will contribute to inferences about ancestral gene order and structure of *Drosophila* species.

Comparative genomic insights into interactions of the bacterial wilt pathogen *Erwinia tracheiphila* with plant and insect hosts

Lori Shapiro

Erwinia tracheiphila (Enterobacteriaceae), the causal agent of bacterial wilt of cucurbits, is a phytopathogen with a narrow host range restricted to susceptible squashes, pumpkins, melons, cucumbers, and gourds. *Erwinia* and other phytopathogenic genera are often overlooked members within Enterobacteriaceae, a group most commonly known for animal-associated bacteria including *Yersinia, Salmonella, Shigella*, and *Escheria* ssp. Bacterial wilt disease is vectored exclusively by two species of cucumber beetles and is an ecologically and economically

important pathogen of cultivated and wild cucurbits in the eastern US. However, mechanisms of virulence to plants hosts and factors mediating growth and attachment to the insect vector are currently unknown in this species. To begin to address this shortcoming, we used 454 Titanium chemistry to sequence the genome of an *E. tracheiphila* culture isolated from the wild gourd *Cucurbita pepo* ssp. *texana* from the Rock Springs Research Farm to 20X coverage. Individual reads were assembled into a single scaffold (~5.12 Mb) using MIRA assembler. The genome was annoted using RAST and the annotation was checked manually in Artemis. We found that the closest sequenced relative is the rosaceous floral symbiont *Erwinia billingeae* based on Blastp. The *E. tracheiphila* genome shows evidence of extensive horizontal gene transfer and is divergent from other sequenced enteric plant pathogens. Through comparative genomic analyses with other sequenced plant-associated *Erwinia*, we describe possible mechanisms of virulence to plants hosts. Comparisons with other enteric bacteria and preliminary horizontal gene transfer analyses are utilized to hypothesize traits mediating attachment and growth to insect vectors. Implications for the expression of virulence, survival in alternating plant and insect hosts, and consequences for transmission are discussed.

DNA Methylation of the Obesity Gene (lep) Promoter in Cancer Cells

Naomi Yamada

Leptin, the 16-kDa protein product of the obesity gene, is normally synthesized by adipocytes and to some extent by placental cells. Although the gene was discovered because of its vital role in maintenance of body mass and energy balance, leptin has since emerged as a master hormone with diverse physiologic functions ranging from regulation of the immune, cardiac, and reproductive systems to controlling bone physiology. As may be expected of such a multifunctional ligand, altered expression of leptin is associated with several pathologies. In recent years, leptin has been found to be up-regulated in breast cancer cells and research conducted at our laboratory has revealed it's ectopic expression in various soft tissue sarcoma subtypes. Enhanced expression and signaling through it's receptor leads to activation of a number of cellular pathways including the JAK2/STAT3 and ERK1/2 cascades and the PI-3K/Akt/GSK3 growth/anti-apoptotic pathway. The resultant increase in angiogenesis, cellular proliferation, migration, and local inflammation, as well as the concomitant suppression of apoptosis, strongly suggest the importance of this cytokine in cancer development. Expression of the leptin transcript is regulated by various processes including reversible epigenetic alteration of the promoter region. While it is know that epigenetic up-regulation of leptin transcription occurs under conditions of stress such as hypoxia and serum deprivation, till date, the role of DNA methylation in the transcriptional regulation of leptin in cancer cells has not been investigated. Hence, our laboratory has initiated a program to investigate methylation patterns in the leptin gene in cancer cell lines representing connective tissue fibrosarcoma as well as carcinoma of the breast and various other tissues. Since methylation of CpG islands near promoters of cancer-associated genes greatly influences the process of malignant conversion, our initial studies have focused on a 570 bp CpG island containing 64 CpG sites located in the region of the leptin promoter. Amplification and sequence analysis of this segment using bisulfite-modified template DNA indicates that malignant cells have a high overall rate of methylation in this DNA segment and exhibit distinct patterns and extent of methylation at specific CpG sites within the island as compared to normal cells. Overall, our study presents and highlights the first assessment of DNA methylation patterns of the leptin gene at the nucleotide level in cancer cells and may provide a molecular epigenetic explanation of the link between obesity, inflammation,

and the increased risk of cancer development.

Global epigenomic changes in histone acetylation and methylation following inhibition of CK2 kinase in acute leukemia cells

Chunhua Song

Optimal Designs for Two-color Microarray Experiments

Marcus Nunes

The explosion of genomic research over the last two decades has generated a huge amount of data. One important source of DNA genomic data has been the use two-color microarrays. Viewing this data through an ANOVA perspective, two-color microarray experiments can be considered Balanced Incomplete Block Designs (BIBD) with two treatments per block. In our study, each microarray is a block and each sample of interest a treatment, which are assigned to one of two different dye colors. The goal of this presentation is to show the class of optimal designs according to A-, D-, and E-optimality criteria. This class of microarray experiments is limited to the cases when the number of blocks and treatments are equal.

Functional assessment of human coding polymorphisms in *SLC45A2* and *SLC24A5* in zebrafish

Zurab Tsetskhladze

Genome-wide association (GWA) studies link candidate genetic polymorphisms with human diseases and physical traits. Experimental verification of causality, however, remains a challenge. In zebrafish, phenotypic rescue using mRNA is an established method of connecting candidate genes with mutants. Rescue of zebrafish phenotypes using human mRNAs can theoretically be used to test human polymorphisms, but is not always possible. Here, we have used pigmentation as a readily scorable phenotype to determine whether "humanized" zebrafish mRNAs can be used for functional testing of human alleles. Zebrafish mRNAs for *slc24a5* and *slc45a2* rescue mutants for their corresponding genes (*golden* and the newly cloned *albino*, respectively). Introducing human coding polymorphisms *A111T* and *L374F*, respectively, abrogates rescue. In contrast, rescue persists in the presence of a second, naturally-coding human polymorphism, *E272K,* in *SLC452*. This approach, "humanized zebrafish orthologue rescue", or HuZOR, may potentially detect functional effects of human alleles in other genes.