UNIVERSITY OF MARYLAND

OMICS DAY

May 22, 2012
8:30 am – 5:30 pm

A Meeting to Celebrate High Throughput Biology Research at Maryland and to Foster New Collaborations

Special Thanks

to the

UNIVERSITY OF MARYLAND

OMICS DAY

Sponsors

nanoString
TECHNOLOGIES

ESAC
www.esacinc.com

illumina®

Covaris™
the sample prep advantage

CBCB
CENTER FOR BIOINFORMATICS & COMPUTATIONAL BIOLOGY

UMIACS

UNIVERSITY of MARYLAND
SCHOOL OF MEDICINE
PROGRAM in PERSONALIZED and GENOMIC MEDICINE

UNIVERSITY SYSTEM OF MARYLAND

OMICS DAY
Schedule

8:30	Registration and Coffee
9:00	Sridhar Hannenhalli – Welcome
9:10	Mihai Pop – The University of Maryland Center for Bioinformatics and Computational Biology
9:30	Tom Kocher – Genomic approaches to studying adaptation and speciation in Lake Malawi cichlid fish
9:50	Scott E. Devine – Transposable elements in humans and model organisms
10:10	John Moult – Genetic Variation and Disease
10:30	Coffee Break
10:50	Jacques Ravel – The Vaginal Microbiota in Health and Disease
11:10	Ed Eisenstein – Elucidating Gene-Metabolite Relationships in Non-model Medicinal Plants
11:30	Alan R. Shuldiner – Amishomics: Linking genomes with phenomes
11:50	Najib El-Sayed – From Genomes to Host-Pathogen Infectomes: Models in profiling host-pathogen interactions
12:10	Lunch
1:30	Susan Dorsey – Using transcriptomic analysis to gain mechanistic insight into muscular dystrophy
1:50	Kan Kao – Alterations in DNA-Nuclear Lamina Interactions and Nuclear Organization in Hutchinson-Gilford Progeria Syndrome
2:10	Hector Corrada-Bravo – Increased methylation variation in epigenetic domains across cancer types
2:30	Junhyong Kim – Keynote Lecture: All the genomes, all the cells: Towards single cell genomics
3:30	Poster Session and Mixer (See Abstracts Pages 3-14)

OMICS DAY
Poster Session Abstracts

POSTER 1

Hierarchical Classification of Biological Sequences

Bo Liu, Mihai Pop

A challenging bioinformatics problem is: given a set of homologous sequences whose labels form a hierarchical tree (e.g., NCBI taxonomy), how to accurately classify anonymous query sequences. Previous approaches usually treat this as a flat multi-class classification problem. We developed a novel algorithm: (1) MetaPhyler explicitly considers the structure of the labels during training and classification; (2) MetaPhyler computes a confidence score, estimated from nonparametric kernel density, for each classification at each hierarchical level; (3) Classification step is very fast, enabling large-scale analysis; (4) Based on our evaluation, it is more accurate than previous approaches. Our software is opensource and is available at: http://metaphyler.cbcb.umd.edu/

POSTER 2

AGORA: Assembly Guided by Optical Restriction Alignment

Henry C. Lin, Steve Goldstein, Lee Mendelowitz, Shiguo Zhou, Joshua Wetzel, David C. Schwartz, Mihai Pop

Genome assembly is difficult due to repeated sequences within the genome, which create ambiguities and cause the final assembly to be broken up into many separate sequences (contigs). Long range linking information, such as mate-pairs or mapping data, is necessary to help assembly software resolve repeats, thereby leading to a more complete reconstruction of genomes. Prior work has used optical maps for validating assemblies and scaffolding contigs, after an initial assembly has been produced. However, optical maps have not previously been used within the genome assembly process. Here, we use optical map information within the popular de Bruijn graph assembly paradigm to eliminate paths in the de Bruijn graph which are not consistent with the optical map and help determine the correct reconstruction of the genome.

POSTER 3

Cell-type specific networks in Brassica napus guard cell responses to drought

Florent Villiers, Faris Alqadah, Jeffrey Anderson, Felix Hauser, Byeong Wook Jeon, Dongdong Kong, Jade Lee, Jack Mullen, Jun- Kuk Na, Sarah M. Assmann, Joel S. Bader, John K. McKay, Scott C. Peck, Julian I. Schroeder, June M. Kwak

To elucidate networks of guard cell signaling pathways and genes regulated in response to abscisic acid (ABA) and drought and to under-stand how endogenous and environmental cues alter their expression and interaction, we will use a cell-type specific systems approach to build network models enabling researchers to conduct cell-type specific genetics and genomics and to manipulate guard cells towards developing practical strategies for improving water stress tolerance of B. napus and other crop plants.

POSTER 4

A polymorphic T-rich element in ATP1B1 is associated with blood pressure and regulates alternative Polyadenylation

Megana Prasad, Kavita Bhalla, Zhen Hua. Pan, Jeff O'Connell, Alan Weder, Aravinda Chakravarti, Bin Tian, and Yen-Pei Christy Chang

Genome-wide linkage and association studies of hypertension (HTN) aim to uncover loci that are involved in blood pressure (BP) regulation and the pathophysiology of HTN. However, even when such loci are successfully identified, the mechanism by which the implicated genes increase HTN susceptibility is not well understood. ATP1B1, which encodes the beta subunit of Na+, K+ ATPase, a cotransporter involved in multiple BP-regulating physiological processes, is located within a well-established BP linkage peak. Although a single nucleotide polymorphism (rs12079745) in the 3'UTR of ATP1B1 has previously been shown to be associated with BP levels, the molecular mechanism underlying this association is unknown. We identified a multi-allelic T-rich element (TRE) in the 3'UTR of ATP1B1 that varies in length and sequence composition (T22-27 and T12GT3GT6). The 3'UTR of ATP1B1 contains 2 functional polyadenylation signals and the TRE is downstream of the proximal site (A2). Because U-rich elements (UREs) in 3'UTRs are known to play a role in the cleavage and polyadenylation of mRNA, and ATP1B1 transcripts with shorter 3'UTRs are translationally more efficient, we hypothesized that alleles of this TRE might influence ATP1B1 expression by regulating alternative polyadenylation. Consistent with our hypothesis, the A2-polyadenylated mRNA isoform was more abundant in human kidneys with at least one copy of T12GT3GT6 than in those homozygous for the T22-27 alleles. Functionally, in vitro, we demonstrated that the T12GT3GT6 allele has greater luciferase activity than the T22 allele and that the T12GT3GT6 allele shows increased polyadenylation at the A2 site than the T22 allele. In addition, the TRE element is associated with systolic BP in European Americans, with the T12GT3GT6 allele being associated with higher BP (GenNet network of the NHLBI Family Blood Pressure Program, effect size = 2.4 mmHg SBP, P = 0.003). Sharing strong linkage disequilibrium with rs12079745, this TRE is likely the functional site underlying the original association. In summary, we have identified a novel multi-allelic TRE in the 3'UTR of ATP1B1. Alleles at this site are associated with HTN and may mediate their effect on BP by regulating polyadenylation of the ATP1B1 mRNA.

POSTER 5

A genome-scale identification of loci required for the fitness of the invasive M1T1 Group A Streptococcus in human blood

Yoann Le Breton(1), Pregnish Mistry(1), Kayla Valdes(1), Nikhil Kumar(2), Hervé Tettelin(2) and Kevin S. McIver(1)
(1) Dept of Cell Biology and Molecular Genetics and Maryland Pathogen Research Institute, University of Maryland, College Park, MD; (2) Institute for Genome Sciences, University of Maryland School of Medicine, Baltimore, MD

Streptococcus pyogenes (Group A Streptococcus or GAS) is a strict human pathogen responsible for a wide spectrum of diseases ranging from uncomplicated superficial infections of the throat (Strep throat) and skin (impetigo) to life-threatening invasive diseases such as streptococcal toxic shock syndrome and necrotizing fasciitis ("flesh-eating" disease). Over the last decade, considerable efforts have been made to investigate the GAS genome and its genetic variability within different clinical isolates. To now explore this wealth of information and understand the molecular mechanisms important for GAS pathogenesis at different stages of infection, high-throughput functional genomic approaches needed. We present the development and application of a whole-genome methodology, Transposon Site Hybridization (TraSH), for the identification of genes required during GAS infection. First, a new mariner-based transposon, osKaR, designed to perform TraSH screens was created and successfully tested in several GAS invasive serotypes. A near-saturation mutant library was then established in the invasive M1T1 GAS strain 5448AN. To identify genes that are important for GAS fitness in human blood, the mutant library (input pool) was subjected to a negative selection in human blood.

Mutants underrepresented in the output pool of mutants were identified using DNA microarray hybridization. Our screen generated a list of genes that are likely to be important for fitness of GAS in its human host; including genes that were previously shown to play an important role in GAS virulence. In addition, other loci related to different aspects of GAS physiology (such as carbon metabolism, nucleotide synthesis, protein synthesis, gene regulation) were identified. We have further validated these results by showing that several individual mutants of genes identified in our TraSH screen are defective for survival in human blood. In the future, the new oskar system will now be used in conjunction with next-generation sequencing technologies (Tn-seq) to investigate the pathogenesis of GAS.

POSTER 6

Kelly O'Quin

Studies of regulatory evolution have emphasized a central role for cis-regulatory elements in the evolution of gene expression and phenotypic adaptation. Mutations within trans-regulatory factors also contribute to regulatory evolution, but these have been harder to study since the genomic position of these factors relative to the genes they regulate is unknown. African cichlid fishes vary adaptively in their regulation of six opsin genes that are necessary for color vision. To identify the genetic factors responsible for this variation, we used experimental crosses of cichlids that differ in opsin expression to scan for Expression Quantitative Trait Loci (eQTL). We generated 115 individuals from three hybrid F2 families and genotyped them at over one-thousand differentially-fixed single nucleotide polymorphisms by sequencing restriction site associated DNA tags. Our results identified four eQTL: two associated with SWS2B and SWS2A expression, one associated with RH2B expression, and one associated with RH2A and LWS expression. Only one of these eQTL occurred in cis (on the same linkage group) to the six opsins genes, though none of the opsins were part of the associated region. The remaining three eQTL all occurred in trans (on different linkage groups) to the opsins. As part of this study, we also included 13 candidate genes that are known to regulate vertebrate opsin expression in trans, though none of these occurred within the four eQTL. Our results demonstrate that trans-regulatory factors can also play an important role in the evolution of gene regulation and phenotypic adaptation. Previous work has shown that several protein-coding mutations and regulatory factors tune vision in these fishes. We then generated a genetic map from these markers and performed a regression analysis of opsin expression and marker genotypes. Future work will fine-map these eQTL to further resolve the role that mutations within protein-coding genes, cis-regulatory elements, and trans-regulatory factors play in the adaptive evolution of cichlid color vision. Comparative mapping of the associated markers to a draft assembly of the cichlid genome reveals other excellent candidate genes within each eQTL.

POSTER 7

Functional characterization of Mucin-Associated Surface Protein (MASP) in the human parasite Trypanosoma cruzi

Jungmin Choi, Maria Cecilia Fernandes Qian Cai, Gustavo Cerqueira, Zu-Hang Sheng and Najib M. El-Sayed

MASPs are members of a multigenic family recently identified during the sequencing of the T. cruzi CL Brener genome. This family contains around 1,400 members, consisting of approximately 6% of the diploid genome. Highly conserved N- and C-terminal domains, which encode a signal peptide and GPI-anchor addition site respectively, and a hypervariable central region, characterize MASPs. Members of this family are predominantly expressed in the infective trypomastigote form. We hypothesized that members of the T. cruzi MASP protein family play a major role in the interaction of the parasite with the host cell. In order to investigate a putative role for T. cruzi MASP at the host-pathogen interface, we have used MASP as a bait protein against the human proteome using a high-throughput platform that we have recently established for identifying protein-protein interactions between pathogens and theirs hosts. Yeast two-hybrid screens identified human SNAPIN as one of two major MASP interacting proteins. SNAPIN is a member of the SNARE protein complex, which may have a role in a calcium-dependent exocytosis. The MASP-SNAPIN interaction was further validated using in vivo co-Affinity Purification and in vitro pull-down assays. Immunofluorescence assays showed human SNAPIN is recruited to the parasite surface during invasion. Co-localization experiments indicated that SNAPIN is associated with the late endosomes and lysosomes. Supporting our initial hypothesis, SNAPIN depletion using siRNA oligomers in HeLa cells and snapin-/- in Mouse Embryonic Fibroblast (MEF) cells significantly inhibited T. cruzi invasion, suggesting a role for SNAPIN in this process. Lysosomes in snapin-/- MEF cells displayed aberrant morphology and distribution and the parasites did not recruit host lysosomes efficiently when compared to wild-type cells. This was likely due to an impaired calcium-dependent lysosome exocytosis in snapin-/- MEF cells. SNAPIN was translocated to the plasma membrane upon calcium influx induced by a calcium ionophore (Ionomycin), resulting in the exposure of the luminal domain of SNAPIN to the extracelluar space. Leishmania tarentolae transgenic strains expressing two different MASP proteins were shown to trigger intracellular calcium transients in HeLa cells, presumably by injuring the cell membrane. We propose that T. cruzi MASP plays a role in wounding the plasma membrane of the host cell, which in turn elicits a transient intracellular calcium flux and leads to the translocation of lysosome-associated SNAPIN to the plasma membrane. Human SNAPIN, through its exposed luminal domain would then provide an anchor for the entry to the parasite into the cell. The mechanism of T. cruzi MASP evoked calcium influx in the host cell membrane remains under investigation.

POSTER 8

Widespread evidence of viral miRNAs targeting host pathways

Joseph W. Carl, Jr., Joanne Trgovcich, Sridhar Hannenhalli

MicroRNAs (miRNA) are regulatory genes that target other RNA molecules via sequence-specific binding and mediate their repression. Several biological processes are regulated by evolutionarily conserved miRNAs across many organisms. In addition, plants and invertebrates are known to employ their miRNA in defense against viruses by targeting and degrading viral. On the other hand, viral miRNAs can hijack or subvert host cellular pathways to enable efficient replication of the virus. Viruses also encode miRNAs and there is evidence to suggest that virus-encoded miRNAs target specific host genes/pathways that may be beneficial for their infectivity and/or proliferation. However, it is not clear whether there are general patterns underlying cellular targets of viral miRNAs. Based on 6809 putative miRNAs encoded by 23 human viruses, our analysis suggests that several human viruses have evolved their miRNA repertoire to target specific human pathways, such as cell growth, axon guidance, and cell differentiation. Many of the same pathways are also targeted in mouse by miRNAs encoded in murine viruses. Overall, our results suggest that viruses may have evolved their miRNA repertoire to target specific host pathways as a means for their survival.

POSTER 9

Which assembly is better? A probabilistic method for de novo assembly evaluation

Christopher Hill

The genome of an organism usually consists of one or a few long DNA sequences. Unfortunately, current sequencing technology can only “read” relatively short pieces of DNA. The goal of genome assembly is to reconstruct the original genome, given these sequence fragments. There have been several attempts at formulating genome assembly as an algorithmic problem: shortest superstring, string graph, De Bruijn graph, etc. Each formulation tries to optimize a slightly different objective function. In addition, there are many other common criteria used for evaluating assemblies, such as the popular N50. These criteria, however, are not suitable because the solution that maximizes them is not the best assembly. For example, the concatenation of all of the reads results in an “assembly” with a very large N50, which is obviously not a good solution. In this work, we describe our objective function for evaluating genome assemblies: the probability of an assembly being the truth, given our observations (i.e. the set of reads). We can prove that the true genome maximizes our probabilistic criteria. Furthermore, we show that this likelihood strongly correlates with results presented by the popular assembler competitions GAGE and Assemblathon1.

POSTER 10

De novo detection of copy number variation

Jurgen F. Nijkamp, Marcel A. van den Broek, Jan-Maarten A. Geertman, Marcel J.T. Reinders, Jean-Marc G. Daran and Dick de Ridder

Comparing genomes of individual organisms using next generation sequencing (NGS) data is, until now, mostly performed using a reference genome. This is challenging when the reference is distant and introduces bias towards the exact sequence present in the reference. Recent improvements in both NGS read length and efficiency of assembly algorithms have brought direct comparison of individual genomes by de novo assembly, rather than via a reference genome, within reach.

Here, we develop and test a Poisson mixture model (PMM) for copy number estimation of contigs assembled from NGS data. We combine this with co-assembly to allow de novo detection of copy number variation between two individual genomes, without mapping reads to a reference genome. In co-assembly, multiple sequencing samples are combined, generating a single contig graph with different traversal counts for the nodes and edges between the samples. In the resulting `colored' contig graph the contigs have integer copy-numbers; this negates the need to segment genomic regions based on depth of coverage, as required for mapping-based detection methods. The PMM is then used to assign integer copy numbers to contigs, after which copy number variation probabilities are inferred.

The copy number estimator and copy number variation detector perform well on simulated data. Application of the algorithms to hybrid yeast genomes showed allotriploid content from different origin in the wine yeast Y12, and extensive copy number variation in the aneuploid brewing yeast genomes. Integer copy number variation was also accurately detected in a short-term laboratory evolved yeast strain.

POSTER 11

The Functional G143E Variant of Carboxylesterase 1 is Associated with Increased Clopidogrel Active Metabolite Levels and Greater Clopidogrel Response

J. P. Lewis, R. B. Horenstein, K. Ryan, J. R. O'Connell, Q. Gibson, B. D. Mitchell, K. Tanner, R. Pakzy, K. P. Bliden, U. S. Tantry, C. J. Peer, W. D. Figg, S. D. Spencer, M. A. Pacanowski, P. A. Gurbel, and A. R. Shuldiner.

Carboxylesterase 1 (CES1) is the primary enzyme responsible for converting clopidogrel into biologically inactive carboxylic acid metabolites. We genotyped a functional variant in CES1, G143E, in participants of the Pharmacogenomics of Anti-Platelet Intervention (PAPI) Study (n=566) and in 350 coronary heart disease (CHD) patients treated with clopidogrel, and performed association analysis of bioactive metabolite levels, on-clopidogrel ADP-stimulated platelet aggregation, and cardiovascular outcomes. Levels of clopidogrel active metabolite were significantly greater in CES1 143E-allele carriers (P=0.001). Consistent with these findings, individuals who carried the CES1 143E-allele had better clopidogrel response as measured by ADP-stimulated platelet aggregation in both PAPI Study participants (P=0.003) and clopidogrel-treated CHD patients (P=0.03). No association was observed between this SNP and baseline measures of platelet aggregation in either cohort. Taken together, these findings suggest, for the first time, that genetic variation in CES1 may be an important determinant of clopidogrel efficacy.

POSTER 12

Identifying deleterious effects on splicing with SplicePort

Stephen M. Mount[1,*], Anna Flynn[1,4], Chiao Feng Lin[2], Li-San Wang[2], Allison Abad[1,3],
Zaneta Franklin[1,4] and Rezarta Islamaj Doğan[5]
1. Dept. of Cell Biology and Molecular Genetics and Center for Bioinformatics and Computational Biology, University of Maryland, College Park, MD 20742. 2. Penn Center for Bioinformatics, 1424 Blockley Hall, 423 Guardian Drive, University of Pennsylvania, Philadelphia, PA 19104, USA. 3. Science, Math, and Computer Science program, Poolesville High School, Poolesville, MD 4. Science and Technology program, Eleanor Roosevelt High School, Beltsville, MD 5. National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland, USA

Sequences in the vicinity of splice sites, but outside of the core splice site consensus nucleotides immediately surrounding the splice site, play an important but underappreciated role in splice site selection. SplicePort (spliceport.org, reference 1) is a tool for splice-site analysis. SplicePort implements a feature generation algorithm for the classification of potential splice sites, scoring each GT or AG dinucleotide using features within a window of 162 nucleotides. We have been exploring the utility of SplicePort for identification of sequence variants that affect gene expression through splicing.

The score provided by SplicePort is not in and of itself directly meaningful. Therefore, in order to distinguish neutral from deleterious mutations we compared the change in SplicePort score to two distributions. One is the distribution of score changes due to every possible single-nucleotide change in all human RefSeq splice sites. Based on this distribution we classified positions into eight categories (exon 5' SS (-80 to -4), core 5' SS (-3 to +6), intron 5' SS (+7 to +82); intron 3' SS (-82 to -27), branchpoint (-26 to -19), pyrimidine tract (-18 to -4), core 3' SS (-3 to +1) and exon 3' SS (+2 to +82). The other relevant distribution of score changes was obtained from the literature. 63 non-core 5' splice site mutations and 103 non-core 3' splice site mutations affect splicing in monogenic genetic diseases. To assess the impact of single nucleotide polymorphisms the SplicePort score change associated with each mutation is compared to these two distributions, and variants are predicted to be deleterious (Candidate Spicing Impact SNPs) if their SplicePort score change was negative and greater than the median score change for mutations known to be deleterious. These critieria yield 66,261 candidate splicing impact SNPs in the human genome (using hg19), and their true effects are currently under investigation. Ongoing work is directed towards the evaluation of this large sets of single nucleotide variants, and the development of a pipeline for analysis of exome data.

References:

1. Dogan RI, Getoor L, Wilbur WJ and Mount SM. 2007. SplicePort -- an interactive splice site analysis tool. Nucleic Acids Res. 35:W285-91

POSTER 13

Gene Order Comparison With Contigs And Scaffolds

Adriana Munoz , Chunfang Zheng , Qian Zhu, Victor A. Albert, Steve Rounsley,
David Sankoff

There has been a trend in increasing the phylogenetic scope of genome sequencing while decreasing the quality of the published sequence for each genome. With reduced "finishing" effort, there are an increasing number of genomes being published in the form of unanchored scaffolds or even as individual contigs. Rearrangement algorithms, including gene order-based phylogenetic tools, require whole genome data on gene order, segment order, or some other marker order. Items whose chromosomal location is unknown cannot be part of the input. The question we address here is, for gene order-based comparisons, how can we use rearrangement algorithms to handle genomes available in contig or scaffold form only? Our method involves optimally filling in genes missing in the scaffolds, while incorporating the augmented scaffolds directly into the rearrangement algorithms as if they were chromosomes. This is accomplished by an exact, polynomial-time algorithm. We then correct for the number of extra fusion/fission operations required to make scaffolds comparable to full assemblies. We model the relationship among scaffold density, rearrangement rate and genomic distance, and carry out simulations to estimate the parameters of this model. Based on these results, we compare the draft sequence of an angiosperm genome with the more polished genome sequence of Vitis vinifera.

POSTER 14

Optical Mapping: High Resolution Whole Chromosome Profiles and Markers for Foodborne Outbreak Strains of Escherichia coli O157:H7, Salmonella enterica, and Cronobacter spp

Michael. L. Kotewicz, David W. Lacher, and Mark. K. Mammel, FDA, CFSAN, OARSA, Laurel, MD 20708

Typical foodborne pathogens such as Escherichia coli, Salmonella enterica, and Cronobacter possess chromosomes ranging from 4.4 to 5.6 million base pair (bp), the optical maps of which contain 400-600 contiguous restriction fragments. Comparisons of sequenced reference genomes to optical maps of outbreak strains allow the precise sizing and mapping of chromosomal changes across entire genomes at a fragment resolution limit of about 800 bp. These variations in bacterial chromosomes include insertions, deletions, and inversions, most often involving prophages. We previously characterized in detail a number of isolates from the 2006 E. coli O157:H7 outbreak associated with spinach (ref. 2). Here we show the usefulness of optical mapping to differentiate strains involving E. coli O157:H7 from cookie dough-associated clinical samples, S. enterica Saintpaul associated with a tomato/pepper outbreak, and S. enterica Typhimurium associated with a peanut paste-related outbreak, as well as clinical and foodborne isolates of Cronobacter species (formerly Enterobacter sakazakii). The optical maps allow differentiation among outbreak species and measurements of changes within an outbreak. Comparisons of genomic changes can be made with information gained from ongoing DNA microarray, single nucleotide polymorphism (SNP), and sequencing analyses of these strains for isolate identification.

POSTER 15

Computational analysis of genetic variation underlying disease risk

Moult Lab–OMICS
Lipika R. Pal[1], Chen-Hsin Yu[1,2], Chen Cao[1,3], Maya Zuhl[1], and John Moult[1,4]
[1]Institute for Bioscience and Biotechnology Research, University of Maryland at College Park, Rockville, MD
[2]Molecular and Cellular Biology Program, University of Maryland at College Park
[3]Computaional Biology, Bioinformatics, and Genomics Program, University of Maryland at College Park, MD
[4]Department of Cell Biology and Molecular Genetics, University of Maryland at College Park, College Park, MD

The Moult Lab uses various computational genomics and bioinformatics approaches to analyze Human disease mechanisms, including those involved in monogenic disease, cancer, and other complex diseases. Genome-wide association studies (GWAS) of human complex disease have identified a large number of disease associated genetic loci, distinguished by a differing frequency of specific single nucleotide polymorphisms (SNPs) among individuals with a particular disease. However, these data do not provide direct information on the biological basis of a disease or on the underlying mechanisms. A variety of mechanisms may link the presence of a SNP to altered in vivo gene product function and hence contribute to disease risk. We have explored the role of two of these mechanisms using associations found in the Wellcome Trust Case Control Consortium (WTCCC1) seven disease study and follow-up work: Missense SNPs (msSNPs) in proteins and SNPs that affect expression level (exSNPs). Hapmap linkage disequilibrium data are first used to identify the set of candidate SNPs for involvement in disease mechanism in each disease related locus. For each candidate msSNP, we estimate impact on protein function using two computational methods we have previously developed, utilizing sequence profile and structure information (SNPs3D, http://www.snps3d.org). For exSNPs, data from diverse genome-wide expression quantitative trait loci (eQTL) association studies were integrated to form a high confidence gold standard set. This set was then mapped on to the WTCCC1 disease mechanism candidate SNPs to identify those causing an expression change. The results support a significant role for both types of SNPs in common human disease susceptibility. Interestingly, GWAS studies do not usually find existing drug targets. Analysis of 1000 genome data suggests this aberration is likely due to purifying selection in drug target genes. We are investigating methods of leveraging GWAS results to identify new potential drug targets. Our research endeavors have been broadened by the creation of a community wide experiment to assess the state of the art in this type of genome interpretation (CAGI, http://genomeinterpretation.org, in collaboration with Steven Brenner). In particular, the riskSNPs component of CAGI pools and compares the methods of multiple research groups in finding the mechanisms underlying the WTCCC1 disease loci. The results show that such consensus approaches are potentially powerful.

POSTER 16

Simultaneous transcriptome profiling of Trypanosoma cruzi parasites and their human host cells

Yuan Li[1], Barbara Burleigh[3], Najib M. El-Sayed*,[1,2]
[1]Department of Cell Biology and Molecular Genetics, and [2]Center for Bioinformatics and Computational Biology, University of Maryland, College Park, MD 20742, USA. [3]Department of Immunology and Infectious Diseases, Harvard School of Public Health, Boston, MA 02115, USA. * Corresponding author: elsayed@umd.edu

We are conducting simultaneous transcriptome profiling of Trypanosoma cruzi, the causative agent of Chagas disease, and their human host cells (foreskin fibroblasts) during the course of infection. We have successfully reconstructed the transcriptomes for the bloodstream (trypomastigote) form and the intracellular form (amastigote) form of the parasite. This has allowed the curation of the existing gene models, the identification of novel open reading frames, and the detection of the trans-splicing and polyadenylation sites at the single nucleotide level. Most interestingly, we have identified differential splicing and polyadenylation events in the pathogen before and after the invasion of human host cells. These events may be associated with the regulation of mRNAs at the post-transcriptional level. We're also examining the subsets of differentially expressed genes both in the parasite and the host cell over the course of the infection to gain insights into mechanisms of invasion and intracellular survival strategy as well as host-pathogen interactions. T. cruzi genes that are significantly regulated during the infection process may present new targets for drug development.

POSTER 17

Next generation sequencing in the study of bacterial pathogenesis: genetic screens and clonal analysis

Gregory T. Crimmins, Erin Green, Sina Mohammadi, Eddie Geisinger, David Mosser, Joan Mecsas, Ralph Isberg

We have explored next generation sequencing as a tool to study persistent, unanswered questions in bacterial pathogenesis. First, we have utilized the newly developed Tn-Seq method, using Illumina sequencing to screen for Yersinia pseudotuberculosis genes required for infection of mice. We screened over 20,000 unique transposon insertions mutants, covering over 3,100 genes (over 75% of the Yersinia pseudotuberculosis genes), for genes required for growth and persistence in mice. We identified over 17 novel Yersinia pseudotuberculosis virulence factors, including a mesenteric lymph node specific virulence factor, MrtAB. Further study of MrtAB and the other newly identified virulence factors will greatly increase our understanding of Yersinia pseudotuberculosis pathogenesis. Second, we generated over 16,000 uniquely tagged plasmids, designed to be amplified for Illumina sequencing. We are currently using these tagged plasmids to investigate the clonality of Yersinia pseudotuberculosis infections under various conditions, including the presence and absence of neutrophils, and low vs. high dose infections. The clonality of infections is unknown for most bacterial infections, leaving us with a large gap in our understanding of bacterial disease: how it is generated, how it is spread from organ to organ, and how the bacteria escape from their host. The data from our study, and the tools we are generating, will provide some preliminary answers to these important questions.

POSTER 18

Genomics and systems biology approaches to unravel secondary metabolism in black cohosh (Actaea racemosa L.)

Martin J Spiering, Bhavneet Kaur, Le Qi, and Edward Eisenstein
Institute for Bioscience and Biotechnology Research, University of Maryland, 9600 Gudelsky Dr, Rockville, MD
Email: spiering@umd.edu

Preparations of rhizomes and roots of black cohosh (Actaea racemosa L.; Ranunculaceae) have long been used by Native American herbal practitioners as analgesic and anti-inflammatory agents. Today black cohosh is used chiefly as dietary supplement to alleviate menopausal vasomotor symptoms. A. racemosa contains a rich diversity of secondary metabolites with biological activities. These include high-abundance signature metabolites, such as cycloartenol-derived triterpenoids and diphenolic esters likely assembled via phenylpropanoid/rosmarinic acid pathways, as well as low-abundance compounds, such as tryptamines and other alkaloids, arising from aromatic amino acid pathways. Here we present a comprehensive approach, utilizing genomics, functional expression of plant proteins, plant tissue culture, and high-throughput analytical chemistry tools to identify key genes in the biosynthesis of the secondary metabolites in A. racemosa. Bioinformatics analysis of a collection of 2000 expressed sequence tags (ESTs) along with homology-based cloning identified several gene candidates for production of each of the medicinally important compound classes in black cohosh. Functional expression of three tryptophan decarboxylase (TDC)-related genes from A. racemosa confirmed that they encode TDC activity, a pivotal step in the biosynthesis of bioactive tryptamines, including serotonin-derivatives. These findings, along with the successful development of a regenerative plant growth platform and a cell suspension culture system for black cohosh, have opened up the way to systems biology approaches to elucidate gene–metabolite relationships in this important medicinal plant with "omics" methods, such as metabolite profiling by LC-MS and quantitative gene expression analysis by RNA-Seq.

POSTER 19

An Anti-Profile Support Vector Machine

Hector Corrada Bravo, Wikum Dinalankara

Recent studies have shown that DNA methylation in genomic regions involved in tissue differentiation and development show hyper-variability in cancer with respect to stable epigenetic regulation in corresponding normal tissues. Using methylation levels displayed in these regions, it is possible to perform classification and prediction between healthy and carcinoma samples. In hyperplastic, but otherwise benign tissues, DNA methylation in these regions show intermediate levels of hyper-variability in comparison with normals and carcinoma. Our goal is to develop classification methods to distinguish samples from these cancer stages (adenoma vs. carcinoma) based on measurements from hyper-variable regions. To this end, we develop the Anti-Profile SVM, a maximum-margin classifier which uses an indirect kernel function to model similarity to an external stable population to distinguish samples from unstable, highly variable populations. We present preliminary results regarding the accuracy and stability of the classifier.

POSTER 20

A model for detecting hyper-variably methylated CpGs in epigenetic domains across cancer types

1.Kwame Okrah, 2.Hector Corrada Bravo and 2. Hao Wang.
1. Center for Bioinformatics and Computational Biology, Department of Mathematics, University of Maryland, College Park 2. Center for Bioinformatics and Computational Biology, Department of Computer Science, University of Maryland, College Park

Increased variability of DNA methylation in genomic domains involved in tissue differentiation and development has been shown recently to be a consistent and defining mark across different cancer types. These hyper-variable regions many times coincide with hypo-methylated regions in cancer. Most statistical procedures for testing unequal variance are not adept at detecting hyper-variability in the presence of changes in variance that are concomitant with changes in mean levels. We introduce a model for detecting differentially variable measurements that models expected variance differences resulting from mean differences in a non-parametric manner, and thus detect those measurements that show variance at levels higher than expected under a given mean difference. We apply our model to methylation data obtained from the Illumina 450k methylation chip and show the advantages of this model over traditional testing methods.

POSTER 21

Time-series analysis of metagenomic data

Hisham N. Talukder[12], Joseph N. Paulson[12], Héctor Corrada-Bravo[23]
[1]Applied Mathematics and Scientific Computing, University of Maryland, College Park
[2]Center for Bioinformatics and Computational Biology, University of Maryland, College Park, [3]Department of Computer Science, University of Maryland, College Park

Metagenomics is the study of genetic material recovered from an environmental sample.
In many studies, the goal is to determine if the abundance of one or more organisms is correlated with some characteristics of the sample, e.g., health or disease status of a host organism. Studying the dynamics of a Microbiome is essential, in particular for longitudinal studies. We provide a statistical framework using smoothing splines to analyze metagenomic time-series data and detect differences between groups over time. We propose a model that uses smoothing splines on normalized counts obtained from sequencing of 16S genes to detect differences in abundance over time in samples from two populations.
We applied our methods to existing data consisting of two groups of mice used to determine differences in gut microbiomes resulting from differences in diet: twelve germ-free mice were gavaged with human fecal microbiota from a healthy donor and fed a low-fat, plant-polysaccharide-rich (LF/PP) diet for four weeks. Subsequently, half of the mice were switched to a high-fat/high-sugar Western diet. We show and discuss the advantage of using a smoothing spline method for this task.

POSTER 22

Metastats: an improved statistical method for analysis of metagenomic data

Joseph N Paulson[12], Héctor Corrada Bravo[23], Mihai Pop[23]
[1]Applied Mathematics and Scientific Computing, University of Maryland, College Park, [2]Center for Bioinformatics and Computational Biology, University of Maryland, College Park, [3]Computer Science, University of Maryland, College Park

Metagenomic studies, originally focused on exploratory/validation projects, are rapidly being applied in a clinical setting. In this setting, researchers are interested in finding characteristics of the microbiome that correlate with the clinical status of the corresponding sample. Comparatively few computational/statistical tools have been developed that can assist in this process, rather most developments in the metagenomics community have focused on methods that compare samples as a whole. Specifically, the focus has been on developing robust methods for determining the level of similarity or difference between samples, rather than identifying the specific characteristics that distinguish different samples from each other.

Metastats [1] was the first statistical method developed specifically to address the questions asked in clinical studies. Metastats allows a comparison of metagenomic samples (represented as counts of individual features such as organisms, genes, functional groups, etc.) from two treatment populations (e.g., healthy vs. disease) and identifies those features that statistically distinguish the two populations.

Here we present major improvements to the Metastats software and the underlying statistical methods. First we describe new approaches for data normalization that enable a more accurate assessment of differential abundance by reducing the covariance between individual features implicitly introduced by the traditionally used ratio-based normalization. These normalization techniques are also of interest for time-series analyses or in the estimation of microbial networks. A second extension of Metastats is a mixed-model zero-inflated Gaussian distribution that allows Metastats to account for a common characteristic of metagenomic data: the presence of many features with zero counts due to under sampling of the community. The number of 'missing' features (zero counts) correlates with the amount of sequencing performed thereby biasing abundance measurements and the differential abundance statistics derived from them.

Using simulated and real data we show that these methods significantly improve the accuracy of Metastats. We also describe the addition of several new statistical tests to our code (including presence/absence and corresponding odds-ratio, penetrance calculations, etc.) that improve the usability of our software in clinical practice.

References:
1. White JR, Nagarajan N, Pop M: Statistical Methods for Detecting Differentially Abundant Features in Clinical Metagenomic Samples. PLoS Comput Biol 5(4): e1000352, doi:10.1371/journal.pcbi.1000352

POSTER 23

Model-based preprocessing and base-calling for second-generation sequencing data

Chiao-wen Hsiao, and Hector Corrada Bravo

Second-generation sequencing, or massive parallel sequencing, has inspired new scientific approaches in basic, applied and clinical research. The high throughput of these technologies poses computational and statistical challenges to sequence data processing. Determining the quality of this data in an accurate and informative manner is instrumental to downstream analyses such as genotyping and de novo assembly. We present a model-based preprocessing and normalization procedure for data from the Illumina second-generation sequencing platform. Estimates derived from our model directly provide quality metrics of the bio-chemical sequencing processes, for example, read-specific incorporation and PCR amplification efficiency.

POSTER 24

Automated Identification and Characterization of Large Gene Families

Theodore R. Gibbons, Tsvetan R. Bachvaroff, Gregory T. Concepcion, Charles F. Delwiche

Rapid, accurate, automated identification of large gene families is a necessary step in the analysis of whole-transcriptome data from organisms containing very highly duplicated genes, such as members of the dinoflagellates and trypanosomatids. The reciprocal best hits algorithm, while often successful at identifying single-copy orthologs, can struggle to cluster members of large gene families when faced with many similarly high quality hits. We have developed an alternative approach that uses dynamic filtering thresholds to identify reciprocal good hits and cluster genes into families of arbitrarily large size. Our algorithm addresses these issues, and our cross-platform Python implementation aims to be more user friendly than alternative packages, while at the same time automatically generating additional files that facilitate intuitive interpretation of the results at the level of both whole-transcriptomes, and individual gene families.