Genome Biology 2011, **12(Suppl 1)**
http://genomebiology.com/supplements/12/S1

Genome Biology

MEETING ABSTRACTS

Beyond the Genome 2011

Washington DC, USA. 19–22 September 2011

Published: 19 September 2011

INVITED SPEAKER PRESENTATIONS

I1

Analysis of 2,440 human exomes highlights the evolution and functional impact of rare coding variation

Joshua Akey
Department of Genome Sciences, University of Washington, Seattle, WA 98195, USA
Genome Biology 2011, **12(Suppl 1):**I18

Deep exome resequencing is a powerful approach for delineating patterns of protein-coding variation among genes, pathways, individuals and populations. We analyzed exome data from 2,440 individuals of European and African ancestry as part of the National Heart, Lung, and Blood Institute's Exome Project, the aim of which is to discover novel genes and mechanisms that contribute to heart, lung and blood disorders. Each exome was sequenced to a mean coverage of 116×, allowing detailed inferences about the population genomic patterns of both common variation and rare coding variation. We identified more than 500,000 single nucleotide variations, the majority of which were novel and rare (76% of variants had a minor allele frequency of less than 0.1%), reflecting the recent dramatic increase in the size of the human population. The unprecedented magnitude of this dataset allowed us to rigorously characterize the large variation in nucleotide diversity among genes (ranging from 0 to 1.32%), as well as the role of positive and purifying selection in shaping patterns of protein-coding variation and the differential signatures of population structure from rare and common variation. This dataset provides a framework for personal genomics and is an important resource that will allow inferences of broad importance to human evolution and health.

I2

Abstract not submitted for online publication.

I3

Are clinical genomes already becoming semi-routine for patient care?

Mark Boguski
Harvard Medical School, Boston, MA, USA
Genome Biology 2011, **12(Suppl 1):**I16

Hardly a month goes by without a new published report of a patient's genome being used diagnostically for clinical management in a diverse spectrum of disease areas, including gastroenterology, nephrology, neurology and oncology. The impression is that clinical genomics is already becoming semi-routine. However, a large and complex set of non-technical barriers needs to be overcome before genomics can truly be integrated into the practice of medicine and made widely available for patient care. Through the use of case studies, my presentation will elucidate issues relating to the needs and requirements of the workforce, the legal and regulatory aspects of 'laboratory-developed tests' and insurance reimbursement for 'multi-analyte diagnostics'. The roles of the Food and Drug Administration, the Centers for Medicare & Medicaid Services and the College of American Pathologists will be highlighted.

I4

Interrogating the architecture of cancer genomes

Peter Campbell
Cancer Genome Project, Wellcome Trust Sanger Institute, Hinxton, UK
Genome Biology 2011, **12(Suppl 1):**I11

Cancer is driven by mutation. Using massively parallel sequencing technology, we can now sequence the entire genome of cancer samples, allowing the generation of comprehensive catalogs of somatic mutations of all classes. Bespoke algorithms have been developed to identify somatically acquired point mutations, copy number changes and genomic rearrangements, which require extensive validation by confirmatory testing. The findings from our first handful of genomes illustrate the potential for next-generation sequencing to provide unprecedented insight into mutational processes, cellular repair pathways and gene networks associated with cancer development. I will also review the possible applications of these technologies in a diagnostic and clinical setting and the potential routes for translation.

I5

Genome-forward oncology: how do we get there?

Matthew Ellis
Breast Cancer Program and the Genome Institute, Washington University in St. Louis, St. Louis, MO, USA
Genome Biology 2011, **12(Suppl 1):**I13

Massively parallel sequencing is transforming our knowledge of cancer, yet the medical value of next-generation approaches has not been fully established. From a technical perspective, it is easy to envisage that, within a few years, the primary diagnostic approach for all cancers will be to assess a partial or whole cancer genome sequence; however, the adoption of this approach will ultimately depend on the development of robust and valid models for the tailoring of therapy. Thus, within a short period, the focus of genomic investigation will shift from the current emphasis on discovery in poorly annotated datasets, such as The Cancer Genome Atlas, to ambitious investigations that focus on precise clinical questions. This transition will occur in two stages. The first stage will be a retrospective, 'genome-backward' approach, in which patients are treated blind to genomics but consent to prospective germline and tumor sequencing, as well as data sharing. In this way, models that use mutation patterns to predict treatment outcomes can be developed. In a later prospective, 'genome-forward' phase, therapeutic postulates that arise from genome sequencing will be used as the basis for clinical trial eligibility or stratification. Specific examples of how these approaches are being studied in breast cancer will be discussed.

I6

Abstract not submitted for online publication.

I7

Microbial reference genomes for human metagenomics

Sarah K Highlander
Department of Molecular Virology and Microbiology, Baylor College of Medicine, Houston, TX 77030, USA
Genome Biology 2011, **12(Suppl 1):**I28

A key resource for metagenomics projects is a representative catalog of annotated microbial genome sequences to serve as a reference for classification and functional annotation [1]. The Human Microbiome Project (HMP) [2], which is funded by the National Institutes of Health, began with the goal of sequencing 600 genomes from culturable prokaryotes that are representative of those inhabiting the major niches within and on the human

BioMed Central © 2011 BioMed Central Ltd

body. This effort has been expanded to include uncultured organisms, small eukaryotes, and viruses that infect prokaryotes or eukaryotes, and the current sequencing target is 3,000 microbes. An automated pipeline for prokaryotic gene calling and annotation has been established for the project, yet very few of the HMP organisms have been subjected to in-depth analysis. Comparisons of several HMP genomes have revealed substantial intraspecies differences and provide clues about the pathogenic and physiological potential of these organisms.

References

1. Human Microbiome Jumpstart Reference Strains Consortium, Nelson KE, Weinstock GM, Highlander SK, Worley KC, Creasy HH, Wortman JR, Rusch DB, Mitreva M, Sodergren E, Chinwalla AT, Feldgarden M, Gevers D, Haas BJ, Madupu R, Ward DV, Birren BW, Gibbs RA, Methe B, Petrosino JF, Strausberg RL, *et al.*: **A catalog of reference genomes from the human microbiome.** *Science* 2010, **328:**994-999.
2. **Data Analysis and Coordination Center for the Human Microbiome Project** [http://www.hmpdacc.org]

I8

Abstract not submitted for online publication.

I9

Next-generation clinical sequencing in a children's hospital

Stephen Kingsmore
Children's Mercy Hospital, Kansas City, MO, USA
Genome Biology 2011, **12(Suppl 1):**I21

Next-generation sequencing and analysis tools are reaching the mature stage at which mentioning their usefulness for clinical testing is not an oxymoron. Indeed, next-generation clinical sequencing has the potential to transform children's health care, because inherited illnesses account for much of the childhood disease burden. I will discuss the first year of the integration of genomic medicine for Mendelian diseases at Children's Mercy Hospital.

I10

A glimpse at tumor genome evolution

Elaine R Mardis
The Genome Institute at Washington University School of Medicine, St. Louis, MO 63108, USA
Genome Biology 2011, **12(Suppl 1):**I8

Next-generation DNA sequencing has dramatically affected cancer genomics efforts in several important ways. Although whole genome sequencing remains an analytical challenge, such efforts are yielding data that elucidate the myriad ways in which a genome can be influenced by single point mutations, focused insertions or deletions, and large structural alterations. In addition to cataloguing somatic alterations, various correlation analyses are indicating the genes whose alterations most profoundly determine patient outcomes, patient responses to therapeutics and other important aspects of disease biology. We have recently begun exploring how the digital nature of next-generation sequencing reads allows important information about tumor cell genomic heterogeneity to be inferred, revealing the earliest mutations and how the composition of the tumor cell mass changes over time under the influence of stressors such as chemotherapy.

I11

Abstract not submitted for online publication.

I12

Comparative genomic analysis of mouse and human mammary tumors

Charles M Perou
Lineberger Comprehensive Cancer Center, 450 West Drive, CB# 7295, The University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA
Genome Biology 2011, **12(Suppl 1):**I10

Human breast cancer is a heterogeneous disease consisting of at least five molecular subtypes. With our increasing understanding of the genetic underpinnings of human breast cancer subtypes, many genetically engineered mouse models (GEMMs) have been created to mimic these subtypes. Traditionally, these GEMMs have been analyzed individually; however, when consolidated into a single dataset, these models have an increased sensitivity for detecting significant overlap with human subtypes. These associations between GEMMs and human subtypes can provide insight into the genetic alterations associated with the human subtypes and can provide a translational resource for preclinical drug testing. We previously performed an analysis of 13 GEMMs [1] and have since expanded our dataset to include 29 murine models. Using DNA expression microarrays and unsupervised hierarchical clustering, we identified 17 distinct murine classes from this set of 29 models. After comparison with the human subtypes by using gene set analysis, we found that seven of our classes show statistically significant overlap with five human subtypes. Although we observed no statistical overlap with the human luminal B subtype, the MMTV-NeuPyMT class shows significant overlap with the combined luminal A and B subtypes. Some of the new models fall into classes that have been defined previously, but many, such as the AIB1 and ETV6 murine models, are associated with new groups. The AIB1 and ETV6 models fall into their own classes, and each shows statistically significant overlap with HER2-enriched tumors, a subtype that was not previously observed to overlap with any GEMM. These expanded analyses have identified new and important models and are laying the groundwork for additional studies focused on DNA copy number changes and mutation status similarities between mice and humans.

References

1. Herschkowitz JI, Simin K, Weigman VJ, Mikaelian I, Usary J, Hu Z, Rasmussen KE, Jones LP, Assefnia S, Chandrasekharan S, *et al.*: **Identification of conserved gene expression features between murine mammary carcinoma models and human breast tumors.** *Genome Biol* 2007, **8:**R76.

I13

Reconstructing microbial communities

Mihai Pop
University of Maryland, College Park, MD, USA
Genome Biology 2011, **12(Suppl 1):**I24

Metagenomic studies have allowed an unprecedented view of the microbial communities that inhabit our world and our bodies. Deep sequencing data have already been generated from several environments, as well as from various human body sites. As more data are generated, we are beginning to understand the structure of our commensal microbial communities and how microbes affect our health. Analyzing the metagenomic data, however, poses significant computational challenges, because few software tools are available that can handle the volume and characteristics of the data being generated. In my talk, I will primarily focus on the challenges posed by metagenomic assembly and will outline recent research in my laboratory aimed at meeting these challenges. I will also describe some of the analyses that can be performed on the assembled data but would not be possible in read-based analyses.

I14

Next-generation human genetics

Jay Shendure
Department of Genome Sciences, University of Washington, Seattle, WA, USA
Genome Biology 2011, **12(Suppl 1):**I17

Over the past five years, a new generation of technologies has reduced the cost of DNA sequencing by more than four orders of magnitude, democratizing the field by putting the sequencing capacity of a major genome center in the hands of individual investigators [1]. To exploit this paradigm shift, we have developed new technical methods and analytical strategies for disease gene discovery based on whole exome and whole genome sequencing. Our results to date include proof of concept [2] and the first demonstration [3] that exome sequencing of a small number of individuals can be applied to solve Mendelian, single-gene, disorders such as Miller syndrome [3] and Kabuki syndrome [4]. Recently, we have also demonstrated that exome or genome sequencing of parent-child trios can be used to rapidly identify candidate genes for complex disorders such as autism [5]. We are currently extending these strategies to additional simple and complex diseases of unknown etiology.

References

1. Shendure J, Ji H: **Next-generation DNA sequencing.** *Nat Biotechnol* 2008, **26:**1135-1145.
2. Ng SB, Turner EH, Robertson PD, Flygare SD, Bigham AW, Lee C, Shaffer T, Wong M, Bhattacharjee A, Eichler EE, Bamshad M, Nickerson DA, Shendure J: **Targeted capture and massively parallel sequencing of 12 human exomes.** *Nature* 2009, **461:**272-276.

3. Ng SB, Buckingham KJ, Lee C, Bigham AW, Tabor HK, Dent KM, Huff CD, Shannon PT, Jabs EW, Nickerson DA, Shendure J, Bamshad MJ: **Exome sequencing identifies the cause of a mendelian disorder.** *Nat Genet* 2010, **42:**30-35.
4. Ng SB, Bigham AW, Buckingham KJ, Hannibal MC, McMillin MJ, Gildersleeve HI, Beck AE, Tabor HK, Cooper GM, Mefford HC, Lee C, Turner EH, Smith JD, Rieder MJ, Yoshiura K, Matsumoto N, Ohta T, Niikawa N, Nickerson DA, Bamshad MJ, Shendure J: **Exome sequencing identifies *MLL2* mutations as a cause of Kabuki syndrome.** *Nat Genet* 2010, **42:**790-793.
5. O'Roak BJ, Deriziotis P, Lee C, Vives L, Schwartz JJ, Girirajan S, Karakoc E, Mackenzie AP, Ng SB, Baker C, Rieder MJ, Nickerson DA, Bernier R, Fisher SE, Shendure J, Eichler EE: **Exome sequencing in sporadic autism spectrum disorders identifies severe *de novo* mutations.** *Nat Genet* 2011, **43:**585-589.

I15

Metatranscriptomics of the human gut microbiome

Thomas Sichertz Pontén
Center for Biological Sequence Analysis, Department of Systems Biology, Technical University of Denmark, DK-2800 Kongens Lyngby, Denmark
Genome Biology 2011, **12(Suppl 1):**I31

Our 'other' genome is the collective genetic information in all of the microorganisms that are living on and within us. Collectively known as the microbiome, these microbial cells outnumber human cells in the body by more than 10 to 1, and the genes carried by these organisms outnumber the genes in the human genome by more than 100 to 1.
How these organisms contribute to and affect human health is poorly understood, but the emerging field of metagenomics promises a more comprehensive and complete understanding of the human microbiome.
In the European-funded Metagenomics of the Human Intestinal Tract (MetaHIT) project [1], we combined next-generation sequencing with high-density microarrays, generating metagenomic and metatranscriptomic data for more than 400 individuals.
The combined data reveal clusters of coexisting species with differences in pathway and gene function activity, suggesting that there is a division of labor between the bacterial species in the human gut microbiome.

Reference

1. The Metagenomics of the Human Intestinal Tract Project [http://www.metahit.eu]

I16

De novo mutations in mental retardation

Joris A Veltman
Department of Human Genetics, Radboud University Nijmegen Medical Centre, PO Box 9101, 6500 HB Nijmegen, The Netherlands
Genome Biology 2011, **12(Suppl 1):**I20

Recent studies have indicated that humans have an exceptionally high per-generation mutation rate of 7.6×10^{-9} to 2.2×10^{-8}. These spontaneous germline mutations can have serious phenotypic consequences when affecting functionally relevant bases in the genome. In fact, their occurrence may explain why cognitive disorders with a severely reduced fecundity, such as mental retardation, remain frequent in the human population, especially when the mutational target is large and comprises many genes. This would explain a major paradox in the evolutionary genetic theory of these disorders. In this presentation, I will describe our recent work on using a family-based exome sequencing approach to test this *de novo* mutation hypothesis in ten patients with unexplained mental retardation [1]. Unique nonsynonymous *de novo* mutations were identified and validated in nine genes. Six of these, identified in different patients, were likely to be pathogenic based on gene function, evolutionary conservation and mutation impact. The clinical relevance of these novel genes, and the ultimate proof that they cause disease, lies in the identification of *de novo* mutations in additional patients with a similar phenotype. As such, we are currently screening approximately 1,200 patients with unexplained mental retardation for mutations in *YY1*, which is one of these newly identified genes. In addition, we are extending our family-based exome sequencing approach to 100 patients to establish the diagnostic yield for *de novo* mutations in patients with unexplained mental retardation. These findings, when replicated, provided strong experimental support for a *de novo* paradigm for mental retardation. Together with *de novo* copy number variation, *de novo* point mutations of large effect could explain the majority of all mental retardation cases in the population. In my presentation, I will explain this work, as well as related work on autism [2] and schizophrenia [3].

References

1. Vissers LELM, de Ligt J, Gilissen C, Janssen I, Steehouwer M, de Vries P, van Lier B, Arts P, Wieskamp N, del Rosario M, van Bon BWM, Hoischen A, de Vries BBA, Brunner HG, Veltman JA: **A *de novo* paradigm for mental retardation.** *Nat Genet* 2010, **42:**1109-1112.
2. O'Roak BJ, Deriziotis P, Lee C, Vives L, Schwartz JJ, Girirajan S, Karakoc E, Mackenzie AP, Ng SB, Baker C, Rieder MJ, Nickerson DA, Bernier R, Fisher SE, Shendure J, Eichler EE: **Exome sequencing in sporadic autism spectrum disorders identifies severe *de novo* mutations.** *Nat Genet* 2011, **43:**585-589.
3. Girard SL, Gauthier J, Noreau A, Xiong L, Zhou S, Jouan L, Dionne-Laporte A, Spiegelman D, Henrion E, Diallo O, Thibodeau P, Bachand I, Bao JY, Tong AH, Lin CH, Millet B, Jaafari N, Joober R, Dion PA, Lok S, Krebs MO, Rouleau GA: **Increased exonic *de novo* mutation rate in individuals with schizophrenia.** *Nat Genet* 2011. doi: 10.1038/ng.886.

I17

Comparative analysis of the vaginal microbiome in health and disease

Bryan A White[1,2,3], Andres M Gomez[1,2], Mengfei Ho[4], Margret Berg Miller[1], Susan M Thomas[1], Carl J Yeoman[1], Suleyman Yildirim[1], Douglas J Creedon[5], Tony L Goldberg[6], Steven R Leigh[1,7], Karen E Nelson[8], Rebecca M Stumpf[1,7] and Brenda A Wilson[1,4]
[1]The Institute for Genomic Biology, University of Illinois, Urbana, IL 61801, USA, [2]Department of Animal Sciences, University of Illinois, Urbana, IL 61801, USA, [3]Division of Biomedical Sciences, University of Illinois, Urbana, IL 61801, USA, [4]Department of Microbiology, University of Illinois, Urbana, IL 61801, USA, [5]Department of Obstetrics and Gynecology, Mayo Clinic, Rochester, MN 55905, USA, [6]Department of Pathobiological Sciences, University of Wisconsin, Madison, WI 53706, USA, [7]Department of Anthropology, University of Illinois, Urbana, IL 61801, USA, [8]J. Craig Venter Institute, Rockville, MD 20850, USA
Genome Biology 2011, **12(Suppl 1):**I26

Diseases of the vaginal tract result from perturbations of the complex interactions among microbes of the host vaginal ecosystem. Recent advances in our understanding of these complex interactions have been enabled by next-generation-sequencing-based approaches, which make it possible to study the vaginal microbiome. In harnessing these approaches, we are beginning to define what constitutes an imbalance of the vaginal microbiome and how such imbalances, along with associated host factors, lead to infection and disease states such as bacterial vaginosis (BV), preterm births, and susceptibility to HIV and other sexually acquired infections. We have exploited various approaches to this end: comparative analysis of reference microbial genomes of vaginal isolates; comparative microbiome, metabolome and metagenome analysis of vaginal communities from subjects deemed to be healthy and individuals with BV; and comparative microbiome analysis of vaginal communities from humans and non-human primate species. The results from comparative genome sequencing have led us to suggest that different strains of the proposed pathogen *Gardnerella vaginalis* have different virulence potentials and that the detection of *G. vaginalis* in the vaginal tract is not indicative of a disease state [1]. Comparative microbiome, metabolome and metagenome analysis of vaginal communities from humans has demonstrated that the microbial communities from subjects with BV have a defined bacterial composition and metabolic profile that is distinct from subjects who do not have BV [2 and unpublished observations]. Our studies of microbial communities from non-human primate species and humans provide a unique comparative context. From an evolutionary perspective, humans and non-human primates differ considerably in mating habits, estrus cycles and gestation period. Moreover, birth is difficult in humans relative to other primates, increasing the risks of maternal injury and infection. In light of these numerous differences between humans and non-human primates, we hypothesize that humans have microbial populations that are distinct from those of non-human primates. Preliminary results show that the vaginal microbiomes of non-human primates are more diverse and are compositionally distinct from human vaginal microbiomes [3,4]. The composition of bacterial genera found in non-human primates is dissimilar to that seen in humans, most notably with lactobacilli being much less abundant in non-human primates. Our observations point to vaginal microbial communities being an important component of an evolutionary

set of adaptations that separates humans from other primates and is of fundamental importance to health and reproductive function.

References

1. Yeoman CJ, Yildirim S, Thomas SM, Durkin AS, Torralba M, Sutton G, Buhay CJ, Ding Y, Dugan-Rocha SP, Muzny DM, Qin X, Gibbs RA, Leigh SR, Stumpf R, White BA, Highlander SK, Nelson KE, Wilson BA: **Comparative genomics of *Gardnerella vaginalis* strains reveals substantial differences in metabolic and virulence potential.** *PLoS ONE* 2010, **5:**e12411.
2. Kim TK, Thomas SM, Ho M, Sharma S, Reich CI, Frank JA, Yeater KM, Biggs D, Nakamura N, Stumpf R, Leigh SR, Tapping RI, Blanke SR, Slauch JM, Gaskins HR, Weisbaum JS, Olsen GJ, Hoyer LL, Wilson BA: **Heterogeneity of vaginal microbial communities within individuals.** *J Clin Microbiol* 2009, **47:**1181-1189.
3. Rivera AJ, Frank JA, Stumpf R, Salyers AA, Wilson BA, Olsen GJ, Leigh S: **Differences between the normal vaginal bacterial community of baboons and that of humans.** *Am J Primatol* 2011, **73:**119-126.
4. Rivera AJ, Stumpf RM, Wilson B, Leigh S, Salyers AA: **Baboon vaginal microbiota: an overlooked aspect of primate physiology.** *Am J Primatol* 2010, **72:**467-474.

I18

The Joint Center for Structural Genomics: exploration of the human gut microbiome

Ian Wilson
The Scripps Research Institute, 10550 North Torrey Pines Road, La Jolla, CA 92037, USA
Genome Biology 2011, **12(Suppl 1):**I25

For more than a decade, the Joint Center for Structural Genomics (JCSG) [1] has been at the forefront of developing tools and methodologies that allow the application of high-throughput structural biology to a broad range of biological and biomedical investigations. In the previous phases of the National Institutes of Health's Protein Structure Initiative (PSI; 2000 to 2010) [2], we explored structural coverage of uncharted regions of the protein universe [3], as well as a single organism, allowing complete structural reconstruction of the metabolic network of *Thermotoga maritima* [4]. In the current phase (PSI:Biology; 2010 to 2015), the JCSG is leveraging its high-throughput platform to explore the structural basis for host-microbe interactions in the human microbiome. The emerging field of metagenomics has been particularly enlightening: the human gut microbiome sequencing projects have already uncovered fascinating new families and expansions of known families for adaptation to this environment. The gut microbiota is dominated by poorly characterized bacterial phyla, which contain an unusually high number of uncharacterized proteins that are largely unstudied. Their influence upon human development, physiology, immunity and nutrition is only starting to surface and is thus an exciting new frontier for structural genomics, where we can structurally investigate the contributions of these microorganisms to human health and disease. The JCSG is located at The Scripps Research Institute, the Genomics Institute of the Novartis Research Foundation, University of California at San Diego, the Sanford-Burnham Medical Research Institute and SSRL/Stanford University.

References

1. The Joint Center for Structural Genomics [http://www.jcsg.org]
2. The Protein Structure Initiative [http://www.nigms.nih.gov/initiatives/psi]
3. Jaroszewski L, Li Z, Krishna SS, Bakolitsa C, Wooley J, Deacon AM, Wilson IA, Godzik A: **Exploration of uncharted regions of the protein universe.** *PLoS Biol* 2009, **7:**e1000205.
4. Zhang Y, Thiele I, Weekes D, Li Z, Jaroszewski L, Ginalski K, Deacon AM, Wooley J, Lesley SA, Wilson IA, Palsson B, Osterman A, Godzik A: **Three-dimensional structural view of the central metabolic network of *Thermotoga maritima*.** *Science* 2009, **325:**1544-1549.

POSTER PRESENTATIONS

P1

RNA-Seq methods for imperfect samples: development, evaluation and applications

Xian Adiconis[1], Lin Fan[1], David DeLuca[1], Andrey Sivachenko[1], Nathalie Pochet[1], Aaron Berlin[1], Sarah Young[1], Gad Getz[1], Aviv Regev[1], Chad Nusbaum[1], Andreas Gnirke[1] and Joshua Z. Levin[1]
[1]Broad Institute of MIT and Harvard, Cambridge, MA 02141, USA
Genome Biology 2011, **12(Suppl 1):**P1

Next-generation sequencing of RNA (RNA-Seq) is a powerful tool that can be applied to a wide range of biological questions. RNA-Seq provides insight at multiple levels into the transcription of the genome. It yields sequence, splicing and expression-level information, allowing the identification of novel transcripts and sequence alterations. We have been developing and comparing methods for samples that present a challenge: that is, those with low quantity and/or quality RNA.

RNA-Seq methods that start from total RNA and do not require the oligo(dT) purification of mRNA will be valuable for such challenging samples. Such methods use alternative approaches to reduce the fraction of sequencing reads derived from rRNA. We will present results from multiple approaches, including the use of not-so-random (NSR) primers for cDNA synthesis, low-C_0t hybridization with a duplex-specific nuclease for light normalization and NuGEN's Ovation RNA-Seq kit. We demonstrated that these three methods successfully reduce the fraction of rRNA to less than 13%, even when starting from degraded RNA. We compared the performance between these methods and with 'gold standard' RNA-Seq data (derived from samples with large quantities of high-quality RNA), using quantitative criteria that evaluate effectiveness for genome annotation, transcript discovery and expression profiling. The application of these methods to samples that contain degraded RNA and/or very low input amounts of RNA will also be presented.

P2

Viral diversity in children with diarrhea in Gambia

Irina Astrovskaya[1], Bo Liu[1] and Mihai Pop[1]
[1]University of Maryland Institute for Advanced Computer Studies, University of Maryland, College Park, MD 20910, USA
Genome Biology 2011, **12(Suppl 1):**P2

Background Despite a decrease in the rate of mortality due to diarrhea in the past few decades, diarrhea remains one of the leading causes of childhood deaths worldwide, especially in developing countries. The known causes of disease include infection with bacteria (for example, *Salmonella* or *Shigella*), viruses (for example, rotaviruses, noroviruses or hepatitis viruses) or parasites (for example, *Giardia lamblia* or *Cryptosporidium*); however, the true agent remains unknown in up to 40% of clinical cases [1].

Recent advances in sequencing technologies allow us to explore microbial diversity in a sample, making metagenomic analysis a promising technique to characterize the viral spectrum (that is, the viral sequences and their abundances) in stool samples. By studying the genomes of particular viruses that are present *in vivo*, we may obtain a complete picture of the causes of diarrhea and potentially identify unknown viral pathogens.

Methods In this project, we explored viral communities present in diarrheal samples from 40 Gambian children of 18 months of age or younger. Each sample contained 4,829 to 57,778,454 pyrosequencing shotgun reads with read lengths varying from 50 to 930 bp.

In our pipeline, we first assembled the genomes of known diarrhea-causing viruses by aligning the reads with the available references in the National Center for Biotechnology Information database and reconstructing the haplotypes from the mapped reads. Additional care needs to be taken for RNA viruses because they exist as a set of closely related but nonidentical genomes (quasispecies). We therefore reconstructed the set of the most plausible haplotypes [2] rather than the consensus genome. Next, we estimated the abundances of the assemblies by employing an expectation-maximization algorithm that takes into account sequencing error, as well as mark reads that are not adequately covered by the assemblies. Then, we focused on assembling the uncovered reads and identifying them. Finally, we analyzed the viral spectrum across all of the samples to decide whether specific genomes are responsible for causing diarrhea.

Results We were able to detect and assemble sequences from known diarrhea-causing viruses (such as rotaviruses, adenoviruses and noroviruses), known human viruses (such as herpesviruses and enteroviruses) and potential diarrhea-causing viruses (such as bocaviruses, astroviruses and parechoviruses). These findings were consistent with independent virology results.
In some clinical cases, sequences from classic viruses were found, but the virology results were negative.
Conclusions Annually, diarrhea causes about 1.8 million deaths worldwide. Although many causative agents are known, as many as 40% of clinical cases are attributed to unknown viral pathogens. The metagenomic analysis of pyrosequencing data allows us to investigate the role of viruses in causing diarrhea.
References
1. Finkbeiner SR, Allred AF, Tarr PI, Klein EJ, Kirkwood CD, Wang D: **Metagenomic analysis of human diarrhea: viral detection and discovery.** *PLoS Pathog* 2008, **4:**e1000011.
2. Astrovskaya I, Tork B, Mangul S, Westbrooks K, Mandoiu II, Balfe P, Zelikovsky A: **Inferring viral quasispecies spectra from 454 pyrosequencing reads.** *BMC Bioinformatics* 2011, **Suppl 6:**S1.

P3

COSMIC: the Catalogue of Somatic Mutations in Cancer

Nidhi Bindal[1], Simon A Forbes[1], David Beare[1], Prasad Gunasekaran[1], Kenric Leung[1], Chai Yin Kok[1], Mingming Jia[1], Sally Bamford[1], Charlotte Cole[1], Sari Ward[1], Jon Teague[1], Michael R. Stratton[1], Peter Campbell[1] and P Andrew Futreal[1]
[1]Wellcome Trust Sanger Institute, Cambridge, UK
Genome Biology 2011, **12(Suppl 1):**P3

The Catalogue Of Somatic Mutations In Cancer (COSMIC) [1] is one of the largest repositories of information on somatic mutations in human cancer. The project has been running for more than ten years as part of the Cancer Genome Project (CGP) at the Wellcome Trust Sanger Institute in the UK.
The data in COSMIC are curated from a variety of sources, primarily the scientific literature and large international consortia. The project includes information from the CGP, along with data from other consortia such as the International Cancer Genome Consortium and The Cancer Genome Atlas. In addition, COSMIC is regularly updated with the genes highlighted in the Cancer Gene Census, which curates the scientific literature for known cancer genes [2].
With the advent of whole exome and genome sequencing technology, the amount of data in COSMIC is increasing rapidly. The recent COSMIC release (version 53; 18 May 2011) contains 608,042 tumor and cell line samples, annotating 176,856 mutations across 19,439 genes, with 352 full exomes, 43 whole genome rearrangement screens and 4 full genomes now available. The data are updated regularly, with new releases scheduled every two months.
COSMIC provides a large number of graphical and tabular views for interpreting and mining the large quantity of information, as well as the facility to export the relevant data in various formats. The website can be navigated in many ways to examine mutation patterns on the basis of genes, samples and phenotypes, which are the main entry points to COSMIC.
COSMIC also provides various options to browse the data in a genomic context. Integration with the Ensembl genome browser allows the visualization of full genome annotations, together with COSMIC data, on the GRCh37 genome coordinates. COSMIC also contains its own genome browser, which facilitates data analysis by combining genome-wide gene structures and sequences with rearrangement breakpoints, copy number variations and all somatic substitutions, deletions, insertions and complex gene mutations.
The main COSMIC website [1] encompasses all of the available data. However, within COSMIC, the Cancer Cell Line Project [3] is a specialized component, which provides details of the genotyping of almost 800 commonly used cancer cell lines, through the set of known cancer genes. Its focus is to identify driver mutations, or those likely to be implicated in the oncogenesis of each tumor.
This information forms the basis for integrating COSMIC with the Genomics of Drug Sensitivity in Cancer project [4], which is a joint effort with the Massachusetts General Hospital [5] to screen this panel of cancer cell lines against potential anticancer therapeutic compounds to investigate correlations between somatic mutations and drug sensitivity.
Data on somatic mutations in cancer are being produced at a rapidly increasing rate, and the combined analysis of large distributed datasets is becoming ever more difficult. However, COSMIC curates and standardizes this information in a single database, providing user-friendly browsing tools and analytical functions, thus ensuring its role as a key resource in human cancer genetics.
References
1. Catalogue Of Somatic Mutations In Cancer [http://www.sanger.ac.uk/cosmic]
2. Cancer Gene Census [http://www.sanger.ac.uk/genetics/CGP/Census/]
3. Cancer Cell Line Project [http://www.sanger.ac.uk/genetics/CGP/CellLines/]
4. Genomics of Drug Sensitivity in Cancer Project [http://www.sanger.ac.uk/genetics/CGP/translation/]
5. Massachusetts General Hospital Cancer Center [http://www.massgeneral.org/cancer/]

P4

A novel functional variant in 8q24 is associated with regulation of prostate stem cell antigen (*PSCA*) gene expression and bladder cancer risk

Yi-Ping Fu[1], Indu Kohaar[1], Adam Mumy[1], Wei Tang[1], Brian Muchmore[1], Patricia Porter-Gill[1], Luyang Liu[1], Jonine Figueroa[2] Montserrat Garcia-Closas[2], Dalsu Baris[2], Mark Purdue[2], Michael Thun[3], Demetrius Albanes[2], Nuria Malats[4], Francisco X Real[4], Manolis Kogevinas[5], Alison Johnson[6], Molly Schwenn[7], Stephen Chanock[1], Nathaniel Rothman[2], Debra Silverman[2] and Ludmila Prokunina-Olsson[1]
[1]Laboratory of Translational Genomics, Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Bethesda, MD 20892, USA; [2]Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Bethesda, MD 20892, USA; [3]Epidemiology Research Program, American Cancer Society, Atlanta, GA 30303, USA; [4]Spanish National Cancer Research Center, Madrid 28029, Spain; [5]Centre for Research in Environmental Epidemiology, Barcelona 08003, Spain; [6]Vermont Cancer Registry, Burlington, VT 05401, USA; [7]Maine Cancer Registry, Augusta, ME 04333, USA
Genome Biology 2011, **12(Suppl 1):**P4

Background Recent genome-wide association studies (GWAS) have identified allele T of a single nucleotide polymorphism (SNP), rs2294008, in the prostate stem cell antigen (*PSCA*) gene as a risk factor for bladder cancer [1,2]. In the present study, we aimed to find additional disease-associated SNPs in the *PSCA* region and to explore their possible molecular function.
Methods Based on information from the 1000 Genomes and HapMap 3 projects, we performed imputation analysis on 3,532 bladder cancer cases and 5,120 healthy controls of European ancestry from the stage 1 bladder cancer GWAS, within ±100 kb of the region flanking the GWAS signal, rs2294008. The average allele dosage and best-guess genotypes were estimated and tested for association between SNP variants and bladder cancer risk by using unconditional logistic regression. Functional follow-up studies included RNA sequencing in normal and tumor bladder samples and electrophoretic mobility shift assays to examine the potentially altered DNA-protein interactions for SNPs of interest.
Results A total of 639 imputed and 37 genotyped SNPs within ±100 kb of the region of the original GWAS signal were tested for genetic association with bladder cancer. In these stage 1 GWAS samples, the SNP rs2294008 had a per-allele odds ratio (OR) of 1.09 (95% confidence interval (CI) = 1.02 to 1.16, $P = 6.93\ 10^{-4}$). Multivariable logistic regression analysis adjusted for the study center, age, gender, smoking status and rs2294008 genotype revealed a novel associated variant, rs2978974 (OR = 1.11, 95% CI = 1.04 to 1.19, $P = 1.62 \times 10^{-3}$). There was low linkage disequilibrium between rs2978974 and the original GWAS signal, rs2294008 (D' = 0.19, $r^2 = 0.02$). Only individuals carrying the risk variant of both SNPs had an increased risk of bladder cancer (OR = 1.24, 95% CI = 1.13 to 1.35, $P = 4.69 \times 10^{-6}$) and not individuals who carried a risk variant of only one of the SNPs ($P > 0.05$). Stratified analysis suggested that this compound effect of rs2294008 and rs2978974 was more significant in males (OR = 1.27, $P = 2.80 \times 10^{-6}$) than in females (OR = 1.08, $P = 0.52$).
rs2978974 resides 10 kb upstream of rs2294008, is marked by an H3K4me3 signal and is in the vicinity of an androgen-receptor-binding site. Using RNA sequencing of bladder samples, we showed that rs2978974 is located within an alternative, untranslated first exon of *PSCA*. Using the electrophoretic mobility shift assay with nuclear proteins from LNCaP and HeLa cells, we

observed that the non-risk-associated allele (G) of rs2978974, but not the risk allele (A), could bind to ELK1, a protein belonging to the ETS family of transcription factors.

Conclusions We identified a SNP, rs2978974, in the *PSCA* region as a novel marker for bladder cancer susceptibility. There was a compound effect in carriers of both the rs2294008 and rs2978974 risk variants. The functional relevance of rs2978974 might be related to the loss of ELK1 regulation by the risk allele (A) and differential regulation of *PSCA* mRNA expression.

References

1. Rothman N, Garcia-Closas M, Chatterjee N, Malats N, Wu X, Figueroa JD, Real FX, Van Den Berg D, Matullo G, Baris D, Thun M, Kiemeney LA, Vineis P, De Vivo I, Albanes D, Purdue MP, Rafnar T, Hildebrandt MA, Kiltie AE, Cussenot O, Golka K, Kumar R, Taylor JA, Mayordomo JI, Jacobs KB, Kogevinas M, Hutchinson A, Wang Z, Fu YP, Prokunina-Olsson L *et al.*: **A multi-stage genome-wide association study of bladder cancer identifies multiple susceptibility loci.** *Nat Genet* 2010, **42:**978-984.
2. Wu X, Ye Y, Kiemeney LA, Sulem P, Rafnar T, Matullo G, Seminara D, Yoshida T, Saeki N, Andrew AS, Dinney CP, Czerniak B, Zhang ZF, Kiltie AE, Bishop DT, Vineis P, Porru S, Buntinx F, Kellen E, Zeegers MP, Kumar R, Rudnai P, Gurzau E, Koppova K, Mayordomo JI, Sanchez M, Saez B, Lindblom A, de Verdier P, Steineck G *et al.*: **Genetic variation in the prostate stem cell antigen gene *PSCA* confers susceptibility to urinary bladder cancer.** *Nat Genet* 2009, **41:**991-995.

P5

Evaluating short-read sequence data from the highly redundant, novel transcriptome of *Polarella glacialis*

Theodore R Gibbons[1], Gregory T Concepcion[1], Tsvetan R Bachvaroff[2] and Charles F Delwiche[1]
[1]Department of Cell Biology and Molecular Genetics, University of Maryland, College Park, MD 20742, USA; [2]Smithsonian Environmental Research Center, Edgewater, MD 21037, USA.
Genome Biology 2011, **12(Suppl 1):**P5

Background Dinoflagellates are a diverse group of ecologically important eukaryotic algae, the global impact of which ranges from the large-scale primary production of oxygen [1] to devastating toxic algal blooms [2]. These organisms have exceptionally large genomes (10^9 to 10^{11} bases) [3] and highly duplicated genes (which can occur thousands of times within a single genome) [4]. These and other unusual characteristics have made dinoflagellates difficult to study using traditional molecular biology techniques. Sequence data for dinoflagellates are correspondingly sparse, and not a single genome sequence has been published to date.

As part of our project called Assembling the Dinoflagellate Tree of Life (DAToL), our laboratory has sequenced the transcriptome of *Polarella glacialis*. Its genome is estimated to be only 3 Gb in size, making it one of the smallest known dinoflagellate genomes. Because we had to rely on *de novo* assemblers that had been tested using data from organisms that are extremely divergent from dinoflagellates, we took special care in our attempts to validate the data. Before expanding our analyses to include additional dinoflagellates, we compared the results from different sequencing and assembly methods.

Methods Total RNA was extracted from cultured *P. glacialis*. This sample was then divided and shipped to Macrogen for rRNA degradation, library preparation and sequencing. One library was sequenced on one-eighth of a Roche/454 GS FLX picotiter plate using Titanium chemistry. A second library was sequenced using one lane on an Illumina GAIIx sequencer for 78 cycles in both directions (paired end). The sequences were assembled using Newbler, MIRA, Oases and Trinity, and they were analyzed using various custom scripts.

Results The total amount of unassembled 454 sequence data added to less than one-third of the combined lengths of only those Trinity transcripts that had a significant BLAST hit against a sequence in GenBank, indicating that we did not achieve complete coverage with our 454 data.

Conclusions Our primary hypothesis was that the longer read lengths of the 454 data might allow the corresponding assemblers to better resolve repetitive sequences, which could be instrumental for assembling conserved regions within highly duplicated genes. Our failure to obtain complete coverage with the 454 dataset undermined our ability to test this hypothesis, although we made several other interesting observations. Notably, despite the vast disparity in the depth of the coverage between the 454 and Illumina assemblies, we observed unique, apparently real sequences within some of the 454 contigs.

References

1. Yang EJ, Choi JK, Hyun JH: **Distribution and structure of heterotrophic protist communities in the northeast equatorial Pacific Ocean.** *Mar Biol* 2004, **146:**1-15.
2. Wang DZ: **Neurotoxins from marine dinoflagellates: a brief review.** *Mar Drugs* 2008, **6:**349e731.
3. Hou Y, Lin S: **Distinct gene number-genome size relationships for eukaryotes and non-eukaryotes: gene content estimation for dinoflagellate genomes.** *PLoS ONE* 2009, **4:**e6978.
4. Bachvaroff TR, Place AR: **From stop to start: tandem gene arrangement, copy number and *trans*-splicing sites in the dinoflagellate *Amphidinium carterae*.** *PLoS ONE* 2008, **3:**e2929.

P6

A cost-effective and universal strategy for complete prokaryotic genome sequencing proposed by computer simulation

Jingwei Jiang[1], Jun Li[1] and Frederick C Leung[1]
[1]School of Biological Sciences, Faculty of Science, The University of Hong Kong, China
Genome Biology 2011, **12(Suppl 1):**P6

Background Pyrosequencing techniques allow scientists to perform prokaryotic genome sequencing and achieve draft sequences within a

Table 1 (abstract P6). Main average indices for different sequencing strategies for 100 genomes (400-bp read length)

ST	GCE (%)	SBE (%)	IDR (%)	FLT (%)	FDT (%)	CN	NB	SN
6xSE+10xPE	98.26971	0.004915	0.000364	0.310807	0.4678237	50.94	331136.7	3.64
10xSE+10xPE	98.30248	0.004265	0.000322	0.2626039	0.5629617	44.75	383793.6	3.51
15xSE+10xPE	98.32861	0.003293	0.000294	0.2518801	0.6041274	43.12	397060.7	3.48
20xSE+10xPE	98.35117	0.00227	0.000293	0.2307405	0.6301239	42.3	411169.2	3.66

ST: Sequencing Strategy; GCE: Genome Coverage Rage; SBE: Single Base Error Rate; IDR: Indel Error Rate; FLT: False Gene Duplication Rate; FDT: False Gene Loss Rate; CN: Contig Number; NB: Contig N50 Size (bp); SN: Scaffold Number.

Table 2 (abstract P6). Main average indices for different sequencing strategies for 100 genomes (100-bp read length)

ST	GCE (%)	SBE (%)	IDR (%)	FLT (%)	FDT (%)	CN	NB	SN
6xSE+10xPE	98.06775	0.00498	0.000339	0.4892094	0.190552	72.11	209661.1	4
10xSE+10xPE	98.09051	0.003982	0.000324	0.4596817	0.180621	63.08	240424.9	3.8367
15xSE+10xPE	98.308065	0.004018	0.000322	0.4731213	0.1733068	61.77	241163.8	3.9184
20xSE+10xPE	98.10211	0.004231	0.000339	0.4754001	0.1754001	59.65	244658.8	3.7642

ST: Sequencing Strategy; GCE: Genome Coverage Rage; SBE: Single Base Error Rate; IDR: Indel Error Rate; FLT: False Gene Duplication Rate; FDT: False Gene Loss Rate; CN: Contig Number; NB: Contig N50 Size (bp); SN: Scaffold Number.

Table 3 (abstract P6). Main average indices for different sequencing strategies for 100 genomes (200-bp read length)

ST	GCE (%)	SBE (%)	IDR (%)	FLT (%)	FDT (%)	CN	NB	SN
6xSE+10xPE	98.17144	0.003195	0.000334	0.4401864	0.2416131	61.15	253000.7	3.625
10xSE+10xPE	98.15661	0.004024	0.000317	0.4076573	0.2861061	54.33	290749.3	3.7188
15xSE+10xPE	98.16915	0.004743	0.000305	0.3916122	0.261398	53.47	301038.3	3.64
20xSE+10xPE	98.17177	0.004877	0.000309	0.409125	0.2509012	52.98	289864.6	3.6

Table 4 (abstract P6). Linear regression results for 100 genomes, between the genome quality indicators and, for various read lengths, the number of repeats in the genome, the total repeat length of the genome and the percentage of the total repeat length of the genome

	Repeat length		Repeat length (>300)		Repeat length (>700)	
	R^2	P-value	R^2	P-value	R^2	P-value
6XSE+10XPE, 400bp						
Number of Contigs	0.5657	2.2E-16	0.7842	2.2E-16	0.7948	2.2E-16
N50 size of Contigs	0.07932	0.00453	0.1107	0.00072	0.1114	0.0006918
Genome coverage	0.1298	0.0002314	0.2295	4.591E-07	0.2545	*.732E-08
SNP error rate	0.04819	0.0282	0.09175	0.002189	0.08484	0.003282
Inde1 error rate	0.002337	0.6329	0.04038	0.53	0.003728	0.5462
se gene duplication	0.2951	5.227E-09	0.2969	4.598E-09	0.2158	0.000001124
False gene loss rate	0.1978	0.00003553	0.338	2.264E-10	0.3408	1.827E-10
Number of Scaffolds	0.3363	2.565E-10	0.462	7.497E-15	0.4845	9.023E-16
10XSE+10XPE, 400bp						
Number of Contigs	0.4762	2E-15	0.6908	2.2E-16	0.7164	2.2E-16
N50 size of Contigs	0.05194	0.02258	0.09437	0.001878	0.09966	0.001377
Genome coverage	0.1185	0.0004542	0.2119	0.000001443	0.2358	0.000000305
SNP error rate	0.02702	0.1022	0.06257	0.01207	0.06363	0.01134
Inde1 error rate	0.0006153	0.8065	0.001432	0.7085	0.001119	0.7411
se gene duplication	0.3133	1.414E-09	0.324	6.457E-10	0.2426	1.936E-07
False gene loss rate	0.1232	0.0003429	0.2021	0.000002708	0.1943	0.000004425
Number of Scaffolds	0.2813	1.384E-08	0.4074	9.141E-13	0.4424	4.417E-14
15XSE+10XPE, 400bp						
Number of Contigs	0.453	1.709E-14	0.6676	2.2E-16	0.7008	2.2E-16
N50 size of Contigs	0.01038	0.3131	0.07265	0.006691	0.07771	0.004978
Genome coverage	0.1149	0.0005616	0.02068	0.00002001	0.2323	3.837E-07
SNP error rate	0.0001226	0.913	0.0004724	0.83	0.0002939	0.8656
Inde1 error rate	0.0001226	0.913	0.0004724	0.83	0.0002939	0.8656
se gene duplication	0.3217	7.638E-10	0.3318	3.595E-10	0.2468	1.465E-07
False gene loss rate	0.1541	0.00005366	0.2604	5.834E-08	0.2642	4.519E-08
Number of Scaffolds	0.4023	1.399E-12	0.5996	2.2E-16	0.5878	2.2E-16
6XSE+10XPE, 400bp						
Number of Contigs	0.448	2.696E-14	0.6554	2.2E-16	0.6869	2.2E-16
N50 size of Contigs	0.05142	0.02328	0.09641	0.001666	0.1006	0.001301
Genome coverage	0.1152	0.000551	0.2076	0.000019	0.2338	3.467E-07
SNP error rate	0.2124	0.000001398	0.3199	8.7E-10	0.3315	3.678E-10
Inde1 error rate	0.00001646	0.968	0.00016	0.9006	0.00006389	0.937
se gene duplication	0.3492	9.627E-11	0.3761	1.182E-11	0.2922	6.453E-09
False gene loss rate	0.1163	0.000515	0.2011	0.000002892	0.1938	0.000004569
Number of Scaffolds	0.3125	1.495E-09	0.458	1.09E-14	0.4898	5.431E-16

few days. However, the sequencing results always turn out to contain several hundred contigs. A multiplex PCR procedure is then needed to fill all of the gaps and to link the contigs into one full-length genome sequence [1-10]. The full-length prokaryotic genome sequence is the gold standard for comparative prokaryotic genome analysis. This study assessed pyrosequencing strategies by using a simulation with 100 prokaryotic genomes.

Results Our simulation shows the following: first, a single-end 454 Jr Titanium run combined with a paired-end 454 Jr Titanium run may assemble about 90% of 100 genomes into <10 scaffolds and 95% of 100 genomes into <150 contigs; second, the average contig N50 size is more than 331 kb (Table 1); third, the average single base accuracy is >99.99% (Table 1); fourth, the average false gene duplication rate is <0.7% (Table 1); fifth, the average false gene loss rate is <0.4% (Table 1); sixth, the total size of long repeats

(both repeat length >300 bp and >700 bp) is significantly correlated to the number of contigs (Table 4); and, seventh, increasing the read length of a pyrosequencing run could improve the assembly quality significantly (Table 1-3).

Conclusions A single-end 454 Jr run combined with a paired-end 454 Jr run is a good strategy for prokaryotic genome sequencing. This strategy provides a solution to producing a high-quality draft genome sequence of almost any prokaryotic organism, selected at random, within days. It could be the first step to achieving the full-length genome sequence. It also makes the subsequent multiplex PCR procedure (for gap filling) much easier, aided by the knowledge of the orders/orientations of most of the contigs. As a result, large-scale full-length prokaryotic genome-sequencing projects could be finished within weeks.

References

1. Arnold IC, Zigova Z, Holden M, Lawley TD, Rad R, Dougan G, Falkow S, Bentley SD, Müller A: Comparative whole genome sequence analysis of the carcinogenic bacterial model pathogen *Helicobacter felis*. *Genome Biol Evol* 2011, **3:**302-308.
2. Stephan R, Lehner A, Tischler P, Rattei T: Complete genome sequence of *Cronobacter turicensis* LMG 23827, a food-borne pathogen causing deaths in neonates. *J Bacteriol* 2011, **193:**309-310.
3. Wibberg D, Blom J, Jaenicke S, Kollin F, Rupp O, Scharf B, Schneiker-Bekel S, Sczcepanowski R, Goesmann A, Setubal JC, Schmitt R, Pühler A, Schlüter A: Complete genome sequencing of *Agrobacterium* sp. H13-3, the former *Rhizobium lupini* H13-3, reveals a tripartite genome consisting of a circular and a linear chromosome and an accessory plasmid but lacking a tumor-inducing Ti-plasmid. *J Biotechnol* 2011, **155:**50-62.
4. Song JY, Jeong H, Yu DS, Fischbach MA, Park HS, Kim JJ, Seo JS, Jensen SE, Oh TK, Lee KJ, Kim JF: Draft genome sequence of *Streptomyces clavuligerus* NRRL 3585, a producer of diverse secondary metabolites. *J Bacteriol* 2010, **192:**6317-6318.
5. Gao F, Wang Y, Liu YJ, Wu XM, Lv X, Gan YR, Song SD, Huang H: Genome sequence of *Acinetobacter baumannii* MDR-TJ. *J Bacteriol* 2011, **193:**2365-2366.
6. Powney R, Smits THM, Sawbridge T, Frey B, Blom J, Frey JE, Plummer KM, Beer SV, Luck J, Duffy B, Rodoni B: Genome sequence of an *Erwinia amylovora* strain with pathogenicity restricted to *Rubus* plants. *J Bacteriol* 2011, **193:**785-786.
7. Nam SH, Choi SH, Kang A, Kim DW, Kim RN, Kim A, Kim DS, Park HS: Genome sequence of *Lactobacillus farciminis* KCTC 3681. *J Bacteriol* 2011, **193:**1790-1791.
8. Chen C, Kittichotirat W, Chen W, Downey JS, Si Y, Bumgarner R: Genome sequence of naturally competent *Aggregatibacter actinomycetemcomitans* serotype a strain D7S-1. *J Bacteriol* 2010, **192:**2643-2644.
9. Seth-Smith HMB, Harris SR, Rance R, West AP, Severin JA, Ossewaarde JM, Cutcliffe LT, Skilton RJ, Marsh P, Parkhill J, Clarke IN, Thomson NR: Genome sequence of the zoonotic pathogen *Chlamydophila psittaci*. *J Bacteriol* 2011, **193:**1282-1283.
10. Lyons E, Freeling M, Kustu S, Inwood W: Using genomic sequencing for classical genetics in *E. coli* K12. *PLoS ONE* 2011, **6:**e16717.

P7

Functional exploration of *CCNE1* splicing forms as a possible link to bladder cancer susceptibility

Indu Kohaar[1], Alexandra Scott-Johnson[1], Yi-Ping Fu[1], Patricia Porter-Gill[1] and Ludmila Prokunina-Olsson[1]
[1]Laboratory of Translational Genomics, Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Bethesda, MD 20892, USA
Genome Biology 2011, **12(Suppl 1):**P7

Background A recent genome-wide association study (GWAS) identified a single nucleotide polymorphism, rs8102137, located 6 kb upstream of the cyclin E1 gene (*CCNE1*) on chromosome 19q12, as a risk factor for bladder cancer (odds ratio (OR) = 1.13, $P = 1.7 \times 10^{-11}$) [1]. *CCNE1* encodes a cell cycle protein that regulates cyclin-dependent kinases and is therefore an important cancer susceptibility gene.

Methods This study used 42 bladder tumor samples and 41 normal bladder tissue samples (24 matched normal-tumor pairs), HeLa cells and several prostate and bladder cancer cell lines. Genotyping of rs8102137 in DNA and rs7257694 in both DNA and cDNA samples was performed using an allelic discrimination genotyping assay. TaqMan and SYBR Green assays were used to measure the expression of the different *CCNE1* isoforms. The *CCNE1* isoforms were cloned into a pFC14A (HaloTag) CMV Flexi Vector. Protein expression of *CCNE1* isoforms in normal and tumor bladder tissues and transfected cells was analyzed by western blotting. Subcellular localization of recombinant *CCNE1* splicing forms was analyzed by confocal microscopy.

Results *CCNE1* mRNA was expressed at a higher level in bladder tumors ($n = 42$) than in adjacent normal bladder tissue samples ($n = 41$, 3.7-fold, $P = 2.7 \times 10^{-12}$). However, no association was found between mRNA expression level and the genotype of rs8102137. We observed strong allelic expression imbalance for a synonymous coding variation located in the last exon (rs7257694, Ser390Ser), which is in high linkage disequilibrium with rs8102137 (normal bladder tissue samples, $n = 41$, $D' = 1.0$, $r^2 = 0.815$; HapMap CEU samples, $n = 60$, $D' = 0.95$, $r^2 = 0.68$). In normal and tumor tissue samples heterozygous for both single nucleotide polymorphisms, the risk variant of rs8102137 was associated with lower expression of allele T of rs7257694 (normal samples, $P = 2.2 \times 10^{-4}$; tumor samples, $P = 1.11 \times 10^{-10}$). Western blotting analysis of bladder tissue and prostate cell line lysates revealed that the allelic expression imbalance is likely to be related to two CCNE1 protein isoforms that showed a differential pattern of expression dependent on the rs8102137 and rs7257694 genotype. We have cloned the alternative splicing forms of *CCNE1* and are currently evaluating their functional relevance.

Conclusions Our results suggest that bladder-cancer-associated genetic variants of the *CCNE1* gene might contribute to altered cell cycle regulation, owing to differential mRNA splicing producing different protein isoforms of CCNE1.

Reference

1. Rothman N, Garcia-Closas M, Chatterjee N, Malats N, Wu X, Figueroa JD, Real FX, Van Den Berg D, Matullo G, Baris D, Thun M, Kiemeney LA, Vineis P, De Vivo I, Albanes D, Purdue MP, Rafnar T, Hildebrandt MA, Kiltie AE, Cussenot O, Golka K, Kumar R, Taylor JA, Mayordomo JI, Jacobs KB, Kogevinas M, Hutchinson A, Wang Z, Fu YP, Prokunina-Olsson L *et al.*: A multi-stage genome-wide association study of bladder cancer identifies multiple susceptibility loci. *Nat Genet* 2010, **42:**978-984.

P8

ScaffViz: visualizing metagenome assemblies

Sergey Koren[1,2], Todd Treangen[2,3] and Mihai Pop[1,2]
[1]Department of Computer Science, University of Maryland, College Park, MD 20742, USA; [2]Center for Bioinformatics and Computational Biology, University of Maryland, College Park, MD 20742, USA; [3]The McKusick-Nathans Institute for Genetic Medicine, The Johns Hopkins University School of Medicine, Baltimore, MD 21205, USA
Genome Biology 2011, **12(Suppl 1):**P8

Background Metagenomics has allowed the study of a wide range of microbial communities, from those within the sea [1,2] to those of the human body [3]. Increasingly, *de novo* assembly is the first step in the analysis of these metagenomic samples. As the targets have increased in complexity, computational tools have started to emerge [4,5] to address the challenges presented by the assembly of these datasets. Although the targets and analyses have become more complex, the means of presenting the results has remained the same: a multi-FASTA text file. This presentation hides the variation that is present in the sampled biological community. The ability to navigate and view the complexity of a genomic sample may help drive novel biological insights. Here, we present a graphical visualization tool that allows the visual inspection of genome assembly graphs and the characterization of the genomic variation that is present in these graphs (that is, the differences between two or more related haplotypes commonly found in metagenomes or higher eukaryotes).

Methods Our software, ScaffViz [6], is open source and was developed as a plug-in for the Cytoscape graph viewer package [7,8]. Our assembly view represents assembly metadata within node/edge attributes. For example, node height corresponds to coverage (the amount of oversampling of a sequence), and node width is proportional to the length of the sequence. We support assemblies from Celera Assembler [9], Newbler [10], Bambus 2 and MetAMOS. The creation and initialization of Cytoscape objects is abstracted to allow a developer to easily add new assembly result formats without knowledge of Cytoscape's API. We developed a layout algorithm based on information from the assembler on node position, orientation and length. ScaffViz allows users to show (or hide) an arbitrary subset of nodes. The

viewer can also output genome sequence that corresponds to any subset of the graph, including all alternative sequences present in all selected subpaths. We believe that this representation may prove to be instrumental in finding and characterizing structural variants such as alternative genes, alternative regulatory units or mobile genomic elements.

Results We evaluated the performance of ScaffViz on seven datasets of varying size and complexity. We report that the run time is approximately linear with respect to the number of elements in the graph (nodes + edges). The memory scales linearly with respect to the number of nodes. Extrapolating from these factors, a graph of 250,000 contigs can be opened in approximately 2 minutes using approximately 2.5 GB of memory. ScaffViz is scalable to large graphs and can be run on a laptop.

Conclusions We have developed a novel open-source assembly graph viewer, ScaffViz, as a plug-in for Cytoscape. ScaffViz supports the output of several popular assembly programs and is scalable to large metagenomic assemblies on a laptop.

References

1. Venter J, Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen JA, Wu D, Paulsen I, Nelson KE, Nelson W, Fouts DE, Levy S, Knap AH, Lomas MW, Nealson K, White O, Peterson J, Hoffman J, Parsons R, Baden-Tillson H, Pfannkoch C, Rogers YH, Smith HO: **Environmental genome shotgun sequencing of the Sargasso Sea.** *Science* 2004, **304:**66-74.
2. Rusch DB, Halpern AL, Sutton G, Heidelberg KB, Williamson S, Yooseph S, Wu D, Eisen JA, Hoffman JM, Remington K, Beeson K, Tran B, Smith H, Baden-Tillson H, Stewart C, Thorpe J, Freeman J, Andrews-Pfannkoch C, Venter JE, Li K, Kravitz S, Heidelberg JF, Utterback T, Rogers YH, Falcón LI, Souza V, Bonilla-Rosso G, Eguiarte LE, Karl DM, Sathyendranath S *et al.*: **The *Sorcerer II* Global Ocean Sampling expedition: northwest Atlantic through eastern tropical Pacific.** *PLoS Biol* 2007, **5:**e77.
3. Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, Manichanh C, Nielsen T, Pons N, Levenez F, Yamada T, Mende DR, Li J, Xu J, Li S, Li D, Cao J, Wang B, Liang H, Zheng H, Xie Y, Tap J, Lepage P, Bertalan M, Batto JM, Hansen T, Le Paslier D, Linneberg A, Nielsen HB, Pelletier E, Renault P *et al.*: **A human gut microbial gene catalogue established by metagenomic sequencing.** *Nature* 2010, **464:**59-65.
4. Laserson J, Jojic V, Koller D: **Genovo: *de novo* assembly for metagenomes.** *J Comput Biol* 2011, **18:**429-443.
5. Peng Y, Leung HC, Yiu SM, Chin FY: **Meta-IDBA: a *de novo* assembler for metagenomic data.** *Bioinformatics* 2011, **27:**i94-i101.
6. **ScaffViz Project** [http://code.google.com/p/scaffold-viewer/]
7. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T: **Cytoscape: a software environment for integrated models of biomolecular interaction networks.** *Genome Res* 2003, **13:**2498-2504.
8. Smoot M, Ono K, Ruscheinski J, Wang P, Ideker T: **Cytoscape 2.8: new features for data integration and network visualization.** *Bioinformatics* 2011, **27:**431-432.
9. Miller JR, Delcher AL, Koren S, Venter E, Walenz BP, Brownley A, Johnson J, Li K, Mobarry C, Sutton G: **Aggressive assembly of pyrosequencing reads with mates.** *Bioinformatics* 2008, **24:**2818-2824.
10. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen Z, Dewell SB, Du L, Fierro JM, Gomes XV, Godwin BC, He W, Helgesen S, Ho CH, Irzyk GP, Jando SC, Alenquer ML, Jarvie TP, Jirage KB, Kim JB, Knight JR, Lanza JR, Leamon JH, Lefkowitz SM, Lei M, Li J *et al.*: **Genome sequencing in microfabricated high-density picolitre reactors.** *Nature* 2005, **437:**376-380.

P9

Systems biology of bacteriophage proteins and new dimensions of the virus world discovered through metagenomics

David M Kristensen[1], Arcady R Mushegian[2,3] and Eugene V Koonin[1]
[1]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA; [2]Department of Bioinformatics, Stowers Institute for Medical Research, Kansas City, MO 64110, USA; [3]Department of Microbiology, Molecular Genetics, and Immunology, University of Kansas Medical Center, Kansas City, KS 66160, USA
Genome Biology 2011, **12(Suppl 1):**P9

Most of the DNA viruses in the gastrointestinal tract are phages, which infect bacterial hosts. Despite phages being the most abundant organisms on Earth, as well as extremely active players in the global ecosystem, much remains unknown about how they function in their natural environments. Advances in whole genome sequencing technologies have generated a large collection of hundreds of phage genomes, allowing deep insight into the genetic evolution of phages, and metagenomics technologies seem to promise more rewarding glimpses into their life cycles and community structures.

Recently, we developed an automated approach to assemble a collection of orthologous gene clusters of double-stranded DNA phages (phage orthologous groups, or POGs). This approach follows the well-known clusters of orthologous groups (COGs) framework to identify sets of orthologs by examining top-ranked sequence similarities between proteins in complete genomes without the use of arbitrary similarity cutoffs, and it thus represents a natural system for examining fast-evolving and slow-evolving proteins alike. This automated approach was designed to keep pace with the rapid and accelerating growth of whole genome information from sequencing projects. In particular, we employ a faster graph-theoretical COG-building algorithm that vastly improves our ability to deal with larger numbers of genomes (N) by reducing the worst-case complexity from $O(N^6)$ to $O(N^3 \times \log N)$. This system encompasses more than 2,000 groups from the almost 600 known phage genomes deposited at the National Center for Biotechnology Information and is in the process of being expanded to include single-stranded DNA phages and single- and double-stranded RNA phages.

Using this approach, we found that more than half of the POGs have no or very few evolutionary connections to their cellular hosts, indicating that these phages combine the ability to share and transduce the host genes with the ability to maintain a large fraction of unique, phage-specific, genes. Such genes are useful for targeted research strategies: for example, as diagnostic indicators and fundamental units of systems biology studies. We employed this set of phage-specific genes to probe the composition of several oceanic metagenomic samples. Although virus-enriched samples indeed contain more homologous matches to phage-specific POGs than a full metagenomic sample also containing cellular DNA, the total gene repertoire of the marine DNA virome is dramatically different from that of known phages. In particular, it is dominated by rare genes, many of which might be contained within virus-like entities such as cellular gene transfer agents rather than true viruses. This result might suggest the necessity of radically rethinking what constitutes the 'virus world', because the major component of (marine) viromes could be gene transfer agents that encapsidate bacterial and archaeal genes.

P10

Genetic basis of common human disease: insight into the role of nonsynonymous SNPs from genome-wide association studies

Lipika R. Pal[1] and John Moult[1,2]
[1]Institute for Bioscience and Biotechnology Research, University of Maryland at College Park, Rockville, MD 20850, USA; [2]Department of Cell Biology and Molecular Genetics, University of Maryland at College Park, College Park, MD 20742, USA
Genome Biology 2011, **12(Suppl 1):**P10

Background Recent genome-wide association studies have led to the reliable identification of single nucleotide polymorphisms (SNPs) at a number of loci associated with an increased risk of developing specific common human diseases. Each such locus implicates multiple possible candidate SNPs as being involved in the disease mechanism, and determining which SNPs actually contribute, and by what mechanism, is a major challenge. 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 analyzed the role of one of these mechanisms, nonsynonymous SNPs (nsSNPs) in proteins, for associations found in the Wellcome Trust Case-Control Consortium (WTCCC) study of seven common diseases [1] and the follow-up work.

Methods Using HapMap data and linkage disequilibrium information, we identified all possible candidate SNPs associated with increased disease risk. We then applied two computational methods [2,3], based on analysis of protein structure and sequence, to determine which of these SNPs has a significant impact on *in vivo* protein function (SNPs3D) [4].

Results Several of these disease-associated loci were found to be linked to one or more high-impact nsSNPs. In some cases, these SNPs are in well-known proteins (such as human leukocyte antigens). In other cases, they are in less well-established disease-associated genes (for example, MST1 for Crohn's disease), and in yet others, they are in proteins that have been poorly investigated (for example, gasdermin B, also for Crohn's disease). Approximately 55% of these disease-associated loci have at least one nsSNP, and about 33% of them have at least one high-impact nsSNP in those regions.

Conclusions Together, these data suggest a significant role for nsSNPs in common human disease susceptibility.

References

1. Wellcome Trust Case-Control Consortium: **Genome wide association study of 14000 cases of seven common diseases and 3,000 shared controls.** *Nature* 2007, **447:**661-678.
2. Yue P, Li Z, Moult J: Loss of protein structure stability as a major causative factor in monogenic disease. *J Mol Biol* 2005, **353:**459-473.
3. Yue P, Moult J: **Identification and analysis of deleterious human SNPs.** *J Mol Biol* 2006, **356:**1263-1274.
4. **SNPs3D Project** [http://www.snps3d.org]

P11

Accurate and fast estimation of taxonomic profiles from metagenomic shotgun sequences

Bo Liu[1,2], Theodore Gibbons[1,3], Mohammad Ghodsi[1,2], Todd Treangen[1] and Mihai Pop[1-3]
[1]Center for Bioinformatics and Computational Biology, University of Maryland, College Park, MD 20742, USA; [2]Department of Computer Science, University of Maryland, College Park, MD 20742, USA; [3]Biological Sciences Graduate Program, University of Maryland, College Park, MD 20742, USA
Genome Biology 2011, **12(Suppl 1):**P11

Background A major goal of metagenomics is to characterize the taxonomic composition of an environment. The most popular approach relies on 16S rRNA sequencing; however, this approach can generate biased estimates owing to differences in the copy number of the gene, even between closely related organisms, and owing to PCR artifacts. In addition, the taxonomic composition can also be determined from metagenomic shotgun sequences by matching reads against a database of reference sequences. One major limitation of the computational methods that have been used for this purpose is the use of a universal classification threshold for all genes at all taxonomic ranks.

Methods We present a novel taxonomic profiler for metagenomic sequences, MetaPhyler [1], which relies on 31 phylogenetic marker genes as a taxonomic reference. Because genes can evolve at different rates and because shotgun reads contain gene fragments of different lengths, we propose that better classification results can be obtained by tuning the taxonomic classifier to the length of the gene fragment, to a particular gene and to the taxonomic rank. Our classifier uses different thresholds for each of these parameters, and these thresholds are automatically learned from the taxonomic structure of the reference database.

Results We have randomly simulated about 300,000 DNA sequences of 60 bp and about 70,000 DNA sequences of 300 bp from phylogenetic marker genes. Table 1 shows the performance of the phylogenetic classifications from MetaPhyler, PhymmBL [2], MEGAN [3] and WebCARMA [4]. The query sequence itself was removed from the reference dataset when running the programs. The sensitivity of MetaPhyler is significantly higher than that of the other tools in all situations because our classifier is explicitly trained at each taxonomic rank.

In addition, we have created a simulated metagenomic sample comprising five genomes. Table 2 shows the taxonomic profiles estimated by different approaches. In this setting, MetaPhyler also outperforms the other approaches by more accurately reconstructing the true taxonomic distribution.

Conclusions We have introduced a novel taxonomic classification method for analyzing the microbial diversity from whole metagenome shotgun sequences. Compared with previous approaches, MetaPhyler is more

Table 1 (abstract P11). Comparison of sensitivity and precision.

Sequence length	Parameter	Taxonomic rank	MetaPhyler (%)	PhymmBL (%)	MEGAN (%)	WebCARMA (%)
60 bp	Sensitivity	Genus	33.45	18.18	15.49	22.66
		Family	54.22	38.75	24.52	25.10
		Order	59.59	49.36	31.74	28.22
		Class	70.72	62.86	50.78	32.12
		Phylum	75.30	68.88	64.19	34.65
	Precision	Genus	96.38	94.42	90.72	35.22
		Family	97.45	97.66	97.18	45.71
		Order	97.39	97.65	98.10	52.51
		Class	98.27	98.15	99.11	66.15
		Phylum	98.83	99.06	99.56	72.90
300 bp	Sensitivity	Genus	52.39	42.97	20.89	45.96
		Family	70.17	58.81	34.27	52.49
		Order	78.09	66.72	45.24	58.56
		Class	84.52	75.42	61.06	62.70
		Phylum	91.18	76.78	81.36	66.49
	Precision	Genus	97.90	96.16	96.09	77.63
		Family	99.14	99.07	99.19	88.69
		Order	99.15	99.15	99.21	92.67
		Class	99.34	99.34	99.57	95.43
		Phylum	99.64	99.64	99.80	96.58

Table 2 (abstract P11). Comparison of taxonomic profile estimations.

Genus	True (%)	MetaPhyler (%)	PhymmBL (%)	MEGAN (%)	WebCARMA (%)
Bifidobacterium	50.0	50.0	34.3	32.8	34.3
Bacteroides	20.0	20.4	32.1	34.3	33.8
Staphylococcus	10.0	10.2	9.4	9.1	8.9
Enterococcus	10.0	10.1	9.0	7.3	10.4
Clostridium	10.0	9.4	11.8	12.1	12.6
Other	0.0	0.0	3.6	4.4	0.1

accurate at estimating the taxonomic profile, especially when taking into account the actual abundance of individual taxonomic groups.

References

1. MetaPhyler Software [http://metaphyler.cbcb.umd.edu/]
2. Brady A, Salzberg SL: **Phymm and PhymmBL: metagenomic phylogenetic classification with interpolated Markov models.** *Nat Methods* 2009, **6:**673-676.
3. Huson DH, Auch AF, Qi J, Schuster SC: **MEGAN analysis of metagenomic data.** *Genome Res* 2007, **17:**377-386.
4. Gerlach W, Junemann S, Tille F, Goesmann A, Stoye J: **WebCARMA: a web application for the functional and taxonomic classification of unassembled metagenomic reads.** *BMC Bioinformatics* 2009, **10:**430.

P12

Abstract not submitted for online publication.

P13

Massively parallel sequencing identifies a previously unrecognized X-linked disorder resulting in lethality in male infants owing to amino-terminal acetyltransferase deficiency

Alan F Rope[1], Kai Wang[2,3], Rune Evjenth[4], Jinchuan Xing[5], Jennifer J Johnston[6], Jeffrey J Swensen[7,8], W Evan Johnson[9], Barry Moore[5], Chad D Huff[5], Lynne M Bird[10], John C Carey[1], John M Opitz[1,5,7,11], Cathy A Stevens[12], Christa Schank[9], Heidi Deborah Fain[13], Reid Robison[13], Brian Dalley[14], Steven Chin[7], Sarah T South[1,8], Theodore J Pysher[7], Lynn B Jorde[5], Hakon Hakonarson[2], Johan R Lillehaug[4], Leslie G Biesecker[6], Mark Yandell[5], Thomas Arnesen[4,15] and Gholson J Lyon[13,16,17]

[1]Department of Pediatrics (Medical Genetics), University of Utah School of Medicine, Salt Lake City, UT, USA; [2]Center for Applied Genomics, Children's Hospital of Philadelphia, Philadelphia, PA, USA; [3]Present address: Zilkha Neurogenetic Institute, Department of Psychiatry and Preventive Medicine, University of Southern California, Los Angeles, CA, USA; [4]Department of Molecular Biology, University of Bergen, N-5020 Bergen, Norway; [5]Eccles Institute of Human Genetics, University of Utah, Salt Lake City, UT, USA; [6]Genetic Disease Research Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD, USA; [7]Department of Pathology, University of Utah, Salt Lake City, UT, USA; [8]ARUP Laboratories, Salt Lake City, Utah, USA; [9]Department of Statistics, Brigham Young University, Provo, Utah, USA; [10]Rady Children's Hospital and University of California, San Diego, Department of Pediatrics, San Diego, CA, USA; [11]Department of Obstetrics and Gynecology, University of Utah, Salt Lake City, UT, USA; [12]Department of Pediatrics, University of Tennessee College of Medicine, Chattanooga, TN, USA; [13]Department of Psychiatry, University of Utah, Salt Lake City, UT, USA; [14]Huntsman Cancer Institute, Salt Lake City, UT, USA; [15]Department of Surgery, Haukeland University Hospital, N-5021 Bergen, Norway; [16]New York University Child Study Center, New York, NY, USA; [17]Present address: Center for Applied Genomics, Children's Hospital of Philadelphia, Philadelphia, PA, USA

Genome Biology 2011, **12(Suppl 1):**P13

Background Individuals II-1 and II-6 from family 1 presented in the mid-1980s to the University of Utah Medical Center. These boys had a striking similarity to each other, with a range of shared clinical manifestations, but the disease they presented with was not a recognized syndrome. Both boys subsequently died in infancy. X-linked inheritance was confirmed in the next generation, when individuals III-4 and III-7 presented with the same disease. Their aged appearance was the most striking part of the disease.

Methods We describe two parallel genetic research efforts that converged on the same gene variant. Exon capture was carried out on samples from two families, using a commercially available in-solution method (Agilent's SureSelect Human X Chromosome kit) as per the manufacturer's guidelines with minor modifications to generate sequencing libraries (Illumina). We also used a recently developed tool, the Variant Annotation, Analysis and Selection Tool (VAAST), which identifies disease-causing variants, to analyze the exon capture data from family 1. Our analysis applied a disease model that did not require complete penetrance or locus homogeneity. We restricted the expected allele frequency of putative disease-causing variants within the control genomes to 0.1% or lower. The background file used in the analysis is composed of variants from dbSNP (version 130), 189 genomes from the 1000 Genomes Project, the 10Gen Data Set, 184 Danish exomes and 40 whole genomes from the Complete Genomics Diversity Panel. VAAST candidate gene prioritization analysis was performed using the likelihood ratio test under the dominant inheritance model, assuming an expected allele frequency of 0.1% or lower for the causal variant in the general population. After masking out loci of potentially low variant quality, single nucleotide variations in each gene were scored as a group. The significance level was assessed using individual permutation tests.

Results We identified a family with a previously undescribed lethal X-linked disorder of infancy comprising a distinct combination of an aged appearance, craniofacial anomalies, hypotonia, global developmental delays, cryptorchidism, cardiac arrhythmia and cardiomyopathy. We used X-chromosome exon sequencing and a recently developed probabilistic disease-gene discovery algorithm to identify a missense variant in *NAA10*, which encodes the catalytic subunit of the major human amino-terminal acetyltransferase (NAT; also known as hNaa10p). More recently, we became aware that a parallel effort on a second unrelated family converged on the same variant. The absence of this variant in controls, the amino acid conservation of this region of the protein, the predicted disruptive change and the co-occurrence in two unrelated families with the same rare disorder suggest that this is the pathogenic mutation. We confirmed this by demonstrating that the mutant hNaa10p had significantly impaired biochemical activity, and we therefore conclude that a reduction in acetylation by hNaa10p causes this disease.

Conclusions This is one of the first uses of next-generation sequencing to identify the genetic basis of a previously unrecognized X-linked syndrome. It is also the first evidence of a human genetic disorder resulting from direct impairment of amino-terminal acetylation, one of the most common protein modifications in humans. We have also demonstrated that a probabilistic disease-gene discovery algorithm (VAAST) can readily identify and characterize the genetic basis of this syndrome.

P14

Abstract not submitted for online publication.

P15

Prostate stem cell antigen (*PSCA*) and risk of bladder cancer: linking genotypes to functional mechanisms

Adam Mumy[1], Indu Kohaar[1], Patricia Porter-Gill[1], Wei Tang[1], Yi-Ping Fu[1] and Ludmila Prokunina-Olsson[1]

[1]Laboratory of Translational Genomics, Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Bethesda, MD 20892, USA

Genome Biology 2011, **12(Suppl 1):**P15

Background Genome-wide association studies (GWAS) have identified a single nucleotide polymorphism, rs2294008 C/T, within the prostate stem cell antigen (*PSCA*) gene as a risk variant for bladder cancer [1]. PSCA is a glycosyl phosphatidylinositol (GPI)-anchored cell surface protein from the Ly-6/Thy-1 family of cell surface antigens. PSCA overexpression has been reported in bladder, prostate and pancreatic tumors. The risk allele (T) of rs2294008 creates a novel translation start site and extends the PSCA leader peptide sequence by 11 amino acids.

Methods The mRNA expression in 42 bladder tumor samples and 39 adjacent normal bladder tissue samples (24 matched normal-tumor pairs) was explored using genome-wide RNA sequencing and targeted *PSCA* mRNA expression assays. For allelic expression imbalance studies, genotyping of rs2294008 both in DNA and cDNA samples was performed using an allelic discrimination genotyping assay. Alternative allele-specific splicing forms of *PSCA* were cloned and transfected into several human cancer cell lines. The endogenous expression of PSCA protein and the expression pattern of the recombinant PSCA allelic isoforms in different cancer cell lines were studied by western blotting, confocal microscopy and fluorescence-activated cell-sorting analysis. PSCA protein expression in normal and tumor bladder tissue samples was examined in relation to rs2294008 genotypes by using immunohistochemistry.

Results *PSCA* mRNA was expressed at a 5.7-fold higher level in tumors than in matching normal bladder tissue samples ($P = 0.0060$). There was a strong allelic expression imbalance in tumor samples ($P = 0.0020$), based on 20 normal and 13 tumor samples that were heterozygous for rs2294008. *PSCA* mRNA expression was associated with the genotype of rs2294008 both in normal and tumor bladder tissue samples. Our preliminary data on the expression of recombinant allele-specific PSCA protein isoforms in transfected cells show a possible difference in the distribution of the cytoplasmic and membrane expression of these isoforms.

Conclusions Our results suggest that the extension of the PSCA leader peptide by 11 amino acids, introduced by the risk allele (T) of rs2294008, may affect subcellular protein localization and the availability of functional GPI-anchored PSCA on the cell surface. These results may have clinical implications because antibodies that target cell-surface-expressed PSCA are in clinical trials for pancreatic and prostate cancer.

Reference

1. Rothman N, Garcia-Closas M, Chatterjee N, Malats N, Wu X, Figueroa JD, Real FX, Van Den Berg D, Matullo G, Baris D, Thun M, Kiemeney LA, Vineis P, De Vivo I, Albanes D, Purdue MP, Rafnar T, Hildebrandt MA, Kiltie AE, Cussenot O, Golka K, Kumar R, Taylor JA, Mayordomo JI, Jacobs KB, Kogevinas M, Hutchinson A, Wang Z, Fu YP, Prokunina-Olsson L *et al.*: **A multi-stage genome-wide association study of bladder cancer identifies multiple susceptibility loci.** *Nat Genet* 2010, **42:**978-984.

P16

Metabolic pathways are affected differently in *Saccharomyces cerevisiae* thiol peroxidase mutants during redox stress with hydrogen peroxide

Amy L Olex[1], Brian M Westwood[2], Leslie B Poole[3] and Jacquelyn S Fetrow[1,4]
[1]Department of Computer Science, Wake Forest University, Winston-Salem, NC 27109, USA; [2]Department of Molecular Genetics and Genomics, Wake Forest University School of Medicine, Winston-Salem, NC 27157, USA; [3]Department of Biochemistry, Wake Forest University School of Medicine, Winston-Salem, NC, 27157, USA; [4]Department of Physics, Wake Forest University, Winston-Salem, NC 27109, USA
Genome Biology 2011, **12(Suppl 1):**P16

Background Thiol peroxidases have been conserved throughout evolution and are found in almost every known organism from bacteria to humans. These proteins play a key role in maintaining redox homeostasis and have been implicated in other processes such as cell signaling and sensing hydrogen peroxide and passing this signal along to transcription factors. To gain a better understanding of the role that each thiol peroxidase plays in redox regulation on a global level, Fomenko and colleagues [1] performed a series of microarray experiments in which different combinations of the genes encoding the eight thiol peroxidases (three glutathione peroxidase homologs (Gpx) and five peroxiredoxins (Prx)) present in yeast were knocked out, including one mutant (8-Δ) in which all eight peroxidases were removed. Surprisingly, all of the mutants, including 8-Δ, were viable and could withstand redox stresses; however, they were unable to activate or repress transcriptional events in response to hydrogen peroxide treatment, which was most evident in the 8-Δ mutant. In our work, network analysis was used to gain a better understanding of the biological networks whose gene expression is affected by these mutations.

Methods Microarray data (provided by [1]) was processed for input into the Cytoscape plug-in jActiveModules. Active sub-networks for select mutants were identified using all yeast interactions found in the Kyoto Encyclopedia of Genes and Genomes (KEGG) [2] as the background network (including protein-protein, metabolic and gene expression interactions). Nodes in each sub-network were input into the Database for Annotation, Visualization and Integrated Discovery (DAVID) [3] to identify which KEGG pathways were present.

Results Two hundred and six genes appeared in one or more of the active sub-networks. Only seven genes were present in the sub-networks of all strains. These were a known oxidative stress-induced aldose reductase (*GRE3*), four putative aryl-alcohol dehydrogenases (*AAD3*, *AAD6*, *AAD10* and *AAD14*), a mitochondrial aldehyde dehydrogenase (*ALD4*) and a xylulokinase (*XKS1*). All of the genes were upregulated on average by 6- to 12-fold in all strains, except for 8-Δ with a 1.5-fold average upregulation and 5Prx-Δ with a 3-fold average upregulation.

Many metabolic pathways were affected by the knockouts; the pathway types affected depended on which peroxidase gene was knocked out. This result suggests that different thiol peroxidases may have a significant and specific impact on the regulation of metabolic pathways during oxidative stress.

Surprisingly, the Gpx3-Δ active sub-network was similar to the Gpx1-Δ and Gpx2-Δ sub-networks. Gpx3 is known to sense hydrogen peroxide and pass that signal along to transcription factors; thus, it was expected that this sub-network would differ from that of the other Gpx mutants. Additionally, our results showed that amino acid metabolism, biosynthesis and degradation pathways were active in wild-type cells but were present in few mutant strains.

Conclusions The results of this work indicate that thiol peroxidases, along with playing a key role in maintaining redox homeostasis, may also play a significant role in the regulation of metabolic pathways in yeast, thus illuminating the global role that thiol peroxidases play in oxidative stress.

References

1. Fomenko DE, Koc A, Agisheva N, *et al.*: **Thiol peroxidases mediate specific genome-wide regulation of gene expression in response to hydrogen peroxide.** *Proc Natl Acad Sci USA* 2011, **108:**2729-2734.
2. **The Kyoto Encyclopedia of Genes and Genomes** [http://www.genome.jp/kegg/]
3. **The Database for Annotation, Visualization and Integrated Discovery** [http://david.abcc.ncifcrf.gov/]

P17

Metastats: an improved statistical method for analysis of metagenomic data

Joseph N Paulson[1,2], Mihai Pop[2,3] and Hector Corrada Bravo[2,3]
[1]Applied Mathematics and Scientific Computing Program, University of Maryland, College Park, MD 20742, USA; [2]Center for Bioinformatics and Computational Biology, University of Maryland, College Park, MD 20742, USA; [3]Department of Computer Science, University of Maryland, College Park, MD 20742, USA
Genome Biology 2011, **12(Suppl 1):**P17

Metagenomic studies were originally focused on exploratory/validation projects but 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 on 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 and functional groups) from two treatment populations (for example, healthy versus 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 allow 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 owing to undersampling 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 the corresponding odds ratio, and penetrance calculations) that improve the usability of our software in clinical practice.

Reference

1. White JR, Nagarajan N, Pop M: **Statistical methods for detecting differentially abundant features in clinical metagenomic samples.** *PLoS Comput Biol* 2009, **5:**e1000352.

P18

Detection of bladder, breast and prostate cancer using serum and tissue miRNA profiling

[1]Patricia Porter-Gill, [1]Yi-Ping Fu, [1]Alpana Kaushiva, [2]Douglas Price, [2]William Dahut, [2]William Figg, [1]Ludmila Prokunina-Olsson
[1]Laboratory of Translational Genomics, Division of Cancer Epidemiology and Genetics, NCI/NIH, Bethesda, 20892, USA; [2]Molecular Oncology Branch, NCI/NIH, Bethesda, 20892, USA
Genome Biology 2011, **12(Suppl 1):**P18

Background miRNAs are short, non-coding regulatory RNA molecules that can bind to complementary sequences on target mRNAs, resulting

in translational repression and gene silencing. miRNAs are attractive as biomarkers, because they are stable in various conditions and are easy to measure using quantitative PCR methods. Biomarkers that can differentiate between normal and tumor states and can be measured in easily accessible body fluids, such as blood and urine, are important for cancer diagnostics and disease monitoring.

Methods and results In this study, we identified a universal panel of miRNAs for cancer detection. This panel can easily be used to screen the serum of healthy individuals and patients with different types of cancer. First, we measured the expression of about 800 miRNAs in 40 control individuals and 60 patients with bladder, breast or prostate cancer using TaqMan Low Density gene expression arrays (Applied Biosystems), starting from 250 µl serum. On the basis of these results, we selected a panel of 24 miRNAs that showed the best discrimination between normal samples and cancer samples. These miRNAs were then retested as a custom-designed mini-panel on serum samples from 44 healthy controls and from patients with cancer (31 with bladder cancer, 25 with breast cancer and 28 with prostate cancer), as well as in relevant normal and tumor tissue samples (42 normal bladder samples and 43 bladder tumors, 44 normal breast samples and 42 breast tumors, and 50 normal prostate samples and 20 prostate tumors). Only miRNAs with changes in expression in the same direction in serum and tissue samples and with a significant association with cancer in both sample types were used for further analysis.

The current panel consists of 16 miRNAs: 15 targets and 1 positive control. Using this panel on serum samples from 77 controls, 52 patients with bladder cancer, 48 patients with breast cancer and 34 patients with prostate cancer, we performed receiver operating characteristic (ROC) analysis and achieved complete discrimination (area under the ROC curve (AUC) of about 1.0) between all types of cancers and controls, as well as good discrimination between different types of cancers (minimal AUC of 0.89 for breast and bladder cancer samples).

Conclusions Our results prove that miRNA detection in serum might be a promising method for cancer detection.

P19

An unusual suspect: an uncommon human-specific synonymous coding variant within the *UGT1A6* gene explains a GWAS signal and protects against bladder cancer

Wei Tang[1], Yi-Ping Fu[1], Jonine D Figueroa[2], Núria Malats[3], Montserrat Garcia-Closas[2,4], Nilanjan Chatterjee[2], Manolis Kogevinas[5-8], Dalsu Baris[2], Michael Thun[9], Jennifer L Hall[10], Immaculata De Vivo[11], Demetrius Albanes[2], Patricia Porter-Gill[1], Mark P Purdue[2], Laurie Burdett[12], Luyang Liu[1], Amy Hutchinson[12], Timothy Myers[12], Adonina Tardón[7,13], Consol Serra[14], Alfredo Carrato[15], Reina Garcia-Closas[16], Josep Lloreta[17], Alison Johnson[18], Molly Schwenn[19], Margaret R Karagas[20], Alan Schned[21], Amanda Black[2], Eric J. Jacobs[9], W Ryan Diver[9], Susan M Gapstur[9], Jarmo Virtamo[22], David J. Hunter[23], Joseph F Fraumeni Jr[2], Stephen J Chanock[1], Debra T Silverman[2], Nathaniel Rothman[2] and Ludmila Prokunina-Olsson[1]

[1]Laboratory of Translational Genomics, Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, MD 20892, USA; [2]Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, MD 20892, USA; [3]Spanish National Cancer Research Centre, Madrid 28029, Spain; [4]Division of Genetics and Epidemiology, Institute of Cancer Research, London SW7 3RP, UK; [5]Centre for Research in Environmental Epidemiology (CREAL), Barcelona 08003, Spain; [6]Municipal Institute of Medical Research, Barcelona 08003, Spain; [7]CIBER Epidemiología y Salud Pública (CIBERESP), Barcelona 08003, Spain, [8]National School of Public Health, Athens 11521, Greece; [9]Epidemiology Research Program, American Cancer Society, Atlanta, GA 30303, USA; [10]Lillehei Heart Institute, Department of Medicine, University of Minnesota, Minneapolis, MN 55455, USA; [11]Channing Laboratory, Department of Medicine, Brigham and Women's Hospital, Boston, MA 02115, USA; [12]Core Genotype Facility, SAIC-Frederick, National Cancer Institute, Frederick, MD 21702, USA; [13]Universidad de Oviedo, Oviedo 33003, Spain; [14]Universitat Pompeu Fabra, Barcelona 08002, Spain; [15]Ramón y Cajal University Hospital, Madrid 28034, Spain; [16]Unidad de Investigación, Hospital Universitario de Canarias, La Laguna 38320, Spain; [17]Hospital del Mar-Institut Municipal d'Investigació Mèdica (IMIM), Universitat Pompeu Fabra, Barcelona 08003, Spain; [18]Vermont Cancer Registry, Burlington, VT 05401, USA; [19]Maine Cancer Registry, Augusta, ME 04333, USA; [20]Dartmouth Medical School, Hanover, NH 03755, USA; [21]Department of Urology, Washington University School of Medicine, St. Louis, MO 63110, USA; [22]National Institute for Health and Welfare, Helsinki 00271, Finland; [23]Department of Epidemiology, Program in Molecular and Genetic Epidemiology, Harvard School of Public Health, Boston, MA 02115, USA

Genome Biology 2011, **12(Suppl 1):**P19

Background A recent genome-wide association study (GWAS) of bladder cancer identified a single nucleotide polymorphism (SNP), rs11892031, within the *UGT1A* gene cluster on chromosome 2q37.1, as a novel risk factor. The *UGT1A* locus encodes nine UGT proteins, which belong to the phase II cellular detoxification system. UGTs are functionally important for the detoxification of aromatic amines, which are found in industrial chemicals and tobacco smoke and are known risk factors for bladder cancer. The UGT-encoding genes have exons 2 to 5 in common but have different first exons, which define the enzymatic activity and substrate specificity of the gene products.

Methods and results We sequenced all nine highly similar alternative first exons for the UGT-encoding genes of up to 2,000 individuals. We identified 26 known nonsynonymous and 17 known synonymous coding variants but no novel variants. Imputation based on the GWAS dataset, a combined reference panel of HapMap 3 and the 1000 Genomes Project, and a subset of GWAS samples genotyped for all of the identified coding variants generated data for 1,170 SNPs within the whole *UGT1A* region. Of these markers, the strongest association was detected for an uncommon protective genetic variant that explained the original GWAS signal (odds ratio (OR) = 0.55, 95% confidence interval (CI) = 0.44 to 0.69, $P = 3.3 \times 10^{-7}$ in 4,035 cases and 5, 284 controls; D′ = 0.96, r^2=0.23 with rs11892031). No residual association in this region was detected after adjustment for this SNP. A typical genetic variant identified by GWAS for a common disease is expected to be a common allele (>10% minor allele frequency) that increases the disease risk. We show that the novel associated variant is an uncommon protective allele (1.14% in cases and 2.5% in controls). Interestingly, the risk allele (G) is conserved in 33 species, whereas the protective allele (T) is a human-specific variant. Even though this SNP is a synonymous coding variant, we show its association with quantitative mRNA expression of a specific functional splicing form of *UGT1A6*, probably through an exonic splicing enhancer.

Conclusions This study exemplifies that uncommon protective genetic variants are unusual suspects that may play important but underestimated functional roles in complex traits.

P20

A comprehensive census of horizontal gene transfers from prokaryotes to unikonts

Pere Puigbò[1], Sergei Mekhedov[1], Yuri I Wolf[1] and Eugene V. Koonin[1]

[1]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD, USA

Genome Biology 2011, **12(Suppl 1):**P20

Background Horizontal gene transfers (HGTs) are pervasive in prokaryotes [1], being the routes of net-like evolution that collectively dominate the evolution of prokaryotes [2]. However, in eukaryotes, the effect of HGT has not been thoroughly analyzed, with the exception of the massive HGT from the endosymbionts [3]. Here, we report a comprehensive analysis of likely HGT events in different groups of unikonts (Amoebozoa, Archamoebae, Mycetozoa, the Fungi/Metazoa group, Choanoflagellida, Fungi and Metazoa).

Methods We analyzed the complete proteomes of 36 species of unikonts: 1 from the Archamoebae, 1 from Mycetozoa, 18 from Fungi, 13 from Metazoa and 1 from Choanoflagellida. These proteomes were manually selected to widely represent the unikont supergroup. Initial pre-candidate genes were obtained by analyzing each proteome using the DarkHorse program [4]. The program BLASTClust was then used to make clusters of putative unique transfer events at the origin of the different groups of unikonts. These clusters were separated into two groups: group I candidate clusters (clusters with no eukaryotic representative other than the unikont group analyzed), and group II candidate clusters (clusters with representatives from prokaryotes, the unikont group analyzed and other eukaryotes). Sequences from group I candidate clusters were analyzed using BLAST versus nr and RefSeq databases, compared with the clusters of orthologous groups for eukaryotic complete genomes (KOGs) [5] and manually curated to remove false positives that result from bacterial contamination of the genomic DNA. Group II candidate clusters were analyzed using a series of automatic, conservative filters to assess the quality of the candidates. Finally, all clusters

were phylogenetically analyzed to define the final candidates and to infer putative donors.
Results Using this methodology, we detected numerous probable HGT events from prokaryotes (mainly Bacteria) to unikonts. These events are not distributed uniformly throughout the evolution of unikonts: for example, almost all HGTs detected in Amoebozoa occurred after the divergence of Archamoebae and Mycetozoa. Importantly, we also detected many HGT events from Bacteria to Fungi, Choanoflagellida and Metazoa**Conclusions** Although HGTs are not as pervasive in eukaryotes as in prokaryotes, the amount of HGT detected in this study suggests that the acquisition of genes from Bacteria played a major role in the evolution of the unikonts.
References
1. Puigbò P, Wolf YI, Koonin EV: **Search for a 'Tree of Life' in the thicket of the phylogenetic forest.** *J Biol* 2009, **8:**59.
2. Puigbò P, Wolf YI, Koonin EV: **The tree and net components of prokaryote evolution.** *Genome Biol Evol* 2010, **2**:745-756.
3. Lane CE, Archibald JM: **The eukaryotic tree of life: endosymbiosis takes its TOL.** *Trends Ecol Evol* 2008, **23:**268-275.
4. Podell S, Gaasterland T: **DarkHorse: a method for genome-wide prediction of horizontal gene transfer.** *Genome Biol* 2007, **8:**R16.
5. Koonin EV, Fedorova ND, Jackson JD, Jacobs AR, Krylov DM, Makarova KS, Mazumder R, Mekhedov SL, Nikolskaya AN, Rao BS, Rogozin IB, Smirnov S, Sorokin AV, Sverdlov AV, Vasudevan S, Wolf YI, Yin JJ, Natale DA: **A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes.** *Genome Biol* 2004, **5:**R7.

P21

A genetic survival network for glioblastoma multiforme

Andy Lin[1] and Desmond J Smith[1]
[1]Department of Molecular and Medical Pharmacology, David Geffen School of Medicine, University of California, Los Angeles, CA 90095, USA
Genome Biology 2011, **12(Suppl 1):**P21

Background Most studies exploring cancer progression have focused on the influence of individual genes, and few efforts have investigated the effects of interactions between genes within the genome. Our hypothesis is that cancer cells thrive by exploiting combinations of genes, in fact by exploiting networks of genes that both protect the cell against destruction and enhance its survival. We believe that these networks involve genes that tend to be coordinated in their copy number alterations, even when they are located at a distance in the genome. Radiation hybrid (RH) cells have a random assortment of genes as triploid rather than diploid. Our recent work studying genetic networks in libraries of RH cells has elucidated key survival-enhancing interactions with high specificity [1]. Because of the hardiness of the RH clones, statistically significant patterns of co-inherited, unlinked triploid gene pairs pointed to the cell survival mechanism. We identified more than 7.2 million significant interactions at single-gene resolution using the RH data.
Methods Our work with the RH data provided the rationale for an investigation of cancer survival networks, in particular for glioblastoma multiforme, a formidable brain cancer for which extensive datasets are available but few treatment options. We investigated correlated patterns of copy number alterations for distant genes in glioblastoma multiforme tumors using the same method we employed to construct the RH survival network. Public data were analyzed from 301 glioblastomas that had been assessed for copy number alterations using array comparative genomic hybridization [2].
Results The glioblastoma and RH survival networks overlapped significantly ($P = 3.7 \times 10^{-31}$). We therefore exploited the high-resolution mapping of the RH data to obtain single-gene specificity in the glioblastoma network. The combined network features 5,439 genes and 13,846 interactions (false discovery rate (FDR) <5%) and suggests novel approaches to therapy for glioblastoma. For example, although the epidermal growth-factor receptor (*EGFR*) oncogene is frequently activated in glioblastoma, EGFR inhibitors have limited therapeutic efficacy [3]. In the combined glioblastoma survival network, there are 46 genes that interact with EGFR, of which ten (22%) happen to be targets of existing drugs. This observation suggests that a flanking attack strategy that strikes at both EGFR and its partner genes in the glioblastoma survival network may be an effective approach to treating these tumors.
Conclusions By elucidating a genetic survival network for glioblastoma, we gained insight into the mechanisms of proliferation of this cancer and opened up new avenues for therapeutic intervention.

References
1. Lin A, Wang RT, Ahn S, Park CC, Smith DJ: **A genome-wide map of human genetic interactions inferred from radiation hybrid genotypes.** *Genome Res* 2010, **20:**1122-1132.
2. The Cancer Genome Atlas Research Network: **Comprehensive genomic characterization defines human glioblastoma genes and core pathways.** *Nature* 2008, **455:**1061-1068.
3. Lo HW: **EGFR-targeted therapy in malignant glioma: novel aspects and mechanisms of drug resistance.** *Curr Mol Pharmacol* 2010, **3:**37-52.

P22

Building a protein interaction (PIN) database for autism candidate genes

Catherine Croft Swanwick[1] and Sharmila Banerjee-Basu[1]
[1]MindSpec, McLean, VA 22181, USA
Genome Biology 2011, **12(Suppl 1):**P22

Background Hundreds of diverse genetic loci have been linked to autism spectrum disorders (ASDs), making large-scale analysis essential for understanding the molecular events underlying the pathogenesis of these disorders. Our laboratory first released the autism database AutDB in 2007 as a bioinformatics tool for systematic curation of all known ASD candidate genes [1-3]. AutDB was designed with a systems biology approach, integrating genetic entries within the Human Gene module with corresponding behavioral, anatomical and physiological data in the Animal Model module. In June 2011, we released a new Protein Interaction (PIN) module of AutDB, which serves as a comprehensive, up-to-date resource on the direct protein interactions of ASD-linked genes.
Methods To curate the PIN module, our researchers utilize a multi-level annotation model to systematically search, collect and extract information entirely from published, peer-reviewed scientific literature. Although we initially consult public molecular interaction databases (HPRD and BioGRID) and commercial molecular interaction software (Pathway Studio, version 7.1), every interaction is manually extracted and verified by evaluating the primary reference articles from PubMed. Our manual curation has proved critical for accurate annotation, because these references were the second largest source of references for the initial PIN dataset, providing more interactions than both HPRD and Pathway Studio. Each ASD gene entry within the PIN module is presented as a multi-level display, with interactive graphical and tabular views of its corresponding interactome.
Results The initial PIN dataset includes interactomes for 86 ASD candidate genes, with a total of 1,311 direct protein interactions garnered from 533 unique primary references. These interactomes are composed of 6 interaction types and 13 species, documented by 402 distinct pieces of evidence. Our researchers will expand and maintain the data content of the PIN module with systematic updates.
Conclusions We have created an integrated bioinformatics tool that can be used for the large-scale analysis of the biological relationships among ASD candidate genes. Such network analysis is envisioned to provide a framework for identifying the key molecular pathways underlying ASD pathogenesis, potentially leading to the development of novel drug therapies.
References
1. Basu SN, Kollu R, Banerjee-Basu S: **AutDB: a gene reference resource for autism research.** *Nucleic Acids Res* 2009, **37:**D832-D836.
2. Kumar A, Wadhawan R, Swanwick CC, Kollu R, Basu SN, Banerjee-Basu S: **Animal model integration to AutDB, a genetic database for autism.** *BMC Med Genomics* 2011, **4:**15.
3. **The AutDB Resource** [http://www.mindspec.org/autdb.html]

P23

Whole transcriptome sequencing of normal and tumor bladder tissue samples

Wei Tang[1] and Ludmila Prokunina-Olsson[1]
[1]Laboratory of Translational Genomics, Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Bethesda, MD 20892, USA
Genome Biology 2011, **12(Suppl 1):**P23

Background Bladder cancer is the 9th most common cancer worldwide and the 13th most common cancer-related cause of death. Bladder cancer frequently recurs after the removal of primary carcinomas. This recurrence

leads to repeated surgeries and long-term treatment and surveillance, making it the most expensive type of cancer to treat. Genetic factors and environmental factors such as cigarette smoking and occupational exposure to aromatic amines are linked to bladder cancer risk. Genome-wide association studies (GWAS) for bladder cancer have identified multiple genetic variants within genes and regions, including *TP63*, *TERT-CLPTMIL* and 8q24.21, to be highly associated with disease risk. Whole transcriptome sequencing (RNA-Seq) is a revolutionary tool for generating a large amount of qualitative and quantitative information, thus helping to explore known and novel transcripts, splicing forms and fusion genes.

Methods To understand the genetic and genomic landscape of the GWAS susceptibility regions, we investigated and characterized the entire transcriptome of normal and tumor bladder tissue samples by using powerful massively parallel RNA sequencing. We used an Illumina HiSeq 2000 instrument to sequence six paired samples of normal and tumor bladder tissues. For each of the samples, we generated 50 Gb of 100-bp reads to represent the whole transcriptome.

Results Using the Bowtie/TopHat and Samtools packages, we successfully aligned approximately 80% of the total sequence reads against the human genome reference sequence (build 19). Our analysis sought to identify alternative splicing forms, novel exons, non-coding transcripts and chimeric fusion events. Total levels of mRNA in normal and tumor samples were evaluated by Cufflinks analysis based on the Ensembl transcripts database. Multiple splicing isoforms were identified for some of the GWAS susceptibility genes, and some of these isoforms were differentially expressed between the tumor and normal samples. We found that novel transcripts and non-coding RNAs corresponding to gene desert regions such as 8q24 were abundantly expressed. Our next step will focus on validation of these differentially expressed genes and novel transcripts by using quantitative RT-PCR on independent samples.

Conclusions Using RNA-Seq, we explored transcripts corresponding to candidate regions identified by bladder cancer GWAS. Some of these transcripts demonstrated splicing variability and differential levels of expression between normal and tumor tissue samples, which might be of importance for bladder cancer.

P24

Identification of functional genetic variants associated with prostate cancer through analysis of genome-wide genetic and epigenetic datasets

McAnthony Tarway[1], Wei Tang[1] and Ludmila Prokunina-Olsson[1]
[1]Laboratory of Translational Genomics, Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Bethesda, MD 20892, USA
Genome Biology 2011, **12(Suppl 1):**P24

Background Recent genome-wide association studies (GWAS) have identified multiple genetic variants associated with the risk of developing prostate cancer (PrCa). At least ten PrCa-associated single nucleotide polymorphisms (SNPs) are located within a gene-poor region on chromosome 8q24, but the functional mechanisms of each of these variants remain unknown. Normal prostate development, as well as tumor initiation and progression, greatly depends on the androgen receptor (AR) and its ligands, testosterone and 5α-dihydrotestosterone. We hypothesized that genetic variants associated with PrCa risk might be important owing to their effects on AR-binding sites.

Methods and results We comprehensively explored 11 PrCa GWAS published as of July 2011 in the National Human Genome Research Institute's GWAS database [1] and in PubMed [2]. We selected ten SNPs from the 8q24 region that were significantly and consistently associated with PrCa in Caucasian datasets ($P < 5 \times 10^{-7}$). By querying the CEU 1000 Genomes Project panel, we generated a list of 224 SNPs in high linkage disequilibrium ($r^2 > 0.8$) with the ten selected GWAS SNPs. Of all of the SNPs on this list, six variants were located in the regions identified as AR-binding sites, based on AR chromatin immunoprecipitation (ChIP)-Seq data from the University of California, Santa Cruz's genome browser [3]. To test for differential binding of AR to alleles of the six SNPs, we developed a protocol for quantitative multiplex allele-specific ChIP (AS-ChIP) assays. Confirmatory AS-ChIP with AR-specific antibodies in the LNCaP cell line showed that five of these SNPs were heterozygous in the LNCaP cell line, and four of them showed statistically significant allele-specific differences in AR binding (*P*-value range = 0.0005 to 0.04, based on four biological replicates of AS-ChIP).

Conclusions Our data suggest that some of the PrCa-associated SNPs within the 8q24 region might create or disrupt binding sites for AR, thereby affecting important regulatory networks in normal and cancerous prostate tissue.

References

1. A Catalog of Published Genome-Wide Association Studies [http://www.genome.gov/gwastudies/]
2. PubMed [http://www.ncbi.nlm.nih.gov/pubmed]
3. University of California, Santa Cruz Genome Browser [http://genome.ucsc.edu/]

P25

MetAMOS: a metagenomic assembly and analysis pipeline for AMOS

Todd J Treangen[1,2], Sergey Koren[1,3], Irina Astrovskaya[1], Dan Sommer[1], Bo Liu[1,3] and Mihai Pop[1,3]
[1]Center for Bioinformatics and Computational Biology, University of Maryland, College Park, MD 20742, USA; [2]The McKusick-Nathans Institute for Genetic Medicine, The Johns Hopkins University School of Medicine, Baltimore, MD 21205, USA; [3]Department of Computer Science, University of Maryland, College Park, MD 20742, USA
Genome Biology 2011, **12(Suppl 1):**P25

Background Metagenomics has opened the door to unprecedented comparative and ecological studies of microbial communities, ranging from the sea [1] to the soil (the terragenome) to within the human body [2,3]. Most analyses begin with assembly, as the short reads that are characteristic of most datasets severely limit the ability to classify the data taxonomically [4-7] and require considerable computational resources to perform comparative analyses (such as BLAST against public databases). In addition, given that many sequences are likely to be from novel organisms, classification methods relying on databases fail to acknowledge most of the novel species present in the dataset. In an attempt to move away from reference-based analysis, computational tools based on promising algorithmic and statistical methods for metagenomic *de novo* assembly have recently started to emerge [8,9]. However, to date, they either are ill-suited to large datasets or have yet to offer significant improvements over existing genome assemblers that were not designed for metagenomic assembly.

Methods Here, we describe MetAMOS [10], an open-source, modular assembly pipeline built upon AMOS and tailored specifically for metagenomic next-generation sequencing data. MetAMOS is the first step toward a fully automated assembly and analysis pipeline, from mated reads (Illumina and 454) to scaffolds and ORFs. Currently, MetAMOS has support for four assemblers (SOAPdenovo [11], Newbler, CABOG and Minimus [12]), three annotation methods (BLAST, PhymmBL and MetaPhyler), two metagenomic gene prediction tools (MetaGeneMark and Glimmer-MG) and one unitig scaffolder engineered specifically for metagenomic data (Bambus 2). We also provide a novel graph-based algorithm to propagate annotations rapidly to all contigs in an assembly using, for example, only the largest contigs or contigs with high-confidence classification. MetAMOS has three principal outputs: subdirectories containing FASTA sequence of the contigs/scaffolds/variant motifs belonging to a specified taxonomic level, a collection of all unclassified/potentially novel contigs contained in the assembly, and an HTML report with detailed assembly statistics and summary charts.

Results and conclusions We compared MetAMOS with other metagenomic assembly tools (Meta-IDBA and Genovo) and with genome assemblers that have previously been used with metagenomic data (CA-met and SOAPdenovo). We used both a mock/artificial dataset generated for the Human Microbiome Project (HMP) project and real metagenomic samples from the HMP and its European counterpart (MetaHIT). On the mock dataset, MetAMOS compares favorably to existing metagenomic and genomic assemblers with respect to several validation metrics that take into account contig accuracy in addition to size. On the real dataset, MetAMOS also outperforms the existing software. These improvements can largely be attributed to heavy reliance on Bambus 2 and to assembly verification techniques that help identify and remove potentially chimeric contigs while running the pipeline.

In terms of biology, we were able to report several novel variant motifs that would be challenging at best to identify and extract from the output of other methods. In addition, much emphasis was placed on making MetAMOS compatible with a variety of next-generation sequencing technologies, genome assemblers and annotation methods, making the pipeline highly customizable for the beginner and advanced bioinformatics user alike.

References
1. Yooseph S, Sutton G, Rusch DB, Halpern AL, Williamson SJ, Remington K, Eisen JA, Heidelberg KB, Manning G, Li W, Jaroszewski L, Cieplak P, Miller CS, Li H, Mashiyama ST, Joachimiak MP, van Belle C, Chandonia JM, Soergel DA, Zhai Y, Natarajan K, Lee S, Raphael BJ, Bafna V, Friedman R, Brenner SE, Godzik A, Eisenberg D, Dixon JE, Taylor SS *et al.*: **The *Sorcerer II* Global Ocean Sampling expedition: expanding the universe of protein families.** *PLoS Biol* 2007, **5:**e16.
2. Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, Manichanh C, Nielsen T, Pons N, Levenez F, Yamada T, Mende DR, Li J, Xu J, Li S, Li D, Cao J, Wang B, Liang H, Zheng H, Xie Y, Tap J, Lepage P, Bertalan M, Batto JM, Hansen T, Le Paslier D, Linneberg A, Nielsen HB, Pelletier E, Renault P *et al.*: **A human gut microbial gene catalogue established by metagenomic sequencing.** *Nature* 2010, **464:**59-65.
3. Turnbaugh PJ, Ley RE, Hamady M, Fraser-Liggett CM, Knight R, Gordon JI: **The Human Microbiome Project.** *Nature* 2007, **449:**804-810.
4. Brady A, Salzberg, SL: **Phymm and PhymmBL: metagenomic phylogenetic classification with interpolated Markov models.** *Nat Methods* 2009, **6:**673-676.
5. McHardy AC, Martin HG, Tsirigos A, Hugenholtz P, Rigoutsos I: **Accurate phylogenetic classification of variable-length DNA fragments.** *Nat Methods* 2007, **4:**63-72.
6. Nalbantoglu OU, Way SF, Hinrichs SH, Sayood K: **RAIphy: phylogenetic classification of metagenomics samples using iterative refinement of relative abundance index profiles.** *BMC Bioinformatics* 2011, **12:**41.
7. Patil KR, Haider P, Pope PB, Turnbaugh PJ, Morrison M, Scheffer T, McHardy AC: **Taxonomic metagenome sequence assignment with structured output models.** *Nat Methods* 2011, **8:**191-192.
8. Laserson J, Jojic V, Koller, D: **Genovo: *de novo* assembly for metagenomes.** *J Comput Biol* 2011, **18:**429-443.
9. Peng Y, Leung HC, Yiu SM, Chin FY: **Meta-IDBA: a *de novo* assembler for metagenomic data.** *Bioinformatics* 2011, **27:**i94-i101.
10. **MetAMOS Source Code** [https://github.com/treangen/metAMOS]
11. Li Y, Hu Y, Bolund L, Wang J: **State of the art *de novo* assembly of human genomes from massively parallel sequencing data.** *Hum Genomics* 2010, **4:**271-277.
12. Sommer DD, Delcher AL, Salzberg SL, Pop M: **Minimus: a fast, lightweight genome assembler.** *BMC Bioinformatics* 2007, **8**:64.

P26

The mutation spectrum in Indian patients with Gaucher disease

Vartika Bisariya[1], Pramod K Mistry[2], Jun Liu[2], Madhumita Roy Chaudhari[1], Neerja Gupta[1] and Madhulika Kabra[1]
[1]Genetics Division, Department of Pediatrics, All India Institute of Medical Sciences, New Delhi, India, [2]Yale University School of Medicine, New Haven, USA
Genome Biology 2011, **12(Suppl 1):**P26

Gaucher disease is the most common lysosomal storage disorder. It results from an inherited deficiency of the enzyme glucocerebrosidase (GBA); accumulation of the substrate of this enzyme has many clinical manifestations. Since the discovery of the *GBA* gene, more than 200 mutations have been identified, but only a handful of mutations are recurrent (L444P, N370S, IVS2, D409H and 55Del). To determine the spectrum of mutations in the Indian population, we performed mutational screening in children with Gaucher disease.

Twenty-four patients from twenty families were enrolled in this study, after written informed consent was obtained. The diagnosis of Gaucher disease was based on mandatory clinical and biochemical analysis. An initial screening for five common mutations was carried out using PCR-RFLP. Patients who were negative for common mutations were screened by sequencing exons 9 to 11 (a mutation hotspot region) [1].

We identified common mutations (L444P, N370S, IVS2 and D409H [2], and 55Del [3]) in approximately 50% of the patients. L444P (c.1448T>C) was the most frequently identified, followed by D409H in our patients. Western data shows that N370S is the most common mutation in Romanian patients [4]. One polymorphism (E340K) was identified in two patients who were compound heterozygotes for A456P/R463C and S237F/A269P, respectively. Our data highlight the spectrum of mutations that lead to Gaucher disease in the Indian population.

References
1. Hatton CE, Cooper A, Whitehouse C, Wraith JE: **Mutation analysis in 46 British and Irish patients with Gaucher's disease.** *Arch Dis Child* 1997, **77:**17-22.
2. Rozenberg R, Araujo FT, Fox DC, Aranda P, Nonino A, Micheletti C, Martins AM, Cravo R, Sobreira E, Pereisa LV: **High frequency of mutation G377S in Brazilian type 3 Gaucher disease patients.** *Braz J Med Biol Res* 2006, **39:**1171-1179.
3. Tayebi N, Stern H, Dymarskaia I, Herman J and Sidransky E: **55-Base pair deletion in certain patients with Gaucher disease complicates screening for common Gaucher alleles.** *Am J Med Genet* 1996, **66:**316-319.
4. Drugan C, Procopciuc L, Jebeleanu G, Grigorescu-Sido P, Dussau J, Poenaru L, Caillaud C: **Gaucher disease in Romanian patients: incidence of the most common mutations and phenotypic manifestations.** *Eur J Hum Genet* 2002, **10:**511-515.

P27

Boundary distinction interpretation of microarray data via discrete correlate summation

Brian M Westwood[1], Amy L Olex[2], James L Norris[2], Leslie B Poole[3] and Jacquelyn S Fetrow[2,4]
[1]Department of Molecular Genetics and Genomics, Wake Forest University School of Medicine, Winston-Salem, NC 27157, USA; [2]Department of Computer Science, Wake Forest University, Winston-Salem, NC 27109, USA; [3]Department of Biochemistry, Wake Forest University School of Medicine, Winston-Salem, NC, 27157, USA; [4]Department of Physics, Wake Forest University, Winston-Salem, NC 27109, USA
Genome Biology 2011, **12(Suppl 1):**P27

Background Given differential gene expression data across divergent mutant strain arrays of two enzyme subgroups, it would be logical to segregate by protein group ablation (PGA). Discrete correlate summation (DCΣ) was utilized to examine the differential effects of a hydrogen peroxide stressor on discrete and total yeast knockouts of the genes encoding glutathione peroxidase (Gpx) and peroxiredoxin (Prx), both groups starting from the wild-type (WT) strain [1]. While the half-life of the total Gpx knockout mutant is intermediate between that of the WT and the transient total Prx knockout mutant, the distribution of passage number of the various mutant strains can be separated into two groups independent of Gpx and Prx state. Based on half-viability, totalPrx <<<< nPrx << Gpx3 = Tsa1 < totalGpx < mPrx <<< Gpx1 < Gpx2 << Ahp1 = WT <<< Tsa2 ($P < 0.0005$, two tailed *t*-test, $n = 5$, 6). DCΣ was also employed for the boundary between robust and gracile cultures. The aim of this study was to find the characteristic response of the transcriptome, from the perspective of PGA versus strain viability (SV).

Methods DCΣ is a method used to score variables that can be classified into two groups [2]. It is a composite score of a gene's mean group change and overall interaction difference relative to all others tested. Transcripts were included in this analysis only if the values for all conditions passed microarray quality control and were present in the Kyoto Encyclopedia of Genes and Genomes (KEGG) network [3]. Randomly sorted edges were sampled for comparison ($P < 0.001$, two tailed *t*-test, $n = 8{,}372$). Edges that were sorted on average DCΣ score and grouped by biological process yielded a distinctive topology ($P < 1e{-}85$, two tailed *t*-test, $n = 8{,}372$). The identified transcripts were subjected to functional annotation in the Database for Annotation, Visualization and Integrated Discovery (DAVID) [4].

Results Application of DCΣ to the individual and complete knockouts of Gpx (3 genes) and Prx (5 genes) identified 92 transcripts based on PGA and 43 based on SV, with a 13 gene overlap (corresponding to the proteins Arg1p, Aah1p, Ade17p, Pgm2p, Cat2p, Cdd1p, Mae1p, Arg3p, Nma2p, Ole1p, Cta1p, Spb1p and Cds1p). Functional annotation analysis of the 92 PGA transcripts identified the following functions: pyrimidine metabolism, steroid biosynthesis, purine metabolism, RNA polymerase and terpenoid backbone biosynthesis. Ergosterol biosynthesis, gluconeogenesis and transcription from Pol I/III promoters were major biological process categories for this set. Interestingly, terpenoids feed into the steroid pathway, which results in the vitamin D2 precursor ergosterol. Analysis of the 43 SV transcripts identified starch and sucrose metabolism, butanoate metabolism, and fructose and mannose metabolism. Stress response was the key biological process for this arm of the study. No functional annotations were statistically significant for the common genes. Transcripts identified by PGA of either the Gpx- or Prx-encoding genes tend toward transcriptional control mechanisms, whereas SV-associated transcripts track with metabolic necessities.

References
1. Fomenko DE, Koc A, Agisheva N, Jacobsen M, Kaya A, Malinouski M, Rutherford JC, Siu KL, Jin DY, Winge DR, Gladyshev VN: **Thiol peroxidases mediate specific genome-wide regulation of gene expression in response to hydrogen peroxide.** *Proc Natl Acad Sci USA* 2011, **108:**2729-2734.

2. Westwood, B, Chappell, M: **Application of correlate summation to data clustering in the estrogen- and salt-sensitive female mRen2.Lewis rat.** In *Proceedings of the First International Workshop on Text Mining in Bioinformatics: November 10 2006; Arlington*. Association for Computing Machinery; 2006:21-26.
3. **The Kyoto Encyclopedia of Genes and Genomes** [http://www.genome.jp/kegg/]
4. **The Database for Annotation, Visualization and Integrated Discovery** [http://david.abcc.ncifcrf.gov/]

P28

A computational approach to identify transposable element insertions in cancer cells

Israel T Silva[1,2], Daniel G Pinheiro[1] and Wilson A Silva Jr[1,3]
[1]Regional Blood Center of Ribeirão Preto, Molecular Biology and Bioinformatics Laboratory, Ribeirão Preto, São Paulo 14051-140, Brazil; [2]Barão de Mauá University, Ribeirão Preto, São Paulo 14026-150, Brazil; [3]Department of Genetics, Medical School of Ribeirão Preto, University of São Paulo, Ribeirão Preto, São Paulo 14049-900, Brazil
Genome Biology 2011, **12(Suppl 1):**P28

Background Transposable elements (TEs) in the human genome may contribute to molecular evolution, hereditary diseases and cancer [1-3]. Therefore, analyzing the impact of TEs in the genome is necessary to better characterize genetic events related to tumorigenesis. Here, we used a computational approach to identify TE insertions in publicly available data for exome sequences in lymphoblastoid and breast tumor cells derived from the same patient.
Methods A total of 29,340, sequences from the cell lines HCC1954 (18,365,271) and HCC1954BL (10,975,107) were used to investigate gene fusion with TEs (gfTEs) [4,5]. The RepeatMasker and Burrows-Wheeler Alignment (BWA) tools were used to identify and to map gfTEs, respectively. We also used BEDTools to find overlaps between gfTEs and genome annotations. Human mRNAs and RepeatMasker tracks were downloaded in BED format from the GRCh37/hg19 assembly. Repbase was used to filter the eukaryotic TEs.
Results RepeatMasker was used to identify gfTEs in the exome reads. Next, the repeat masked reads were aligned against the reference genome using BWA. Finally, we filtered the aligned reads to exclude those without TEs (length of Ns <15, Ns means block of nucleotides masked), those with alignments showing low sequence identity (<95%) or those with a small hit length (<50 nucleotides). The study focused on the detection of TEs in coding sequence gene regions. A total of 3,307,608 reads were excluded, and 23,841 reads were predicted as cancer-specific gfTEs. Table 1 shows the number of gfTEs distributed among the TE families and highlights the members with higher frequency in both cell lines. Insertions of LINE/L1 and SINE/Alu were the most frequent. The Gene Ontology analysis for the biological process and molecular function terms showed a bias toward membrane receptor and cell adhesion proteins.
Conclusions We used a computational approach to identify putative cancer-specific gfTEs using human exome capture sequences. Interestingly, the total number of gfTEs was similar in normal and tumor cell lines, but the Gene Ontology analysis revealed an enrichment of insertions in genes encoding protein receptors and cell adhesion molecules. These results suggest that TEs could be contributing to cancer development.

References

1. Cordaux R, Batzer MA: **The impact of retrotransposons on human genome evolution.** *Nat Rev Genet* 2009, **10:**691-703.
2. Callinan PA, Batzer MA: **Retrotransposable elements and human disease**. *Genome Dyn* 2006, **1:**104-115.
3. Zhang W, Edwards A, Fan W, Deininger P, Zhang K: **Alu distribution and mutation types of cancer genes.** *BMC Genomics* 2011, **12:**157.
4. Zhao Q, Kirkness EF, Caballero OL, Galante PA, Parmigiani RB, Edsall L, Kuan S, Ye Z, Levy S, Vasconcelos AT, Ren B, de Souza SJ, Camargo AA, Simpson AJ, Strausberg RL: **Systematic detection of putative tumor suppressor genes through the combined use of exome and transcriptome sequencing.** *Genome Biol* 2010, **11:**R114.
5. Galante PA, Parmigiani RB, Zhao Q, Caballero OL, de Souza JE, Navarro FC, Gerber AL, Nicolás MF, Salim AC, Silva AP, Edsall L, Devalle S, Almeida LG, Ye Z, Kuan S, Pinheiro DG, Tojal I, Pedigoni RG, de Sousa RG, Oliveira TY, de Paula MG, Ohno-Machado L, Kirkness EF, Levy S, da Silva WA Jr, Vasconcelos AT, Ren B, Zago MA, Strausberg RL, Simpson AJ *et al.*: **Distinct patterns of somatic alterations in a lymphoblastoid and a tumor genome derived from the same individual**. *Nucleic Acids Res* 2011. doi: 10.1093/nar/gkr221.

Table 1 (abstract P28). Number of genes containing insertion of TEs from different families

Class/Family	HCC1954BL (N)	HCC1954 (T)
DNA	4	3
DNA/MuDR	5	1
DNA/PiggyBac	2	2
DNA/TcMar-Mariner	10	9
DNA/TcMar-Tc2	6	8
DNA/TcMar-Tigger	90	96
DNA/hAT	2	8
DNA/hAT-Blackjack	7	19
DNA/hAT-Charlie	107	137
DNA/hAT-Tip100	12	19
LINE/CR1	23	25
LINE/Dong-R4	1	1
LINE/L1	863	641
LINE/L2	163	175
LINE/RTE	9	13
LINE/RTE-BovB	1	0
LTR	1	2
LTR/ERV1	134	145
LTR/ERVK	11	17
LTR/ERVL	70	77
LTR/ERVL-MaLR	148	186
LTR/Gypsy	6	7
Other	5	4
RNA	1	3
SINE	6	17
SINE/Alu	264	406
SINE/Deu	5	14
SINE/MIR	109	149
SINE/tRNA	0	3
Satellite	7	15
Satellite/acro	2	1
Satellite/centr	52	112
Unknown	6	8
rRNA	14	11
scRNA	4	2
snRNA	0	2
srpRNA	4	1
tRNA	0	1
Total	2.154	2.340

P29

Joint analysis of genome-wide genetic variants associated with gene expression and disease susceptibility

Chen-Hsin Yu[1,2] and John Moult[1,3]
[1]Institute for Bioscience and Biotechnology Research, University of Maryland, Rockville, MD, USA; [2]Molecular and Cellular Biology Program, University of Maryland, College Park, MD, USA; [3]Department of Cell Biology and Molecular Genetics, University of Maryland, College Park, MD, USA
Genome Biology 2011, **12(Suppl 1):**P29

Background Genome-wide association studies (GWAS) of human complex disease have identified a large number of disease-associated genetic loci, which are distinguished by distinctive frequencies of specific single nucleotide polymorphisms (SNPs) in 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. Many studies have shown that variations in gene expression among individuals, as well as among cell types, contribute to phenotype diversity and disease susceptibility. Recent genome-wide expression quantitative trait loci (eQTL) association (GWEA) studies have provided information on genetic factors, especially SNPs, that are associated with gene expression variation. These expression-associated SNPs (exSNPs) have already been utilized to explain some results of GWAS for diseases, but interpretation of the data is handicapped by low reproducibility of the genotype-expression relationships.

Methods To address this problem, we established several gold standard sets of high-reliability exSNPs based on multiple occurrences in different GWEA studies in various human populations and cell types. We then related these data to results from GWAS for diseases, to find a set of disease-associated loci that are likely to have an underlying expression mechanism. HapMap linkage disequilibrium data were utilized to allow the comparison of GWEA results from studies that employed different microarray SNP sets.

Results We integrated the current gold standard data with SNPs in disease-associated loci from the Wellcome Trust Case-Control Consortium (WTCCC) GWAS of seven common human diseases. Approximately one-third of these disease-associated loci in the WTCCC GWAS were found to be consistent with an underlying expression change mechanism. Comparing separate gold standard sets for Caucasian (CEU), African (YRI) and Asian (ASN) populations also allowed us to investigate which exSNPs contribute to population-specific eQTLs.

Conclusions Use of the gold standard set of SNP-expression relationships has enabled us to more reliably determine the role of expression changes in common human diseases.

P30

Phylogenomics of prokaryotic ribosomal proteins

Natalya Yutin[1], Kira S Makarova[1], Yuri I Wolf[1], Eugene V Koonin[1]
[1]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA
Genome Biology 2011, **12(Suppl 1):**P30

Background Archaeal and bacterial ribosomes contain more than 50 proteins. Thirty-four ribosomal proteins (r-proteins) are universally conserved in the three domains of cellular life (Bacteria, Archaea and Eukarya), and 33 r-proteins are shared between Archaea and Eukarya to the exclusion of Bacteria; there are also 23 Bacteria-specific, 1 Archaea-specific and 11 Eukarya-specific r-proteins [1]. Despite the high sequence conservation of r-proteins, the annotation of r-protein genes is often difficult because of their short lengths and biased sequence composition.

Methods To perform a comprehensive survey of prokaryotic r-proteins, we developed an automated computational pipeline for the identification of r-protein genes and applied it to 995 completely sequenced bacterial genomes and 87 archaeal genomes available in the RefSeq database. The pipeline employs curated seed alignments of r-proteins to run position-specific scoring matrix (PSSM)-based BLAST searches against six-frame genome translations, thus overcoming possible gene annotation errors. Likely false positives are identified using comparisons against the original seed alignments.

Results In the course of this analysis, we gained insight into the diversity of prokaryotic r-protein complements, such as missing and paralogous r-proteins and distributions of r-protein genes among chromosomal partitions. A phylogenetic tree was constructed from a concatenated alignment of 50 almost-ubiquitous bacterial r-proteins. The topology of the tree is generally compatible with the current high-level bacterial taxonomy, although we detected several inconsistencies, possibly indicating uncertain or erroneous classification of the respective bacteria. Similarly, a concatenated alignment of 57 ubiquitous archaeal proteins was used for an archaeal phylogenetic tree reconstruction. In both Bacteria and Archaea, the patterns of the presence/absence of non-ubiquitous r-proteins suggest several independent losses and/or gains of these proteins. According to parsimony reconstruction, three bacterial and five archaeal r-proteins do not appear to be ancestral. Remarkably, all five non-ancestral archaeal r-proteins are present in Eukarya.

Conclusions Extended sets of prokaryotic r-proteins were created. Alignments of these sets may be used as new seed profiles for the identification of r-proteins in new genomes and for comparative genomics studies.

Reference

1. Lecompte O, Ripp R, Thierry JC, Moras D, Poch O: Comparative analysis of ribosomal proteins in complete genomes: an example of reductive evolution at the domain scale. *Nucleic Acids Res* 2002, **30:**5382-5390.

P31

Genomes in a bottle: creating standard reference materials for genomic variation - why, what and how?

Justin M. Zook[1] and Marc Salit[1]
[1]Biochemical Science Division, National Institute of Standards and Technology, Gaithersburg, MD 20899, USA
Genome Biology 2011, **12(Suppl 1):**P31

Broad clinical application of ultra-high-throughput sequencing is imminent. In a few notable cases, actionable information has been discovered from sequencing, and the number of such cases is likely to increase. At present, there are no widely accepted genomic standards or quantitative performance metrics. These are needed to achieve the confidence in measurement results that is expected for sound, reproducible research and regulated applications. The National Institute of Standards and Technology (NIST) has been approached about considering development in this area by several commercial entities and regulatory agencies. There is great enthusiasm for translation of sequencing from the research community to clinical practice, and standards that can be used to inform confidence in measurement results (for instance, through validation studies, proficiency testing and routine quality assurance) may be an enabling factor in that goal.

NIST is currently gathering input from the genomics community about which reference materials and data would be useful. For example, NIST and the Coriell Institute for Medical Research may develop genomic reference material from cell lines from families that have already been characterized by a variety of sequencing methods (for example, the cell line from which NA12878 DNA is derived). In addition, we may build synthetic DNA constructs to test specific questions about measuring different types of variants or combinations of variants in different genomic contexts. For example, we might create pairs of constructs with single nucleotide polymorphisms, indels and/or structural variants in GC- or AT-rich regions or repeat regions. To ensure the design of appropriate standards, we are interested in discussing the design and application of genomic reference materials with any interested parties.

P32

Deciphering the reproductive protein-protein interaction network in *Anopheles gambiae* with *Drosophila melanogaster* as a framework

Daniel Achinko[1], Paul Mireji[2], Flaminia Catteruccia[3] and Dan Masiga[1]
[1]Molecular Biology and Biotechnology Department, International Center of Insect Physiology and Ecology (icipe), Nairobi 30772-00100, Kenya; [2]Biochemistry and Molecular Biology Department, Egerton University, Njoro, Kenya; [3]Division of Cell and Molecular Biology, Imperial College London, London SW7 2AZ, UK
Genome Biology 2011, **12(Suppl 1):**P32

Background Protein-protein interactions (PPIs) are the most fundamental biological processes at the molecular level. The experimental methods for testing PPIs are time-consuming and are limited by analogs for many reactions. As a result, a computational model is necessary to predict PPIs and to explore the consequences of signal alterations in biological pathways. Reproductive control of the vector *Anopheles gambiae* using transgenic techniques poses a serious challenge. To meet this challenge, it would help to define the biological network involving the male accessory gland (MAG) proteins responsible for successful formation of the mating plug [1]. This plug forms in the male and is transferred to the female during mating, hence initiating the PPIs in both sexes. As is the case in *Drosophila melanogaster*, a close relative of *A. gambiae*, some MAG proteins responsible for the formation of the mating plug have been shown to alter the post-mating behavior of females.

Methods and results The STRING database for known PPIs was used to identify orthologs of *A. gambiae* proteins in *Drosophila* (Table 1). Twenty-seven proteins are known to form the mating plug in *A. gambiae*, and 16 others were obtained as strings in the STRING database. Chromosome synteny comparisons for proteins with more than 50% identity between species were carried out using the Artemis Comparison Tool (ACT version 9.0), and this showed 24.39% matches (M), 12.20% mismatches (MM) and 63.41% unmatched (NM). The network built in Cytoscape (version 2.8.0) with the UniProt IDs for these *Drosophila* orthologs showed 14 complexes, with 4 of them being for *Drosophila*. The network showed 555 nodes and 2,344 edges. The top 50 identified hubs in the network showed a range of 3 to 30 interactions. The expression values for these proteins in FlyAtlas showed that

Table 1 (abstract P32). Orthologs of *Anopheles gambiae* proteins in *Drosophila* identified using the STRING database.

A. gambiae ID (plug proteins)	STRING	Chromosome	Sex	Ortholog in *Drosophila*	UniProt ID	Chromosome	STRING score	Chromosome synteny
AGAP009099		3R	Male	CG7356	Q9VLU2	2L	108	MM
					Q8IPH0	2L	108	MM
AGAP009368		3R	Male	CG15005	Q9VZG4	3L	40	NM
AGAP009370		3R	Male					
AGAP012830		Unknown	Male			Unknown		
AGAP008276		2R	Male	CG12350	Q7JPN9	3R	139	NM
AGAP008277		2R	Male	CG12350	Q7JPN9		137	MM
AGAP013150 (AGAP004671)			Male	CG4738	Q9VKJ3		362	NM
AGAP005791		2L	Male	CG32834	Q9WIW6	2R	74.7	NM
					D3DMG3			
AGAP007041		2L	Male	CG6676	Q95SM8	2R	172	NM
AGAP006418		2L	Male	CG32679	Q8IRL3	X	166	NM
					D9PTU6			
AGAP009673		3R	Male	CG5976	Q7KTY3	3L	317	MM
					Q0GT94			
AGAP003083		2R	Male	CG6113	Q9VKT9	2L	180	NM
AGAP001649		2R	Male	CG31414	Q8IMY3	3R	537	M
				CG3647	Q4V4A3			
					Q4V4J1			
	AGAP0012412	3L		CG6437	Q9W297	2R	583	NM
	AGAP002055	2R		CG3132	Q9VGE7	3R	642	NM
AGAP009584		3R	Both	CG31884	Q9V429	2L	134	NM
	AGAP000565	X		CG2151	P91938	X	687	M
				CG11401	Q9VNT5	3L	687	
	AGAP011107	3L		CG6852	B7Z076	3L	144	NM
					Q9VVT6	3L	144	
	AGAP010517	3L		SOD1	B8YNX4	3L	302	NM
	AGAP007201	2L		TRX-2	Q6HI1	2L	169	NM
	AGAP007827	3R		CG17654	P15007	2L	736	M
	AGAP009623	3R		CG8893	P07487	X	544	M
	AGAP007120	2L		CG2210	P08879	3R	292	NM
	AGAP006818	2L		CG8975	P48592	2R	588	NM
				CG17797	O46197	2L		
	AGAP010198	3R		CG5371	P48591	2L	1311	M
	AGAP001325	2R		CG32920	Q960M4	3R	238	NM
				CG7217	Q6XHE3	3R		
AGAP012407		3L	Both	CG6988	P54399	3L	642	M
	AGAP007393	2L		CG8983	Q3YMU0	2R	696	NM
	AGAP002816	2R		CG1333	Q9V3A6	3L	582	M
AGAP011630		3L	Female	CG33998	Q6IG52	2R	66	NM
					B3DN29	2R	66	
AGAP004533		2R	Both	CG10992	Q9VY87	X	464	M
AGAP005194		2L	Female	CG5255	Q9VEM5	3R	154	NM
AGAP005195		2L	Female	CG4053	Q9VEM7	3R	136	NM
					Q9VIT3	2L		
					Q8IGA0	2L		
AGAP006904		2L	Female	CG4859	Q9W122	2R	762	M
				CG4859	Q8MLN6	2R		
	AGAP003319	2R		CG6281	Q9VH14	3R	148	MM
AGAP007347		2L	Female	CG7798	A1ZAB8	2R	131	NM
AGAP003139		2R	Both	CG18525	Q9VFC2	3R	293	NM
				CG18525	Q9U1I4	3R		
AGAP006964		2L	Female	CG32147	Q8SZB7	3L	155	NM
	AGAP009172	3R		CG5355	Q9VKW5	2L	980	M
AGAP006420		2L	Both	CG32679	Q8IRL3	X	137	NM
AGAP009212		3R	Both	CG7219	A4V9T5	2L	206	NM
NOVEL ACP1 (from female)		3R: b/w 9370 & 9371	Male					
NOVEL ZCP7 (AGAP008071)		3R: b/w 5051000 & 5067900	Male	CG8564	Q9VS63	3L	164	NM

they are upregulated in the reproductive tissues of both sexes. To understand the processes involved in plug formation, the Reactome database was used, and the hub proteins were identified in 49 of the 2,021 known processes in *Drosophila*. Twelve proteins were involved in the following processes: metabolism of proteins (8.8e–13), gene expression (2.0e–06), 3'-UTR-mediated translational regulation (7.7e–08), regulation of β-cell development (1.3e–06), diabetes pathways (6.8e–06), signal recognition (preprolactin) (5.0e–07) and membrane trafficking (1.3e–03). Of the top 50 proteins, 92% had orthologs in *A. gambiae*, with one identified in the mating plug and four others identified as strings to AGAP009584, which is found in the mating plug. Acp29AB was identified in the network and is known to induce post-mating responses in *Drosophila*, confirming that the network is reproductive and giving an insight into the possible pathways involved. The CG9083 (Q8SX59) protein was ranked first among the hub proteins but has no ortholog in *A. gambiae*. Interestingly, it has the same protein properties as the Plugin protein (AGAP009368) in *A. gambiae*, suggesting that Plugin may be the main protein in the PPI reproductive network in *A. gambiae*. The Whelan and Goldman (WAG) maximum likelihood tree evaluations of the plug proteins in *A. gambiae* and their orthologs in *Drosophila* showed that these proteins are involved in similar biological processes in both species, but the *A. gambiae* protein evaluation provided a better explanation for the expected process as it clustered in both pre-mated and post-mated PPIs.

This table shows the 27 proteins known to be in the mating plug of *A. gambiae* [1], derived predominantly from the male. The 16 strings predicted as orthologs in *Drosophila*, using the STRING database, have varying scores. Scores above 60 can be trusted following their alignments. Plugin, which has the lowest score, has no good ortholog in *Drosophila*. Most of the proteins are encoded on chromosome arms 2L and 3R in both species. The chromosome synteny comparisons using ACT showed 24.39% matches (M), 12.20% mismatches (MM) and 63.41% unmatched (NM). The presence of gaps between the alignments resulted in the observed MM and NM. The nucleotide sequences at the chromosomal locations where the proteins NOVEL ACP1 and NOVEL ZCP7 are encoded were used to identify similar proteins and their orthologs.

Conclusions The identification of *A. gambiae* proteins in this network creates more targets for functional analysis and reproductive control of the malaria vector.

Reference

1. Rogers DW, Baldini F, Battaglia F, Panico M, Dell A, Morris HR, Catteruccia F: **Transglutaminase-mediated semen coagulation controls sperm storage in the malaria mosquito.** *PLoS Biol* 2009, **7:**e1000272.

P33

InSilico DB: an online platform to collaboratively structure and export publicly available datasets from the Gene Expression Omnibus database

A Coletta[1], C Molter[1], R Duqué[1], D Steenhoff[2], J Taminau[2], V de Schaetzen[2], C Lazar[2], S Meganck[2], A Nowé[2], H Bersini[1] and D Weiss[1]
[1]IRIDIA, Université libre de Bruxelles, Brussels 1050, Belgium; [2]COMO, Vrije Universiteit Brussel, Brussels 1050, Belgium
Genome Biology 2011, **12(Suppl 1):**P33

There are more than 20,000 genomic studies comprising 500,000 samples freely available in the Gene Expression Omnibus (GEO) database [1]. However, accessing these data requires complex computational steps, including structuring and formatting the clinical vocabulary used to annotate the samples. These complex steps hinder the accessibility of genomic datasets through visualization and analysis software platforms, such as GenePattern and R/Bioconductor, therefore hampering the pace of research.

InSilico DB [2] is an online platform that provides a complete collaborative solution for structuring and formatting clinical annotations from GEO, making GenePattern and R datasets one click away for researchers.

InSilico DB has made available powerful and intuitive online curation tools to structure the metadata of GEO datasets. The database is automatically updated daily, through GEO import pipelines. Datasets can have multiple annotations given by different users, and one user can have multiple versions of an annotation to suit different experimental questions.

The InSilico DB platform supports datasets from Affymetrix human gene expression platforms, which account for 2,900 studies comprising 110,000 samples, making InSilico DB the largest public database of manually curated human gene expression samples. In addition to the web interface, InSilico DB offers programmatic access through an R/Bioconductor package [3].

Future releases of InSilico DB will include Illumina RNA-Seq platform data and Affymetrix mouse gene expression data.

References

1. Galperin MY, Cochrane GR: **The 2011 Nucleic Acids Research Database Issue and the online Molecular Biology Database Collection.** *Nucleic Acids Res* 2011, **39:**D1-D6.
2. **InSilico DB** [http://insilico.ulb.ac.be]
3. **R/Bioconductor package for InSilico DB** [ftp://ftp.questnet.net.au/pub/bioconductor/packages/2.9/bioc/html/ inSilicoDb.html]

P34

Introducing the non-B DNA Motif Search Tool (nBMST)

Regina Z Cer[1], Kevin H Bruce[1], Duncan E Donohue[1], Alpay N Temiz[1], Albino Bacolla[1,2], Uma S Mudunuri[1], Ming Yi, Natalia Volfovsky, Brian T Luke[1], Jack R Collins[1] and Robert M Stephens[1]
[1]Advanced Biomedical Computing Center, Information Systems Program, SAIC-Frederick, National Cancer Institute-Frederick, Frederick, MD 21702, USA; [2]The Dell Pediatric Research Institute, Department of Pharmacy, The University of Texas at Austin, 1400 Barbara Jordan Boulevard, Austin, TX 78723, USA
Genome Biology 2011, **12(Suppl 1):**P34

DNA sequence motifs with the ability to form non-B (non-canonical) structures have been linked to a variety of regulatory and pathological processes. Although the exact mechanism is unknown, recent work has provided significant evidence that non-B DNA structures may play a role in DNA instability and mutagenesis, leading to both DNA rearrangements and increased mutational rates, which are hallmarks of cancer. We have developed algorithms to identify a wide variety of non-B-DNA-forming motifs, including G-quadruplex-forming repeats, direct repeats and slipped motifs, inverted repeats and cruciform motifs, mirror repeats and triplex motifs, and A-phased repeats. After identifying these motifs in the mammalian reference genomes of human, mouse, chimpanzee, macaque, cow, dog, rat and platypus, the data were made publicly available in non-B DB [1]. However, it soon became apparent that it was not feasible to annotate the ever-growing list of genomic data and that it would be more effective to provide researchers with a systematic tool to predict these motifs in their own genomic data. Thus, the non-B DNA Motif Search Tool (nBMST) was created, and it is freely available online [2]. nBMST is a web interface that enables researchers to interactively submit any DNA sequence for searching for non-B DNA motifs. Once a user submits one or more DNA sequences in FASTA format, nBMST returns a comprehensive results page that contains the following: downloadable files in both a tab-delimited format and a generic feature format (GFF); a visualization, including PNG images; and a dynamic genome browser created using the Generic Genome Browser (GBrowse) [3] (version 2.0). Currently, nBMST allows file sizes of up to 20 MB of DNA sequence to be uploaded and stores the results for registered users for up to six months. In summary, the purpose of nBMST is to help provide insight into the involvement of alternative DNA conformations in cancer and other diseases, as well as into other potential biological functions.

References

1. Cer RZ, Bruce KH, Mudunuri US, Yi M, Volfovsky N, Luke BT, Bacolla A, Collins JR, Stephens RM: **Non-B DB: a database of predicted non-B DNA-forming motifs in mammalian genomes.** *Nucleic Acids Res* 2011, **Suppl 1:**D383-D391.
2. **non-B DNA Motif Search Tool** [http://nonb.abcc.ncifcrf.gov/apps/nBMST/]
3. Stein LD, Mungall C, Shu S, Caudy M, Mangone M, Day A, Nickerson E, Stajich JE, Harris TW, Arva A, Lewis S: **The generic genome browser: a building block for a model organism system database.** *Genome Res* 2002, **10:**1599-1610.

P35

Integrating GWAS with gene expression data to dissect the genetic architecture of triple-negative breast cancer

Chindo Hicks, Ranjit Kumar, Antonio Pannuti and Lucio Miele
Cancer Institute, University of Mississippi Medical Center, 2500 North State Street, Jackson, MS 39216, USA.
Genome Biology 2011, **12(Suppl 1):**P35

Background Research focused on genome-wide association studies (GWAS) has resulted in the identification of genetic variants associated with risk of developing breast cancer. These genetic variants are providing valuable insight into the genetic susceptibility landscape of breast cancer. However,

to date, data generated from GWAS have not been maximally leveraged and integrated with gene expression data to identify the genes and pathways associated with the most aggressive subset of breast cancers, triple-negative breast cancer (TNBC), which accounts for about 20% of all breast cancers. TNBC disproportionately affects young premenopausal women and has a higher mortality rate among African-American women. At present, no targeted treatments exist for TNBC, and standard chemotherapy remains the only therapeutic option. Integration of genetic mapping results from GWAS with gene expression data could lead to a better understanding of the genetic mechanisms underlying the molecular basis of the TNBC phenotype and to the identification of potential biomarkers for the development of novel therapeutic strategies.

Methods We mined data from 43 GWAS involving over 250,000 patients with breast cancer and 250,000 controls, reported through April 2011, to identify genetic variants (single nucleotide polymorphisms (SNPs)) and genes associated with risk for breast cancer. We then integrated GWAS information with gene expression data from 305 subjects (162 cases and 143 controls) to stratify TNBC and other breast cancer subtypes, as well as to identify functionally related genes and multi-gene pathways enriched by SNPs that are associated with risk for breast cancer and are relevant to TNBC. To stratify TNBC and to identify functionally related genes, we performed supervised and unsupervised analysis of gene expression data. We used a false discovery rate to correct for multiple testing. Pathway prediction and networking visualization was performed using Ingenuity Systems' software.

Results Combining GWAS information with gene expression data, we identified 448 functionally related genes that stratified breast cancer subtypes into TNBC. A subset of these genes (130 genes) contained SNPs associated with risk for breast cancer; of these 130 genes, 122 correctly stratified TNBC. Pathway prediction revealed multi-gene pathways enriched by SNPs that are significantly associated with risk for breast cancer. Key pathways identified include the p53, nuclear factor-κB, DNA repair and cell cycle regulation pathways.

Conclusions Our results demonstrate that integrating GWAS information with gene expression data can be an effective approach for identifying biological pathways that are relevant to TNBC. These could be potential targets for the development of novel therapeutic strategies.

P36

Abstract not submitted for online publication.

P37

An amalgamated risk estimation model (REM) and assay integration into future REMs

Peter Cartwright[1], Erica Ramos[1], India Bradley[1] and Eric Hanson[1]
[1]Advanced Medical Imaging and Genetics (Amigenics), 5495 South Rainbow Boulevard, Suite 102, Las Vegas, NV 89118, USA
Genome Biology 2011, **12(Suppl 1):**P37

The clinical reality of the post-genomic era is that we now face even more complex disease processes when provided with genomic information, including multifactorial genetic and genomic influences, and epigenetic and environmental factors. A useful example of the promise and perils of genomic technologies and information is breast cancer. By the mid-1990s, two genes (*BRCA1* and *BRCA2*) had been identified, accounting for approximately 5% of affected individuals. Since then, surprisingly few genetic breast cancer risk factors have been identified to account for the remaining 95%. To efficiently and cost-effectively identify individuals at high risk, a combination of information components is required: a patient-reported personal and family medical history; clinical data (for example, a physical exam, pathology results, laboratory test results and imaging); and genetic/genomic results. Gaining comprehensive data from all of these areas provides the best risk assessment and management options for patients. Furthermore, high quality patient and clinical information is essential for the accurate and reliable interpretation of genomic results.

We have clinically implemented a platform that integrates all three informational components with multiple risk estimation models (REMs) to produce an effective automated method for risk-stratifying patients. Although this platform can be and has been applied to a wide range of genetic conditions, this presentation will use breast cancer to illustrate the approach. This system consists of three primary components: a secure web-based questionnaire used by patients to enter personal and family medical history; a tablet-based system for collecting clinical and genomic information; and an analysis engine that seamlessly integrates REMs that have been developed to calculate either a woman's risk of developing breast cancer during her lifetime (Claus, Gail II, BRCAPRO, BOADICEA and IBIS) or the probability of detecting a hereditary breast cancer gene mutation (Myriad, Penn II, BRCAPRO, BOADICEA and IBIS).This use of multiple or amalgamated REM (aREM) results offers one of the most comprehensive breast cancer risk assessments available for predicting the lifetime risk of developing breast cancer or the presence of *BRCA* mutations. Additional uses for aREMs include rapid analyses of existing breast cancer datasets, external validation of new REMs, and prospective outcome comparisons based on initial aREM results.

Numerous biomarkers for breast cancer, in addition to *BRCA1* and *BRCA2* mutations, have been reported, but few molecular markers or assays have been adopted for clinical use. The addition of novel REMs that integrate a new molecular assay or classifiers can facilitate the identification of an enriched population for screening (for example, lowering the number needed to screen) or for diagnostic, prognostic or therapeutic purposes. REMs are rapidly integrating multiple genetic influences, whole genome sequencing data and epigenetic modifications, so structured comparisons of the performance of existing and emerging predictive REMs are required for safe and effective clinical application.

P38

G-CODE: enabling systems medicine through innovative informatics

Subha Madhavan[1], Yuriy Gusev[1], Michael A Harris[1], David M Tanenbaum[2], Robinder Gauba[1], Krithika Bhuvaneshwar[1], Andrew Shinohara[2], Kevin Rosso[2], Lavinia A Carabet[1], Lei Song[1], Rebecca B Riggins[1], Sivanesan Dakshanamurthy[1], Yue Wang[3], Stephen W Byers[1], Robert Clarke[1] and Louis M Weiner[1]
[1]Lombardi Comprehensive Cancer Center, Georgetown University Medical Center, 2115 Wisconsin Avenue NW, Suite 110, Washington DC 20007, USA; [2]ESAC, 155 Gibbs Street, Suite 420, Rockville, MD 20850, USA; [3]Bradley Department of Electrical & Computer Engineering, Virginia Tech Research Center - Arlington, 900 North Glebe Road, Arlington, VA 22203, USA
Genome Biology 2011, **12(Suppl 1):**P38

The new and emerging field of systems medicine, an application of systems biology approaches to biomedical problems in the clinical setting, leverages complex computational tools and high dimensional data to derive personalized assessments of disease risk. Systems medicine offers the potential for more effective individualized diagnosis, prognosis and treatment options. The Georgetown Clinical & Omics Development Engine (G-CODE) is a generic and flexible web-based platform that serves to allow basic, translational and clinical research activities by integrating patient characteristics and clinical outcome data with a variety of high-throughput research data in a unified environment to enable systems medicine. Through this modular, extensible and flexible infrastructure, we can quickly and easily assemble new translational web applications with both analytic and generic administrative features. New analytic functionalities specific to the needs of a particular disease community can easily be added within this modular architecture. With G-CODE, we hope to help enable the creation of new disease-centric portals, as well as the widespread use of biomedical informatics tools by basic, clinical and translational researchers, through providing powerful analytic tools and capabilities within easy-to-use interfaces that can be customized to the needs of each research community. This infrastructure was first deployed in the form of the Georgetown Database of Cancer (G-DOC) [1], which includes a broad collection of bioinformatics and systems biology tools for analysis and visualization of four major omics types: DNA, mRNA, microRNA and metabolites. Although several rich data repositories for high dimensional research data exist in the public domain, most focus on a single data type and do not support integration across multiple technologies. G-DOC contains data for more than 2,500 patients with breast cancer and almost 800 patients with gastrointestinal cancer, all of which are handled in a manner that allows maximum integration. We believe that G-DOC will help facilitate systems medicine by allowing easy identification of trends and patterns in integrated datasets and will hence facilitate the use of better targeted therapies for cancer.

One obvious area for expansion of the G-CODE/G-DOC platform infrastructure is to support next-generation sequencing (NGS), which is a highly enabling and transformative emerging technology for the biomedical sciences. Nonetheless, effective utilization of these data is impeded by the substantial

handling, manipulation and analysis requirements that are entailed. We have concluded that cloud computing is well positioned to fill these gaps, as this type of infrastructure permits rapid scaling with low input costs. As such, the Georgetown University team is exploring the use of the Amazon EC2 cloud and the Galaxy platform to process whole exome, whole genome, RNA-Seq and chromatin immunoprecipitation (ChIP)-Seq NGS data. The processed NGS data will be integrated into G-DOC to ensure that they can be analyzed in the full context of other omics data. Likewise, all G-CODE projects will simultaneously benefit from these advances in NGS data handling. Through technology re-use, the G-CODE infrastructure will accelerate progress in a variety of ongoing programs that are in need of integrative multi-omics analysis and will advance our opportunities to practice effective systems medicine in the near future.

Reference

1. The Georgetown Database of Cancer [http://gdoc.georgetown.edu]

P39

Timing chromosomal abnormalities using mutation data

Steffen Durinck[†1], Christine Ho[†2], Nicholas J Wang[†1], Wilson Liao[3], Lakshmi R Jakkula[1], Eric A Collisson[1], Jennifer Pons[3], Sai-Wing Chan[3], Ernest T Lam[3], Catherine Chu[3], Kyunghee Park[4], Sung-woo Hong[4], Joe S Hur[5], Nam Huh[4], Isaac M Neuhaus[3], Siegrid S Yu[3], Roy C Grekin[3], Theodora M Mauro[3], James E Cleaver[3], Pui-Yan Kwok[3], Philip E LeBoit[6], Gad Getz[7], Kristian Cibulskis[7], Jon C Aster[8], Haiyan Huang[2], Elizabeth Purdom[2], Jian Li[9,10], Lars Bolund[9,10], Sarah T Arron[3], Joe W Gray[1,11], Paul T Spellman[1] and Raymond J Cho[3]

[1]Life Sciences Division, Lawrence Berkeley National Laboratories, CA, USA; [2]Department of Statistics, University of California, Berkeley, CA, USA; [3]Department of Dermatology, University of California, San Francisco, CA, USA; [4]Emerging Technology Research Center, Samsung Advanced Institute of Technology, Korea; [5]Samsung Electronics Headquarters, Seoul, Korea; [6]San Francisco Dermatopathology Service, San Francisco, CA, USA; [7]The Eli and Edythe L Broad Institute of Harvard and MIT, Cambridge, MA, USA; [8]Department of Pathology, Brigham and Women's Hospital, Boston, MA, USA; [9]Beijing Genomics Institute-Shenzhen, Shenzhen, China; [10]Institute of Human Genetics, Aarhus University, Aarhus, Denmark; [11]Biomedical Engineering Department, Oregon Health and Science University, Portland, OR, USA

[†]Equal contributors

Genome Biology 2011, **12(Suppl 1):**P39

Background Tumors accumulate large numbers of mutations and other chromosomal abnormalities because of a breakdown in genomic repair mechanisms, which is a hallmark of tumors. Not all of these abnormalities are thought to be crucial for tumor growth and progression, and it is a question of great importance to try to identify critical abnormalities, particularly as possible targets for treatment. A strong indicator of the importance of an abnormality is the order in which it occurred relative to other abnormalities, with triggering events likely to have occurred earlier.

Methods In general, we cannot directly observe the temporal progression of a tumor; however, for some types of chromosomal gains and losses, the mutations within the event can be classified as having occurred before or after the event by virtue of being homozygous or heterozygous. The simplest case is copy-neutral loss of heterozygosity (CN-LOH), in which it is reasonable to assume that homozygous mutations occurred before the LOH event and that heterozygous mutations occurred after the LOH event. Using sequencing data, we developed a probabilistic model for the observed allele frequency of a mutation, which allows us to estimate the true proportion of pre- and post-event mutations. Specifically, we modeled the number of reads with the mutation as a mixture model of binomials and estimated the mixing proportion. On the basis of this model, we can estimate this proportion for all LOH events within a sample and give a temporal ordering to the events within a sample. We applied this method to exome capture sequencing data that were obtained from eight primary cutaneous squamous cell tumors and matched normal pairs [1].

Results An immediate novel result of the analysis was that CN-LOH of chromosome 17p was temporally ordered as the first event (among CN-LOH events) in all four of the eight tumors that had CN-LOH of this region. The well-known tumor suppressor gene *TP53* is located in the CN-LOH region and has pre-CN-LOH mutations in all of the samples, further strengthening the role of *TP53* as a trigger for tumor progression.

Conclusions Our method gives novel insight into the biology of tumor progression through a quantitative evaluation of temporal ordering of chromosomal abnormalities. Moreover, it yields a quantitative measure for comparing samples to highlight driver mutations and events.

Reference

1. Durinck S, Ho C, Wang N, Liao W, Jakkula LR, Collisson EA, Pons J, Chan SW, Lam ET, Chu C, Park K, Hong SW, Hur JS, Huh N, Neuhaus IM, Yu SS, Grekin RC, Mauro TM, Cleaver JE, Kwok PY, LeBoit PE, Getz G, Cibulskis K, Aster JC, Huang H, Purdom E, Li J, Bolund L, Arron ST, Gray JW *et al.*: Temporal dissection of tumorigenesis in primary cancers. *Cancer Discov* 2011, **1:**OF1-OF7.

P40

RNA-Seq-based transcriptional map of the bovine respiratory disease pathogen *Histophilus somni* 2336

Ranjit Kumar[1,2,3], Mark L Lawrence[1,2], James Watt[4], Amanda Cooksey[2], Shane C Burgess[1,2] and Bindu Nanduri[1,2]

[1]College of Veterinary Medicine, Mississippi State University, Mississippi State, MS 39762, USA; [2]Institute for Genomics, Biocomputing and Biotechnology, Mississippi State University, Mississippi State, MS 39762, USA; [3]Center of Biostatistics and Bioinformatics, University of Mississippi Medical Center, Jackson, MS 39216, USA; [4]Eagle Applied Sciences, Clinical Research Laboratory, 81st Medical Group, Keesler AFB, MS, USA

Genome Biology 2011, **12(Suppl 1):**P40

Background Genome structural annotation - that is, the identification and demarcation of the boundaries of all of the functional elements in a genome (such as the genes, non-coding RNAs, proteins and regulatory elements) - is a prerequisite for systems level analysis. Current genome annotation programs do not identify all of the functional elements of the genome, especially small non-coding RNAs (sRNAs). Transcriptome analysis is a complementary method for identifying 'novel' genes, small RNAs, regulatory regions and operon structures, thus improving structural annotation in bacteria. In particular, the identification of non-coding RNAs has revealed their widespread occurrence and functional importance in gene regulation, stress and virulence. However, very little is known about non-coding transcripts in *Histophilus somni*, one of the causative agents of bovine respiratory disease, as well as bovine infertility, abortion, septicemia, arthritis, myocarditis and thrombotic meningoencephalitis.

Methods In this study, we generated a single-nucleotide resolution transcriptome map of *H. somni* strain 2336 using RNA-Seq (Illumina). A Perl script was written to convert Illumina reads into FASTQ format. The software tools MAQ, Bowtie and SAMtools were used to process the raw data and generate pileup format, which provides the signal map file in per-base format coverage. In-house Perl scripts were written to identify novel sRNAs, putative novel proteins and operon structures. Comparative genomic analysis of *H. somni* strain 2336 and the avirulent strain 129Pt was performed using the tool Mauve. The processed data were submitted to the Gene Expression Omnibus database with accession number GSE29578.

Results The RNA-Seq-based transcriptome map identified 94 sRNAs in the *H. somni* genome, of which 82 had not been predicted or reported in earlier studies. We also identified 38 novel potential protein-coding ORFs that are not in the current genome annotation. The transcriptome map allowed the identification of 278 operon structures (for a total of 730 genes) in the genome. Compared with the genome sequence of a non-virulent strain, 129Pt, a disproportionate number of sRNAs (about 30%) were located in a genomic region unique to strain 2336 (accounting for about 18% of the total genome). This observation suggests that a number of the newly identified sRNAs in strain 2336 may be involved in strain-specific adaptations that could include virulence.

Conclusions Overall, this study describes an RNA-Seq-based transcriptome map of *H. somni*, an important agricultural pathogen, that was constructed to identify functional genomic elements. Our genome-wide survey predicts numerous novel expressed regions that need to be characterized biologically to improve our understanding of disease pathogenesis. A description of all of the functional elements in the *H. somni* system is a prerequisite for using holistic systems approaches to understand the complex pathogenesis of bovine respiratory disease.

P41

On using optical maps for genome assembly

Henry Lin[1] and Mihai Pop[1]
[1]Center for Bioinformatics and Computational Biology, University of Maryland, College Park, MD 20742, USA
Genome Biology 2011, **12(Suppl 1):**P41

Background In this work, we study the benefits of using optical maps to improve genome assembly. Many modern assembly algorithms rely on a de Bruijn graph paradigm to reconstruct a genome from short reads. Ambiguities caused by repeats within the genome cause the final assembly to be broken up into many contigs, because the assembler does not have enough information to find the one correct traversal of the graph. Optical mapping technology can be useful for determining the correct path in the de Bruijn graph, through providing estimates on the locations of one or more restriction enzyme patterns in the genome, thereby constraining the possible traversals of the graph to only those that are consistent with the map. A particular traversal that does not align well with the optical map can be discarded as incorrect. Previous work has shown how to construct optical maps [1,2] for scaffolding contigs [3].
Methods Our algorithm relies on a depth-first search strategy. As the depth-first search proceeds and its corresponding sequence is extended, we check whether the resultant sequence would generate an optical map that matches the optical map of the genome. If the candidate *in silico* optical map matches the optical map of the genome, we proceed with the depth-first search. Otherwise, we backtrack in the depth-first search until we find a path that covers the entire graph and whose sequence has an optical map that matches the optical map of the entire genome. Although the total number of paths in the de Bruijn graph can be exponential in the number of nodes and edges in the graph [4], a reference optical map can effectively prune the search space of paths. To improve performance, we start by finding edges in the de Bruijn graph that can be uniquely placed on the optical map. These edges, which we call landmark edges, can also help guide our depth-first search. Although there may be multiple paths in the de Bruijn graph that can yield sequences with optical maps that match the genome's optical map, these paths all yield very similar sequences in most cases.
Results Given modest assumptions about the errors in the optical map, initial simulations show that our algorithm is very effective at assembling bacterial genomes, given read lengths of 100 or longer. The majority of our assemblies match the original sequences used in our simulations very closely. We will also present the results of simulations aimed at measuring the effect of errors on the correctness of the reconstruction and at measuring how the choice of restriction enzymes can improve the sequence assembly.
Conclusions Our work shows that optical maps can be used effectively to aid in genome assembly. We are currently extending our approach to handle much larger graphs and to tolerate higher amounts of mapping error. In our final assembly, we would also like to be able to detect and mark regions that we are less certain about and regions that we are confident are correct.

References

1. Aston C, Mishra B, Schwartz DC: **Optical mapping and its potential for large-scale sequencing projects.** *Trends Biotechnol* 1999, **17:**297-302.
2. Valouev A, Schwartz DC, Zhou S, Waterman MS: **An algorithm for assembly of ordered restriction maps from single DNA molecules.** *Proc Natl Acad Sci USA* 2006, **103:**15770-15775.
3. Nagarajan N, Read TD, Pop M: **Scaffolding and validation of bacterial genome assemblies using optical restriction maps.** *Bioinformatics* 2008, **24:**1229-1235.
4. Kingsford C, Schatz MC, Pop M: **Assembly complexity of prokaryotic genomes using short reads.** *BMC Bioinformatics* 2010, **11:**21.

P42

Utilizing whole genome sequences to study population genomics of gene networks: a case study of the *Arabidopsis thaliana* immune-signaling network

Mridu Middha[1,2], Yungil Kim[3], Peter Morrell[4], Chad Myers[3] and Fumiaki Katagiri[2]
[1]Biomedical Informatics and Computational Biology, University of Minnesota, Minneapolis, MN 55455, USA; [2]Department of Plant Biology, Microbial and Plant Genomics Institute, University of Minnesota, St. Paul, MN 55108, USA; [3]Department of Computer Science & Engineering, University of Minnesota, Minneapolis, MN 55455, USA; [4]Department of Agronomy & Plant Genetics, University of Minnesota, St. Paul, MN 55108, USA
Genome Biology 2011, **12(Suppl 1):**P42

Arabidopsis thaliana is a member of the mustard (Brassicaceae) family that is widely used as a model organism in plant biology. The 1001 Genomes Project [1] has been sequencing the genomes of *Arabidopsis* strains (accessions) and has made these sequences available. We selected the genomes of 30 *Arabidopsis* accessions with diverse geographical and environmental origins for our analysis. Using the TAIR8 annotation of the *Arabidopsis* reference genome, for the accession Col-0, we generated a dataset of approximately 27,000 protein-coding genes for all of the 30 genomes. With such population genomic data, it is feasible to ask whether a group of genes is under a different type of selection from the rest of the genome.
The plant immune-signaling network is robust to network perturbations. We hypothesized that genes that constitute a robust network tend to be under neutral selection because deleterious mutations in such genes do not strongly affect the immune phenotype owing to the robustness of the network. We identified the component genes of the plant immune-signaling network in a relatively unbiased manner by mining AraNet [2], which is a functional gene network built without using phenotype information. We compared population genetic summary statistics for the network component genes and those for all of the genes in the genome. For example, Tajima's D is such a summary statistic, and positive, negative and zero values of Tajima's D suggest diversifying, purifying and neutral selection, respectively, when the effective population size does not change. The Tajima's D value distribution for all of the genes in the genome has a single clear peak with a negative value, suggesting that purifying selection is the genomic norm.
Our preliminary results showed that the plant immune-signaling network genes are significantly enriched with genes whose Tajima's D values are near zero compared with all of the genes in the genome. This finding suggests that there is a lower level of purifying selection among the network component genes than other genes.

References

1. **1001 Genomes Project** [http://www.1001genomes.org/]
2. **AraNet** [http://www.functionalnet.org/aranet/about.html]

P43

Genomes or exomes: evaluation of cost, time and coverage

Sumit Middha[1], Jeanne L Theis[2], Adele H Goodloe[2], Timothy M Olson[2] and Jean-Pierre A Kocher[1]
[1]Division of Biomedical Statistics and Informatics, Mayo Clinic, Rochester, MN, USA; [2]Division of Cardiovascular Diseases, Mayo Clinic, Rochester, MN, USA
Genome Biology 2011, **12(Suppl 1):**P43

Next-generation sequencing technology platforms are driving the development of a variety of approaches to study genomic variation associated with disease. One of these approaches, exome sequencing, specifically targets the coding regions of the genome, which are captured and sequenced. Compared with whole genome sequencing, exome sequencing offers the advantages of being cost- and time-effective while providing deeper coverage of coding variants, which are more likely to affect function.
However, the protocol is known to be only partially reliable and might miss some of the coding regions. To assess how much coding region could be missed or off target, we compared whole genome and exome sequencing data derived from one sample that was processed by the Illumina GA-IIx platform. Our in-house-developed workflow named TREAT (Targeted RE-sequencing and Annotation Tool) was used to align and annotate the data. We provide a summary of the comparison between the two datasets, including the total number of reads produced, the time needed for sequencing and analysis, the coverage of coding regions and the agreement between called variants.

P44

Integrating whole transcriptome sequence data and public databases for analysis of somatic mutations in tumors

Amin A Momin[1], Brian P James[2], Thomas C Motter[2], Humam N Kadara[3], Garth Powis[2] and Ignacio I Wistuba[3,4]

[1]Department of Bioinformatics and Computational Biology, The University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA; [2]Department of Experimental Therapeutics, The University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA; [3]Department of Thoracic/Head and Neck Thoracic Oncology, The University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA; [4]Department of Pathology, The University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA

Genome Biology 2011, **12(Suppl 1):**P44

The annotation of pathologically relevant somatic variations has gained importance with the wide use of next-generation sequencing in biomedical studies. At present, this evaluation is performed using public tools such as SAMtools and ANNOVAR by comparing predicted mutations and small nucleotide variations (SNVs) with databases such as 1000 Genomes and dbSNP, as well as with paired normal data if available. However, these analytical methods lack the ability to integrate information from the different analyses into a single output. Additionally, many approaches are filter based and remove data that does not match specific criteria, thus leading to the removal of variations that would otherwise be reconsidered later. To this end, we have developed a Perl wrapper script that utilizes standard next-generation sequencing output files along with SAMtools and ANNOVAR to produce an annotated tumor variant file with sequence calls from related tumor and matched normal samples.

We performed SOLiD paired-end sequencing of the whole transcriptome of one lung adenocarcinoma and seven normal lung samples (including one matched normal). BioScope 1.3 was used to map the reads, and the SNVs were identified by the diBayes package. The map files in binary-sequence alignment format (BAM) and SNV files in generic feature format (GFF) were used to annotate the tumor SNVs with matched normal sequence information at each position (diBayes and SAMtools), as well as other normal samples (both position and gene based). Furthermore, SNVs were annotated with positional information, including whether intronic, exonic, or synonymous versus nonsynonymous, as well as with data from the 1000 Genomes Project (allele frequency), the dbSNP database (rs identifiers) and the Catalogue of Somatic Mutations in Cancer (COSMIC) database. Of the 1,804 SNVs initially identified in the tumor sample, 138 SNVs were found in non-coding RNA, and 75 did not appear in the normal samples according to diBayes or in the specific matched normal sample according to SAMtools. Because the capacity to sequence the whole transcriptome is subject to the expression level, the possibility of failure to detect variations in normal lung samples cannot be ignored. To address this concern, we analyzed 1000 Genomes data and found that only 23 of the 75 potential tumor-specific SNVs exhibited allele frequencies <1%, and 6 of these exist in dbSNP. All of these steps can be rapidly performed by a researcher, and modifying the approach to identify other types of SNVs is easily achievable.

The use of a single script that tracks input file names and locations is expected to improve data handling and reporting. Notably, all variant data are present in a single file, allowing straightforward modification of criteria and instant hypothesis testing and therefore reducing the need for an informed end user to re-engage a bioinformatician to address another biological question.

P45

Effective detection of rare variants in pooled DNA samples using cross-pool tail-curve analysis

Tejasvi S Niranjan[†1,2], Abby Adamczyk[†1], Hector Corrada Bravo[†3,4], Margaret A Taub[5], Sarah J Wheelan[5,6], Rafael Irizarry[5] and Tao Wang[1]

[1]McKusick-Nathans Institute of Genetic Medicine and Department of Pediatrics, Johns Hopkins University School of Medicine, Baltimore, MD 21205, USA; [2]Predoctoral Training Program in Human Genetics, Johns Hopkins University School of Medicine, Baltimore, MD 21205, USA; [3]Center for Bioinformatics and Computational Biology, Department of Computer Science, University of Maryland, College Park, MD 20742, USA; [4]Present address: Center for Bioinformatics and Computational Biology, Department of Computer Science, University of Maryland, College Park, MD 20742, USA; [5]Department of Biostatistics, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD 21205, USA; [6]Department of Oncology, Johns Hopkins University School of Medicine, Baltimore, MD 21287, USA

[†]Equal contributors

Genome Biology 2011, **12(Suppl 1):**P45

Rare genetic variants of large effect may confer a substantial genetic risk for common diseases and complex traits. There is considerable interest in sequencing limited genomic regions such as candidate genes and target regions identified by genetic linkage and/or association studies. Next-generation sequencing of pooled DNA samples is an efficient way to identify rare variants in large sample sets. Although sample pooling can reduce the labor and cost of sequencing, it also reduces the sensitivity and specificity for effective and reliable identification of rare variants. It remains a challenge to solve these problems using the available computational genomics tools. We have developed an effective Illumina-based sequencing strategy using pooled samples and have optimized a novel base-calling algorithm, Srfim, and a variant-calling algorithm, SERVIC^{4}E (Sensitive Rare Variant Identification by Cross-pool Cluster, Continuity & Tail-Curve Evaluation). SERVIC^{4}E analyzes base composition by cycle or tail-curves across sample pools and employs multiple filtering strategies, including quality and continuity cluster analysis, average quality filtering, tail-curve filtering and error proximity filtering, to accurately identify rare sequence variants. We validated these algorithms using two independent Illumina sequence datasets generated from different pool sizes, read lengths and sequencing chemistries. Using these programs, we identified 32 coding variants, including 14 present only once over 24 exon-containing regions in one sample cohort (n = 480), and 41 coding variants, including 16 present only once in the same regions in an unrelated cohort (n = 480). Validation of these variants by Sanger sequencing revealed an excellent combination of sensitivity (97.8% and 96.4%) and specificity (84.9% and 93.8%) for variant detection in pooled samples from both cohorts, respectively. Data from these studies showed that our algorithms compare favorably with the available programs, including SAMtools, SNPSeeker, CRISP and Syzygy, for the effective and reliable detection of rare variants in pooled samples.

P46

Abstract not submitted for online publication.

P47

Microbial community function and biomarker discovery in the human microbiome

Nicola Segata, Sahar Abubucker, Johannes Goll, Alyxandria M Schubert, Jacques Izard, Brandi L Cantarel, Beltran Rodriguez-Mueller, Levi Waldron, Jeremy Zucker, Mathangi Thiagarajan, Bernard Henrissat, Owen White, Scott T Kelley, Barbara Methé, Patrick D Schloss, Wendy S Garrett, Dirk Gevers, Makedonka Mitreva and Curtis Huttenhower

Harvard School of Public Health, Biostatistics, 655 Huntington Avenue, Boston, 02115, USA

Genome Biology 2011, **12(Suppl 1):**P47

Microbial communities carry out the majority of the biochemical activity on the planet, and they play integral roles in processes such as metabolism and immune homeostasis in the human microbiome. Whole genome shotgun sequencing of such communities' metagenomes is becoming an increasingly feasible complement to obtaining organismal information from taxonomic markers. However, the resultant dataset typically comprises short reads from hundreds of different organisms, making it challenging to assemble and functionally annotate these sequences in the standard manner for single-organism genomes.

We describe an alternative to this approach to infer the functional and metabolic potential of a microbial community metagenome by determining whether gene families and pathways are present or absent, as well as their relative abundances, directly from short sequence reads. We validated this methodology using synthetic metagenomes, recovering the presence and abundance of large pathways and of small functional modules with high accuracy. We subsequently applied this approach to the microbial communities of 649 metagenomes drawn from 7 primary body sites on 102 individuals as part of the Human Microbiome Project (HMP), demonstrating the scalability of our methodology and the critical importance of microbial metabolism in the human microbiota. This provided a framework in which to

define functional diversity in comparison to organismal ecology, including an example of microbial metabolism linked to specific organisms and to host phenotype (vaginal pH) in the posterior fornix. We provide profiles of 168 functional modules and 196 metabolic pathways that were determined to be specific to one or more niches within the human microbiome, including details of glycosaminoglycan degradation in the gut.

Understanding how and why these biomolecular activities differ among environmental conditions or disease phenotypes is, more broadly, one of the central questions addressed by high-throughput biology. We have thus developed the linear discriminant analysis (LDA) effect size algorithm (LEfSe) to discover and explain microbial and functional biomarkers in the human microbiota and other microbiomes. We demonstrate this method to be effective for mining human microbiomes for metagenomic biomarkers associated with mucosal tissues and with different levels of oxygen availability. Similarly, when applied to 16S rRNA gene data from a murine ulcerative colitis gut community, LEfSe confirms the key role played by *Bifidobacterium* in this disease and suggests the involvement of additional clades, including the Clostridia and *Metascardovia*. A quantitative validation of LEfSe highlights a lower false positive rate, consistent ranking of biomarker relevance, and concise representations of taxonomic and functional shifts in microbial communities associated with environmental conditions or disease phenotypes.

Implementations of both methodologies are available at the Huttenhower laboratory's website [1,2]. Together, they provide a way to accurately and efficiently characterize microbial metabolic pathways and functional modules directly from high-throughput sequencing reads and, subsequently, to identify organisms, genes or pathways that consistently explain the differences between two or more microbial communities. This has allowed the determination of community roles in the HMP cohort, as well as their niche and population specificity, which we anticipate will be applicable to future metagenomic studies.

References

1. The HMP Unified Metabolic Analysis Network [http://huttenhower.sph.harvard.edu/humann]
2. LEfSe Algorithm [http://huttenhower.sph.harvard.edu/lefse]

P48

A high-throughput-sequence analysis infrastructure technology investigation framework for the evaluation of next-generation sequencing software

Xiaoyu Liu, Sumit Middha, Steven Hart, Asha Nair, Ahmed Hadad, Patrick Duffy, Jean-Pierre Kocher

Division of Biomedical Statistics and Informatics, Mayo Clinic, Rochester, MN 55905, USA

Genome Biology 2011, **12(Suppl 1):**P48

High-throughput sequencing (HTS) is an emerging technology that promises to deliver unparalleled information on genomic variations. As technology evolves and matures, and as a deeper understanding of this technology is gained, new and upgraded tools for analyzing HTS will become available and will need to be evaluated and validated. To facilitate this cumbersome task, we have developed an HTS validation framework into which both in-house-generated synthetic datasets and well-characterized experimental datasets have been incorporated for controlled testing and evaluation of these analysis tools. Currently, the framework can be used to assess algorithms for short-read mapping, variant calling and RNA-Seq-derived gene expression measurements. The framework is deployed in the Amazon EC2 cloud so that it is available to the broader research community. Using our framework, researchers can further validate interfaced applications with preferred parameters, upload their own datasets for processing, and interface new applications with the framework for validation and comparison.

We report the performance of several alignment, variant calling and RNA-Seq analytic tools that have been tested with our framework. We also provide feedback on the challenges and benefits of Amazon EC2 deployment.

Cite abstracts in this supplement using the relevant abstract number, e.g.: Liu X, *et al.*: A high-throughput-sequence analysis infrastructure technology investigation framework for the evaluation of next-generation sequencing software. *Genome Biology* 2011, **12(Suppl 1):**P48.