Beyond the Genome 2012

Boston, MA, USA. 27-29 September 2012

ORAL PRESENTATIONS

O1
Abstract not submitted for online publication.

O2
Abstract not submitted for online publication.

O3
Pathway-based analysis of mutation impact
Joshua Stuart
University of California Santa Cruz, USA
BMC Proceedings 2012, 6(Suppl 6):O3

Genomic damage plays an important role in cancer onset and progression. But how exactly do mutations, copy number changes, and epigenetic events affect the wiring of otherwise normal-functioning cellular pathways? Genomic probing with technologies such as RNA sequencing, copy number profiling, and DNA promoter methylation arrays are providing unprecedented views of cells. Uncovering predictive models from these datasets to number key altered pathways in cancer is a major challenge that could offer breakthroughs for personalized medicine.

I’ll describe a new pathway-based strategy that gives novel insight into the functional impact of mutations in particular genes. The major mechanism by which cancer arises is through somatic mutations. Individual tumors can contain hundreds to thousands of mutations. It is critical to distinguish mutations that have an important role defining the cancer – driver mutations – from mutations that are unimportant to the tumor – passenger mutations. A pathway-based approach is able to detect a shift in the downstream effects of an altered gene compared to what is expected from its upstream regulatory input. Application to several datasets across multiple tissues revealed several important driver mutations even among rarely mutated genes. Thus, pathway analysis shows promise in differentiating driver and passenger events that will increase our understanding of cancer disease mechanisms, which can help identify novel targets for treatment.

O4
Illuminating the genetics of complex human diseases
Michael C Schatz
Simons Center for Quantitative Biology. Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, USA
BMC Proceedings 2012, 6(Suppl 6):O4

In 2007, Zhao et al. [1] proposed a unified genetic theory of sporadic and inherited autism spectrum disorders to explain the complex familial patterns observed. The model explains why most families have a low risk of autistic children with an overall incidence rate of approximately 1 in 150 births, but a small minority of families have a 50% risk for male offspring. Their model links these two risk classes by their genetic origins: sporadic autism in low-risk families is mainly caused by highly penetrant spontaneous mutations in autism-related genes, whereas inherited autism in high risk families is mainly caused by unaffected parents carrying a causative mutation that is transmitted in a dominant fashion to their offspring. The evidence for the model was based on the available genotyping assays that showed an elevated rate of de novo copy number variants (CNVs) in children with autism spectrum disorders.

Since then, we and other collaborators have been following up with a high-resolution exome sequencing study of 2,800 families from the Simons Simplex Collection to pinpoint the genetic components of the disorder. Unlike the early studies that could only detect large copy number events spanning tens or hundreds of thousands of basepairs across multiple genes, our new study has power to examine single nucleotide and indel mutations within individual genes. From our preliminary analysis of approximately 350 of these families [2], as well as the reports of three other groups, we have collected strong evidence for the role of ‘likely gene-disrupting’ (LGD) mutations (nonsense, splice site and frame shifts), with affected individuals having twice as many LGDs compared to unaffected siblings, and five ‘double hit’ genes (CHD8, DYRK1A, KATNAL2, PDGZ, SCN2A) having two de novo mutations in unrelated individuals. From this analysis, we estimate approximately 400 genes are targets of autism spectrum disorders. Interestingly, from the gene set we have already identified, we have discovered a strong association between the targets LGD mutations in autism and in vivo targets of the RNA-binding translational regulator FMRP (encoded by FMR1), which result in Fragile X Syndrome when silenced or mutated. The large number of samples in the study necessitated that we develop a high performance parallel sequence analysis pipeline that could scale to the large volume of data and make use of local disk storage. We were able to make use of several existing tools for the preliminary analysis (BWA, SAMTools, GATK, etc), and we also developed additional components for genotyping within a family and across the population using a multinomial statistical approach. This includes a novel sequence analysis algorithm for discovering insertion & deletion variants using a localized sequence assembly approach that is superior to standard mapping algorithms. We continue to refine the algorithms to improve scalability, sensitivity, and specificity, and are beginning to apply it towards analyzing the genomes of families with other complex cognitive disorders.

References

O5
Abstract not submitted for online publication.

O6
Hypothesis-generating clinical genomics research and predictive medicine
Leslie Biesecker
National Human Genome Research Institute, USA
BMC Proceedings 2012, 6(Suppl 6):O6

The advent of affordable genome and exome sequencing provides incredible opportunities and poses significant challenges for clinical research and clinical care. For the first time, it is technically feasible to access the entire genetic architecture of a phenotype. The dissection of
this genetic architecture of disease will yield unprecedented insights into molecular pathophysiology and provide numerous therapeutic targets. Soon, the primary etiology of all Mendelian traits will be elucidated and modifiers will follow. Translating this genetic architecture into diagnostics and therapeutics will be feasible and will require creative, aggressive and thoughtful approaches to numerous challenges. One of the first applications will be predictive medicine, which should initially focus on high-penetration Mendelian phenocopies of common diseases and disorders with effective interventions. These include cancer susceptibility syndromes, cardiomyopathies and dysrhythmias, malignant hyperthermia, dyslipidemias and a host of other disorders. These approaches will require improved abilities to predict phenotype from genotype and a clinical paradigm shift that supports a disease screening approach (as distinct from a differential diagnosis approach).

As well, researcher and clinicians will need to develop creative approaches to dealing with data overload - no physician (research or clinical) can address three million variations. Prioritizing these variants into clinically appropriate categories is urgent, as is developing an unbiased assessment of penetrance in variable expressivity. I will give examples of several of these approaches and their attendant challenges from the ClinSeq project in the intramural NIH.

07 Whole-genome sequencing and disease-gene detection
Lynn B Jorde
Department of Human Genetics, University of Utah School of Medicine, USA
BMC Proceedings 2012, 6(Suppl 6):O7

Recent advances in DNA sequencing technology are transforming our understanding of the genetic basis of rare human diseases. It is now possible to rapidly and cost-effectively interrogate the majority of protein-coding bases in the human genome (known collectively as the exome), finding mutations that would have been difficult if not impossible to discover with the traditional approaches of linkage and candidate gene sequencing. However, unambiguously identifying the disease-causing mutations in a patient’s exome remains challenging. Next-generation sequencing, while powerful, still requires careful filtering to remove errors and is underpowered for discovery of larger insertions/deletions (indels) and complex variants; coverage of genes is incomplete due to biases in DNA capture and sequencing; and predicting the likely functional impact of observed variants is still an immature science. Importantly, existing catalogues both of “normal” variation in reference populations and of reported disease-causing mutations are incomplete and biased. Finally, methods for communicating the results of large-scale sequencing to key target audiences – clinicians, patients and researchers from other fields – remain poorly developed.

In this presentation I describe recent advances in variant-calling from next-generation sequencing technology, and their application to exome data from over 15,000 individuals from multiple different disease-specific studies. Functional annotation of sequence variants across these large samples illustrates the surprising degree of putatively functional genetic variation even in apparently healthy individuals. However, the existence of a very large and accurately called reference panel of exomes provides a powerful resource for interpreting the probability of disease causation for variants observed in rare disease samples. I will discuss new approaches to analyzing and presenting the results from family-based studies of protein-coding mutations in rare disease patients.

08 Sifting disease-causing signal from genomic noise
Daniel G MacArthur1,2
1Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA, 02114, USA; 2Broad Institute of Harvard and MIT, Cambridge, MA, 02142, USA
BMC Proceedings 2012, 6(Suppl 6):O8

Recent advances in DNA sequencing technology are transforming our understanding of the genetic basis of rare human diseases. It is now possible to rapidly and cost-effectively interrogate the majority of protein-coding bases in the human genome (known collectively as the exome), finding mutations that would have been difficult if not impossible to discover with the traditional approaches of linkage and candidate gene sequencing. However, unambiguously identifying the disease-causing mutations in a patient’s exome remains challenging. Next-generation sequencing, while powerful, still requires careful filtering to remove errors and is underpowered for discovery of larger insertions/deletions (indels) and complex variants; coverage of genes is incomplete due to biases in DNA capture and sequencing; and predicting the likely functional impact of observed variants is still an immature science. Importantly, existing catalogues both of “normal” variation in reference populations and of reported disease-causing mutations are incomplete and biased. Finally, methods for communicating the results of large-scale sequencing to key target audiences – clinicians, patients and researchers from other fields – remain poorly developed.

In this presentation I describe recent advances in variant-calling from next-generation sequencing technology, and their application to exome data from over 15,000 individuals from multiple different disease-specific studies. Functional annotation of sequence variants across these large samples illustrates the surprising degree of putatively functional genetic variation even in apparently healthy individuals. However, the existence of a very large and accurately called reference panel of exomes provides a powerful resource for interpreting the probability of disease causation for variants observed in rare disease samples. I will discuss new approaches to analyzing and presenting the results from family-based studies of protein-coding mutations in rare disease patients.

09 De novo diagnostics of patients with intellectual disability
Joris A Veltman
Department of Human Genetics, Radboud University Nijmegen Medical Centre, PO Box 9101, 6500 HB Nijmegen, The Netherlands
BMC Proceedings 2012, 6(Suppl 6):O9

Germline coding de novo mutations (SNVs, indels as well as CNVs) are an important cause of moderate to severe forms of intellectual disability (ID) and associated syndromes. Exome sequencing now allows us to reliably identify these mutations using a single genomic test, and we have recently implemented exome sequencing in the diagnostic follow-up of these patients. In this presentation, I will first discuss the role of de novo mutations in genetic disease and the associated risk factors such as local genomic structure and paternal age. Next, I will describe our recent work using a diagnostic family-based exome sequencing approach to test this de novo mutation hypothesis in 100 patients with unexplained ID, as well as targeted follow-up studies of several candidate ID genes in 750 additional patients. A total of 79 unique coding de novo mutations were identified and validated in 52 patients. Damaging de novo (n = 10) as well as X-linked maternally-inherited (n = 3) mutations were detected in known ID genes, resulting in a minimal diagnostic yield of 13% in this cohort. In addition, potentially causative de novo mutations in novel candidate ID genes were detected in 22 patients. For three of these candidate genes, recurrent de novo mutations were identified in patients with similar phenotypes, confirming that they are true ID genes. To further expand the possibilities of exome sequencing for mutation detection, we have recently implemented automated CNV detection on exome data, and compared its performance to that of high-resolution genomic microarrays. This analysis shows that exome sequencing can reliably detect the large majority of pathogenic de novo CNVs, responsible for approximately 15% of ID.

In conclusion, de novo mutations therefore represent an important cause of ID, and exome sequencing is an effective diagnostic strategy for their detection.

References

O10 Abstract not submitted for online publication.
O11 Clinical diagnostic whole genome sequencing in a paediatric population: experience from our WGS genetics clinic

Elizabeth Worthen1,2, David Blick1,2, David Offord1,2, Mary Shimoyama1,4, Regan Veith2,3, Mary Kim1,4, Howard Jacob1,2,5, Regan Veith2,3, George Kowalski1, Bradley Taylor1,5

1Human and Molecular Genetics Center, The Medical College of Wisconsin, Milwaukee 53226, USA; 2The Department of Pediatrics, The Medical College of Wisconsin, Milwaukee 53226, USA; 3The Children's Hospital of Wisconsin, 9000 W. Wisconsin Avenue, Milwaukee, 53226, USA; 4The Department of Surgery, The Medical College of Wisconsin, Milwaukee 53226, USA; 5The Department of Physiology, The Medical College of Wisconsin, Milwaukee 53226, USA

Background: We are in an unprecedented era where the genome of an individual can be sequenced in a day for low cost in a single sequencing run. The availability this information is changing and challenging the practice of medicine. At The Medical College of Wisconsin (MCW) we have been running a whole genome sequencing (WGS) based genetics clinic for over a year. WGS together with the processes and tools developed for analysis and interpretation are being used to identify disease-causing mutations in rare paediatric disorders in children being treated at our affiliated Children's Hospital of Wisconsin.

Materials and methods: We have developed a fully integrated WGS process for clinical diagnosis: from sequence generation through to interpretation and reporting, including follow-up and counselling tailored specifically to WGS testing. Our initial pilot program has been extended to a Children's Hospital of Wisconsin clinic providing dedicated genetics staff. This required development of a series of CAP/CLIA compliant protocols and pipelines for dealing with samples and returning reports, and has spurred ongoing development of tools including a platform for electronic health record (EHR) data extraction. Clinically validated WGS analysis tools were also developed to support determination of whether the variant is: (1) a likely error, (2) known to be causally linked with human disease, (3) of a type likely to lead to altered function thereby giving rise to a phenotypic change, and/or (4) within a gene with a function associated or relevant to the patient phenotype (including identification of causative disease genes). They support consideration of variant and functional information simultaneously, as required for interpretation and reporting. Development of this clinic also required formation of institutional review bodies, definition of a structure for counselling, and development of criteria for appropriate reporting and return of WGS-based findings.

Results: Of 22 cases approved for clinical WGS in the MCW pilot program, 15 have been analyzed and/or reported. Our protocols and procedures have been requested and shared with other sites moving into this area. Our tools have been clinically validated and are in use, and continue to be developed to add more sophisticated algorithms to support interpretation. Use of these tools has rendered diagnoses in some but not all cases and in some cases; these diagnoses have led to alteration in treatment in some but not all cases [1,2]. In addition these tools are being applied to uncovering causative mutations for dozens of additional patients as part of various research project collaborations.

Conclusions: WGS and subsequent analysis can be used to identify causative mutations and result in diagnosis in patients with rare disorders today. Reaching this stage requires expert use of sophisticated bioinformatics tools for interpretation. The diagnosis can be used to alter treatment in some cases. We will: (1) provide examples where WGS analysis has rendered a molecular diagnosis in initial patients or families in the setting of novel disease, (2) highlight challenges faced, and (3) focus on the nature of the tools developed and how they are used in a clinical setting.

References

O12 Characterizing epistatic hotspots of human disease

Talilah Andrews, Caleb Webber

MRC Functional Genomics Unit, Department of Physiology Anatomy and Genetics, University of Oxford, Oxford, Oxfordshire OX1 3PT, UK

Background: We recently reported that single de novo copy number variants (CNVs) in patients with developmental disorders frequently affected multiple functionally similar genes [1]. Since functional clusters are present in many eukaryotic species, including humans [2], a single CNV can affect multiple genes, potentially incurring additional deleterious effects due to epistatic interactions. Furthermore, the position of genes within a biological pathway may be relevant to the phenotypic impact of their disruption [3]. In this study, we investigated the roles that epistatic interactions and the network position of disrupted genes play in developmental disorders.

Materials and methods: We obtained a set of 626 de novo CNVs identified in patients with developmental disorders from the Database of Chromosomal Imbalances and Phenotype in Humans using Ensemble Resources (DECIPHER). Employing a novel integrated functional linkage network, we examined clustering and importance (centrality within the network) of the genes affected by these de novo CNVs. A fast neighbor joining clustering algorithm was used to identify gene groups within each CNV defined as those linked by the top 1% of shortest paths within the functional linkage network. To ensure functional clusters were not simply the result of tandem duplications, we collapsed paralogous genes into a single copy within the genome. Significant enrichments in clustering and central network position were used to build a predictive model able to scan the genome and identify similar regions whose copy number change may also predispose to disease.

Results: DECIPHER de novo CNVs were significantly enriched for large functional clusters with low within-cluster similarity compared with gene-number matched randomizations. Functional clusters were even more significantly enriched when paralogous genes were collapsed. Clusters were present in 357 of 626 CNVs, with an average size of four genes. In addition all measures of network centrality were significantly high, with average maximum betweenness (bottlenecks) the most significant. Bottleneck genes tended to be haplo-insufficient and highly pleiotropic when knocked out in mice. Functional clusters frequently contained bottleneck genes and these regions were frequently affected by CNVs in more than one patient.

Conclusions: DECIPHER de novo CNVs identify putative epistatic hotspots, which are clusters of functionally related genes whose disruptions are associated with developmental disorders. In addition, these hotspots are enriched in bottleneck genes that may play a role in the diversity of phenotypes observed for these patients.

Acknowledgement: We thank Frank Honti for supplying the integrated functional network used in this study.

References

O13 How to avoid one thousand opportunities to do harm in genomic medicine

Isaac Kohane

Children's Hospital Boston, USA

BMC Proceedings 2012, 6(Suppl 6):O13

With the advent of whole genome sequencing made clinically available, the number of incidental findings is likely to rise. False positive incidental findings are of particular clinical concern, and they can usefully be classified into four categories. In order of increasing...
challenge, there is first, the substantial proportion of ‘textbook cases’ of mutations documented to cause human disease in a highly penetrant Mendelian fashion, which are incorrectly annotated in the databases. The second is the technical/measurement error rate in genome-scale sequencing. Third is the incorrect assignment of prior probabilities for much of our genetic and genomic knowledge. The fourth derives from testing multiple hypotheses across millions of variants. I will describe the nature of these components, provide rough estimates for the magnitude of the problem and point out existing approaches that will serve to control the growth of these aspects of the incidentalome.

O14 Analyzing genomes: is there a duty to disclose?
Amy McGuire
Baylor College of Medicine, USA
BMC Proceedings 2012, 6(Suppl 6):O14

The potential to discover unanticipated clinically significant genetic variants during the course of research has stimulated much debate about the obligations of researchers to communicate such findings to study participants. Published guidelines suggest that researchers have an obligation to communicate valid and significant results, especially if they are clinically actionable. I will present data on the practices and perspectives of investigators conducting genome-wide association studies (GWAS) and will caution against the establishment of a general obligation to return results in research. I will compare the role-specific responsibilities of researchers with those of clinicians and discuss emerging standards for genomic analysis and communication of test results in the clinical setting.

O15 Abstract not submitted for online publication.

O16 Interrogating of cancer genomes: towards more profile-based therapeutics
John David Carpten
Division of Integrated Cancer Genomics, Translational Genomics Research Institute, Phoenix, AZ, USA
BMC Proceedings 2012, 6(Suppl 6):O16

Although death rates due to common diseases such as heart disease, stroke and infectious diseases have declined over the last half-century, there has been only a small change in cancer mortality rates. The availability of the draft human genome sequence and a number of technological advances now provide us with the opportunity to explore the genomic landscape of cancer in an unprecedented way. Here, we will describe the application of multiple genomic technologies towards the interrogation of a number of cancer genomes, in order to discover molecular determinants of cancer that might be associated with clinical outcome and those that might be candidates for targeted therapy. It is our hope that these data would one day be translated into clinical practice to improve therapeutic decision-making for more knowledge-based clinical management.

O17 Abstract not submitted for online publication.

O18 Abstract not submitted for online publication.

O19 Abstract not submitted for online publication.

O20 Abstract not submitted for online publication.

O21 Abstract not submitted for online publication.

O22 MutaScope: a high-sensitivity variant caller dedicated to high-throughput PCR amplicons sequencing
Shawn yeast1, Hakkan Alakan1, Hiroko Matsumoto1, Kristen Jepsen1, Richard Schwab3, Kelly Frazier3,3, Olivier Harismendy2,3,1
1Bionformatics and Systems Biology Graduate program, University of California San Diego, La Jolla, USA; 2Division of Genome Information Sciences, Department of Pediatrics and Rady Children’s Hospital, University of California San Diego, La Jolla, CA, USA; 3Moore UCD Cancer Center, La Jolla, CA, USA; 4Department of General, Visceral and Cancer Surgery, Center for Integrated Oncology, University of Cologne, Germany; 5Clinical and Translational Research Institute, University of California San Diego, La Jolla, CA, USA; 6Institute for Genomic Medicine, University of California San Diego, La Jolla, CA, USA; 7Human Genome Sequencing Center, Harvard Medical School, Boston, MA 02115, USA; 8Institute for Genomic Medicine, University of California San Diego, La Jolla, CA, USA; 9Bioinformatics and Systems Biology Graduate program, University of California San Diego, La Jolla, USA

With the progress of genomics and targeted therapies, an increasing number of cancer somatic mutations are becoming clinically actionable: predictive of drug sensitivity or resistance. However, clinical samples are often suboptimal for their comprehensive detection. Indeed, contamination with normal cells or the presence of diverse subclones affects their detection by high-throughput sequencing. Ultra-deep targeted sequencing (UDT-Seq) is an assay combining microdroplet PCR amplification of exonic sequences followed by a direct, oriented sequencing at high depth of coverage, therefore allowing the detection of low prevalence mutations [1]. Standard sequencing analysis tools that were developed to analyze whole genome and exome shotgun sequencing do not take advantage of UDT-Seq’s specific design where each sequencing read originates from a known strand and location. We propose a complete analysis package (MutaScope) dedicated to the analysis of UDT-Seq or similar PCR-based high-throughput sequencing. After alignment to the genome, MutaScope separates the sequencing reads with respect to the amplicons and strand of origin. This allows the experimental measurement of an error rate along the amplicons, which is used to calculate a variant likelihood and rank candidate mutations. Using a set of reference samples, or matched normal DNA, MutaScope then identifies germline and somatic variants and reports them in a unified expanded variant call format. The performance of MutaScope was evaluated on 676 amplicons using a set of calibration samples harboring variants at defined prevalence down to 1%. Overall, MutaScope’s sensitivity and positive predictive value (PPV) were >96% and >75%, respectively; which is higher than the standard variant calling strategies (approximately 70% and 70%, respectively). MutaScope detects more than 73% of the variants with an alternate allele frequency ≤5%, while the other methods only detect 30% of the variants. MutaScope offers an analysis strategy specifically dedicated to the identification of low prevalence somatic mutations in high-throughput direct sequencing of PCR amplicons. As a result MutaScope increases the overall technical performance of such approaches that are currently being implemented in clinical diagnostics laboratories.

Reference

O23 Analysis of somatic retrotransposition in human cancers
Eunjung Lee1,2, Rebecca Iskow1, Lixing Yang1, Omer Gokcumen1, Psalm Hasley1,2, Lovelace J Luquette III1, Jens L Kohr1,2, Christopher C Harris4, Li Ding5, Richard K Wilson1, David A Wheeler1, Richard A Gibbs1, Raju Kucherlapati1, Charles Lee3, Peter V Kharchenko3, Peter V Park2
1Center for Biomedical Informatics, Harvard Medical School, Boston, MA 02115, USA; 2Division of Genetics, Brigham and Women’s Hospital, Boston, MA 02115, USA; 3Department of Pathology, Brigham and Women’s Hospital, and Harvard Medical School, Boston, MA 02115, USA; 4The Eli and Edythe Broad Institute, Cambridge, MA 02412, USA; 5Dana-Farber Cancer Institute, Boston, MA 02115, USA

Background: Close to half of the human genome is derived from transposable elements (TEs), and some TE families continue to generate new insertions through RNA-mediated mechanisms. Due to its mutagenic potential, such retrotransposition is normally suppressed by epigenetic and post-transcriptional mechanisms. However, the epigenetic and regulatory disruptions commonly observed in cancers may allow for TE activation, and a few examples have been reported in lung and colon cancer previously.

Materials and methods: To systematically evaluate the frequency of such events across different tumor types and assess their impact in human cancers, we developed Tea (Transposable element analyzer), a computational pipeline to detect TE insertions at single nucleotide level and extract their mechanistic signatures. We applied Tea to the high-coverage (>30x) tumor and matched normal genome pairs from 43 cancer patients across five tumor types as well as three healthy individuals.

Results: We identified 194 high-confidence somatic TE insertions (183 L1, 10 Alu, 1 ERV), most of which were generated through endonuclease-mediated retrotransposition mechanism. The novel L1 and Alu insertions were all found in the epithelial cancers (colorectal, prostate, ovarian), and none were detected in the examined blood or brain tumor samples. The somatic L1 insertions tend to occur in genes that are commonly mutated in cancer, and disrupt the expression of the targeted genes. To further illustrate the distinct genomic distribution of the somatic TE landing sites, we compared their placement with the 7,449 non-reference polymorphic TE insertions that we have identified from 44 normal genomes. The TE landing sites are strongly biased towards genomic regions that exhibit cancer-specific decrease in DNA methylation.

Conclusions: Our analysis illustrates the functional impact of somatic TE insertions and suggests resulting positive selection toward tumorigenesis.

O24 Exploring the cancer methylome
Benjamin P Berman1, Daniel J Weisenberger2, Toshinori Hinoue3, Houtan Noushmehr4, Yaping Liu4, Joseph F Aman4, Toshinori Hinoue2, Hui Shen4, Simeen Malik1, Swapna Mahurkar1, Timothy Triche Jr4, Zachary Ramjan1, Charles M Nicolet1, David Van Den Berg2, Leslie Cope2, Benjamin P Berman1, Stephen B Baylin2, Peter W Laird4
1USC Epigenome Center, University of Southern California, Keck School of Medicine, Los Angeles, CA, USA; 2Johns Hopkins School of Medicine, Baltimore, MD, USA; 3Department of Genetics, Harvard Medical School, Boston, MA 02115, USA; 4BMC Proceedings 2012, 6(Suppl 6):O23

Cancer develops not only as a result of genetic mutations and genomic rearrangements, but also as a consequence of numerous epigenetic alterations, including extensive changes in the distribution of DNA methylation throughout the genome. DNA methylation changes contribute directly to cancer by transcriptional silencing of tumor-suppressor genes such as promoter CpG island hypermethylation. Broad epigenomic analysis of human tumors can reveal relationships between large numbers of epigenetic events and can provide insight into the mechanisms underlying concerted epigenetic change. Genomic loci targeted by Polycomb Group Repressors in embryonic stem cells, and involved in cellular differentiation, are predisposed to aberrant DNA methylation in cancer cells, suggesting that an epigenetic block to cellular differentiation may sometimes be an initiating event in carcinogenesis. The very strong associations between distinct epigenetic subtypes, such as CpG Island Methylator Phenotypes (CIMP) and specific somatic genetic events, such as BRF4 mutation in colorectal cancer and IDH1 mutation in glioblastoma multiforme are consistent with an early role for DNA methylation alterations, providing a favorable cellular context for the subsequent somatic mutation. The analysis of whole methylomes at single-basepair resolution reveals that cancer-associated changes occur differentially across defined regions of the genome associated with the nuclear lamina. It is apparent that epigenomic analysis is essential for a full understanding of the relationship between alterations in the cancer genome and the origin and clinical diversity of individual tumors.

O25 The hunt for mammalian epialleles
Vardhman Rakyan
Barts and the London School of Medicine and Dentistry, UK
BMC Proceedings 2012, 6(Suppl 6):O25

Epialleles are genomic loci at which the epigenetic state can stably vary amongst individuals of a given population. Although first described and still best understood in plants, in recent years we have come to realize that epigenomic landscapes in mammals can also show considerable inter-individual variation. Such mammalian epialleles could arise through genetic influences, or have non-genetic origins as a result of stochastic events, environmental factors such as exposure to a compromised in utero environment, or adult lifestyle-associated factors such as smoking. My lab is currently pursuing several complementary lines of investigation that integrate molecular genetics and epigenomics in mouse models and human cohorts to understand the role of epialleles in complex phenotypes and diseases. In my talk, I will present a synthesis of the latest findings from several ongoing studies in my lab on epiallelic variation in mammals.

O26 Folding principles of genomes
Job Dekker
Program in Systems Biology, Department of Biochemistry and Molecular Pharmacology, University of Massachusetts Medical School, 364 Plantation Street, Worcester, USA
BMC Proceedings 2012, 6(Suppl 6):O26

My laboratory studies how chromosomes are organized in three dimensions. The three-dimensional organization of the genome is critical for regulating gene expression by bringing genes in close spatial proximity to distal regulatory elements such as enhancers. We have developed powerful molecular approaches, based on our Chromosome Conformation Capture technology, to determine the folding of genomes at unprecedented resolution (Kb) and scale (genome-wide). We have applied these methods to determine the spatial folding of 1% of the human genome (the ENCODE pilot regions) across a panel of cell lines. We discovered that chromosomes fold into extensive long-range interaction networks in which genes are interacting with distal gene regulatory elements. These results start to place genes and regulatory elements, that are often separated by large genomic distances, in three-dimensional context to reveal their functional relationships.

Our analysis of chromosome folding also revealed that chromosomes are compartmentalized in a series of "Topological Association Domains" (TADs) that are hundreds of Kb in size. Loci located within a TAD mingle freely, but interact far less frequently with loci located outside their TAD. TADs appear involved in gene expression, as we found that genes located within the same TAD tend to be co-expressed, but the mechanism(s) by which these domains affect gene regulation is still unknown. TADs represent novel universal and genetically encoded building blocks of chromosomes.

O27 Ultra-high resolution mapping of protein-genome interactions using ChIP-exo
B Franklin Pugh
Penn State University, Biochemistry and Molecular Biology, University Park, PA 16802, USA
BMC Proceedings 2012, 6(Suppl 6):O27

With the advent of high-throughput and high resolution genome-wide protein-DNA detection assays, the interrelationships between
chromatin and the transcription machinery are now becoming clearer. Here I will discuss our recent findings using MNase ChiP-seq to map nucleosome positions, and a novel ultra-high resolution mapping technique called ChiP-exo that we recently developed [1]. What is apparent from these studies is the following: firstly, transcription factors bind to many more locations in the genome than previously appreciated. Secondly, PICs form at the interface between nucleosomes and nucleosome-free promoter regions. Finally, Chromatin remodeling complexes target specific nucleosome positions, working in concert to organize nucleosomes at the beginning and end of genes. Many remodelers, including the SWI/SNF complexes, interact asymmetrically with the nucleosome core across the genome, which may be important for the directional passage of RNA polymerase II.

Reference
1. Rhee HS, Pugh BF: Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution. Cell 2011, 14:1408

O28 Epigenetic reprogramming in the epithelial-to-mesenchymal transition
Stephen Hoang1, Marcin Cieślak1, Sanjay Chodaparambil, Natalya Baranova, Manish Kumar, David Allison, Jake Wamsley, Lisa Gray, Marty Mayo, Stefan Bekiranov
Department of Biochemistry and Molecular Genetics, University of Virginia School of Medicine, VA, USA
*Equal contributors
BMC Proceedings 2012, 6(Suppl 6):O28

The epithelial-to-mesenchymal transition (EMT) is a cellular dedifferentiation process that is critical to development, wound healing and metastasis. Like other cell state transitions, such as differentiation, EMT is accompanied by genome-wide epigenetic reprogramming. However, the relationship between reprogramming and functional changes in the cell is poorly understood. In an A549 non-small cell lung cancer EMT model system we observed changes in chromatin state between epithelial and mesenchymal states. Multivariate analyses were applied to paired (epithelial and mesenchymal) ChiP-seq data for 18 histone modifications/variants and expression microarray data. We observed epigenetic co-regulation of genes associated with EMT, as well as their proximal enhancers. We also observed epigenetic activation or repression of functionally distinct sets of enhancers. These genes and enhancers are regulated and bound by a small set of transcription factors, specifically AP-1, NF-kB and c-Myc. These transcription factors themselves also show an epigenetic profile similar to the EMT-related genes. Together, these observations suggest a chromatin-mediated transcriptional feedback mechanism that establishes and maintains the phenotypic switch.

O29 Development of a computational strategy to compare repetitive element enrichment between experimental conditions from high-throughput sequencing datasets
Steven Criscione1, Nicola Neretti1,2
1Molecular Biology, Cell Biology, and Biochemistry, Brown University, 75 Waterman Street, Providence, RI 02912, USA; 2Center for Computational Molecular Biology, and Biochemistry, Brown University, Providence, RI 02912, USA
BMC Proceedings 2012, 6(Suppl 6):S29

Repetitive and transposable elements comprise more than half of the human genome and play diverse roles in many biological processes. Mobile elements including retrotransposons are implicated in the organization of the epigenetic landscape, the progression of tumorgenesis, and the enhancement of genetic diversity. Despite the importance of repetitive and transposable elements these sequences are traditionally ignored in high-throughput sequencing analysis due to the technical difficulty of uniquely mapping reads from repeat DNA sequences. Here we report a new computational method for the analysis of repetitive elements from high-throughput sequencing datasets that accounts for all mapping reads. In our approach, we examine reads that map uniquely and to multiple locations of the genome using two separate strategies to determine a complete estimate of enrichment for repetitive elements. Included in our computational method is an output defined by reads per kilobase of repeat element per million mapped reads (similar to RPKM definition for the exon model) [1]. The calculated repeat element enrichment RPKM allows for the comparisons between repetitive elements as well as between experimental conditions. Our new method for examining repetitive elements from high-throughput sequencing datasets represents an improvement over existing methods because we do not exclude reads from the analysis and we can make comparisons between experimental conditions. To test our method we have examined repetitive element enrichment in the embryonic and adult mouse across different tissues using a variety of high-throughput mouse sequencing datasets available from the mouse ENCODE project and Shen et al. that provide a thorough snapshot of the epigenetic landscape of the embryonic and adult mouse [2]. We compare our method with an existing strategy for estimating repetitive element enrichment proposed by Day et al. [3], and demonstrate the advantages to our strategy. In addition, we test the robustness of our approach for determining differences in enrichment between experimental samples by conducting a comparison between the embryonic and adult mouse.

References

O30 Towards a patient-based drug discovery
Stuart L. Schreiber
Broad Institute of Harvard and MIT, USA
BMC Proceedings 2012, 6(Suppl 6):O30

Small-molecule drugs were originally discovered using compound-based drug discovery: opportunistic discovery of a biologically active compound, often a natural product (such as penicillin) followed by a search for a disease that might be treated with the compound. This remains a common approach to modern drug discovery (for example, rapamycin and analogs for use as antifungal agents, immune suppression agents, anti-cancer agents, and possibly others in the future).

The advent of recombinant DNA accelerated a second approach – target-based drug discovery – where the therapeutic target is selected and subjected to methods that yield candidate drugs (mechanism-based design, structure-based design, screening). But this approach has its shortcomings – 97% of drug candidates that enter into clinical investigation eventually fail, many due to unanticipated toxicity and many others due to a lack of efficacy despite successful modulation of the target. Selecting therapeutic targets based on information derived from surrogates of patients has proved challenging. Advances in human biology, including human genetics and physiology, and in small-molecule science, including chemistry and chemical biology, are now accelerating a third approach: patient-based drug discovery. I will present examples that aim to use: 1) information from heritable or somatic human genetics in human disease, for example, in Crohn’s disease and cancer, 2) advances in diversity-oriented synthetic chemistry and chemical biology to accelerate the discovery of safe and effective small-molecule therapeutics, and 3) an understanding of the relationship of human genetic variation with drug efficacy.
POSTER PRESENTATIONS

P1
Life stage and tissue speciation of cathepsin B (AGAP004533) derives different functional properties in the G3 strain of the mosquito Anopheles gambiae

Daniel Archinko1, Dan Masiga1, Paul Mirej2, Flaminia Catteruccia1
1Molecular Biology and Biotechnology Department, International Center of Insect Physiology and Ecology, Nairobi, Kenya; 2Biochemistry Department, Egerton University, Njoro, Kenya. PO Box 20115-336.

Abstract: The mosquito Anopheles gambiae, a primary vector of human malaria, has two G3 strains, K1 and Sgam. Here, we focus on the K1 strain, since the Sgam strain is host specific and not found in the USA. This study aimed to identify differences between the K1 and Sgam strains. We investigated whether these differences are related to mosquito sex or mosquito tissue. We analyzed the expression of the cathepsin B gene (AGAP004533) in all reproductive tissues (ovary, spermatheca, testes, and accessory glands) of the male and female mosquitoes of each strain. We observed significant differences in the expression and translation of this gene between the strains and the sexes. In the male, the K1 strain had higher expression and translation of the gene, while in the female, the Sgam strain had higher expression and translation. These results suggest that the cathepsin B gene plays a role in the reproductive process of the mosquito and that the K1 strain may be more suitable for use in malaria control.

Reference:

P2
Abstract not submitted for online publication.

P3
Targeted next-generation sequencing of colorectal cancer identified metastatic specific genetic alterations

A Rose Branzino1, Ersieva Vakani1, Sasinya Scott1, Brooke Sylvester1, Krishan Kania1, Agnes Viale1, David Solit2, Michael Berger1
1Department of Pathology, Memorial Sloan Kettering Cancer Center, New York, NY 10065, USA; 2Human Oncology and Pathogenesis Program, Memorial Sloan Kettering Cancer Center, New York, NY 10065, USA; 3Genomics Core Laboratory, Memorial Sloan Kettering Cancer Center, New York, NY 10065, USA

Abstract: We conducted a targeted next-generation sequencing (NGS) study of colorectal cancer (CRC) metastases and primary tumors to identify specific genetic alterations that are unique to metastatic CRC. We sequenced a panel of 230 cancer genes in 43 primary CRC tumors and 17 matched metastatic tumors from 15 patients. We identified 838 cancer-susceptible mutations, indels, and copy number events in the primary tumors and 121 events in the metastases. We found 63 mutations, indels, and copy number events specific to metastatic tumors, with 11 being unique to both the primary and metastatic tumors. These alterations were enriched in pathways associated with cell death and proliferation, and were more likely to cause missense mutations. Our findings suggest that targeted NGS can effectively identify specific genetic alterations that are unique to metastatic CRC, and may have implications for treatment decisions.

Reference:

P4
Moving beyond gene expression: identification of lung-disease-associated novel transcripts and alternative splicing by RNA sequencing

John Brothers1, Rebecca Kusko2, Gang Liu3, Lingqi Luo4, Brenda Juan Guardela4, John Tedrow4, Yuriy Alekesyev2, Ivana V Yang5, Mick Cornell6, Mark Geraci7, John Qajackenbush8, Paola Sebastiani1, Marc Lenburg1,2,3,5, Naftali Kaminski4, David A Schwartz6, Yuriy Alekesyev2,5, Ivana V Yang6, Paola Sebastiani1, Marc Lenburg1,2,3,5, Naftali Kaminski4, David A Schwartz6, Aurum Spira1,2,3, Jennifer Beane1
1Bioinformatics Program, Boston University, Boston, MA 02115, USA; 2Genetics and Genomics Graduate Program, Boston University School of Medicine, Boston, MA 02118, USA; 3Section of Computational Biomedicine, Department of Medicine Boston University School of Medicine, Boston, MA 02118, USA; 4Simmons Center for Interventional Lung Disease and Department of Medicine, University of Pittsburgh Medical Center, Pittsburgh, PA 15233, USA; 5Department of Pathology and Laboratory Medicine, Boston University School of Medicine, Boston, MA 02118, USA; 6Center for Genes, Environment and Health and Department of Medicine, National Jewish Health, Denver, CO 80206, USA; 7Dana-Farber Cancer Institute and Harvard School of Public Health, Boston, MA 02115, USA

Abstract: We used RNA sequencing to identify novel transcripts and alternative splicing events associated with lung disease. We sequenced RNA from 45 lung tissues, including normal and diseased tissues, and from 15 different lung disease types. We identified 146 novel transcripts and 23 alternative splicing events that were specifically associated with lung disease. These events included transcription factor predictions, secondary structure modifications, and extensions at the carboxylic end of the protein. The latter translates into a glutamine for leucine (Q6L), which eventually affects its 3D conformation. The sequences in the N-terminal region of the protein sequence, meanwhile, the former translates into a proline for a leucine (P337L) substitution, which results in secondary structure modification. The sequence for the N-terminal region of the protein sequence, meanwhile, the former translates into a proline for a leucine (P337L) substitution, which eventually affects its 3D conformation.

Reference:
Background: Chronic lung diseases affect a significant portion of the population, and the incidences of chronic obstructive pulmonary disease (COPD)/emphysema and idiopathic pulmonary fibrosis (IPF) are increasing. COPD is the fourth leading cause of death in the USA and the incidence of IPF has doubled over the past decade. Identification of novel transcripts and transcript isoforms (alternative splicing patterns) associated with these diseases may help us better understand their molecular pathogenesis, and identify both novel disease-specific biomarkers and therapeutic targets.

Materials and methods: Using lung tissue sections from the NHLBI Lung Tissue Research Consortium, we sequenced the mRNA (75 or 99 nt paired-end sequencing; Illumina GAIIx or HiSeq) from 145 lung tissue samples that were subsequently split into an initial training cohort of 89 samples and an independent filtering set of 56 samples. Genome-guided transcriptome reconstruction using Cufflinks was performed on the training and independent filtering set. A final conservatively filtered dataset of 257 samples was generated using TopHat 2.0.1 and PCR-Trim from each of the 145 individual samples. We then used an independent filtering set of 56 samples to further filter the data. A final conservatively filtered dataset of 257 samples was generated using TopHat 2.0.1 and PCR-Trim from each of the 145 individual samples. We then used an independent filtering set of 56 samples to further filter the data.

Results: The filtered transcriptome assembly (overlap set) is more similar to known genes (based on comparisons with Ensembl) than the initial training and independent filtering set. A set of 38 novel gene candidates were selected based on gene structure parameters computed from Ensembl annotation. Differential expression (DE) analysis was performed, and 39 similar known genes were DE in both diseases. Several examples of disease-associated differential splicing were also identified. These new disease-associated isoforms are further being investigated to identify their biological function and relevance to COPD and IPF.

Conclusions: RNA-Seq of a large number of lung tissue samples has allowed us to identify novel disease-associated genes and alternative splicing patterns that may contribute to our understanding of the pathogenesis of IPF and COPD.

Although the FDA drug label recommends testing for TPMT deficiency prior to dosing and the PharmGKB CPIC group published a guideline [2] with a recommended dosing strategy and interpretation, testing is not universal because these guidelines are difficult to translate into a clinical decision support (CDS) system and integrate with the EHRs. We developed models and specifications to execute PGx CDS rules based on a patient’s genotype. Rules are modeled at four levels of abstraction: (1) unstructured (narrative), (2) semi-structured, (3) structured, and (4) executable.

As genomic sequencing becomes routine, standardized methods to interpret the data and make clinical decisions are paramount. In conjunction with the BCH DNA Diagnostic Laboratory, we streamlined the TPMT testing process to fit into the usual clinical routine (including ordering, testing-in-house and return of results to the clinician). We consolidated all genetic sequencing testing into a single clinical workflow (blood to report) that is run, analyzed and interpreted in a Clinical Laboratory Improvement Amendments (CLIA) certified laboratory using the codified CDS rules. The interpretation reports are generated automatically directly from the genotype calls and then manually reviewed for accuracy. Once cleared by the laboratory director, the reports are uploaded into the EHR (Cerner). Specialty flow sheets enable providers to easily view the allele status and interpretation report. We intend to expand the PGx platform to include additional drug/gene pairs.

References
in the 5’ start position (isomiRs). Negative binomial generalized linear models were used to identify microRNAs and isomiRs differentially expressed between phenotypes. The large RNA from a subset of these samples was also sequenced using the Illumina GAIIx (one sample per lane). A network was generated by identifying all microRNA-mRNA pairs that were significantly anti-correlated (Spearman, false discovery rate (FDR) <0.05) and had a predicted microRNA-binding site in the mRNA 3’ UTR (TargetScan 6.0).

Results: An average of 26.3 million and 7.1 million reads were sequenced per sample using the singleplex and multiplex protocols, respectively. An average of 73% of reads per sample aligned to the human genome with one or fewer mismatches at 10 or fewer locations. 287 novel microRNA precursors were predicted using the miRDeep algorithm. One of these candidates was validated by quantitative RT-PCR and found to be expressed across a range of human tissues. The expression of 309 canonical microRNAs was significantly different between patients with disease and controls (FDR <0.05), in addition, we found that 242 isomiRs from 159 different microRNA loci were also differentially expressed (FDR <0.05). We developed a network of microRNA-mRNA interactions by integrating small RNA and large RNA sequencing data generated on the same samples (n = 72). 2,133 genes in the network (65%) were predicted to be regulated by at least one type of small RNA. All canonical forms of miR-338-3p were significantly downregulated in ILD. Predicted targets of the canonical form of miR-338-3p were enriched in extracellular matrix genes while the predicted targets of a miR-338-3p isomiR were enriched in the Wnt signaling pathway (P<0.001), suggesting different roles for multiple forms of this microRNA in ILD.

Conclusions: Our results demonstrate the power of deep sequencing to reveal additional complexity in the microRNA transcriptome, such as novel microRNAs and isomiRs. The disease-related patterns of microRNA expression can provide insights into the molecular pathogenesis of chronic lung diseases and novel targets for therapy.
calculated using Haploview software [2]. Subsequently, a gene-based test statistic which adjusted for the LD across the SNPs was calculated using VEGAS [3].

Results: Of 26 LRP1 SNPs, 11 were significantly associated with the change in chylomicron concentration after a false discovery rate correction for multiple testing (Q<0.05), across two haplotype blocks. The subsequent gene-based test, corrected for LD and multiple testing, was also significant ($P = 0.01$).

Conclusions: These results implicate the role of LRP1 in postprandial lipoprotein uptake and/or clearance. Given the role of chylomicron clearance in subsequent fat accumulation, if these results are replicated, this information may eventually help tailor dietary advice, aimed at reducing BMI, in the pursuit of personalized medicine paradigm in the treatment of obesity.

Acknowledgements: We are grateful to the staff of the GOLDN study for the assistance in data collection and management.

References

**P9**

**Genome-wide analysis of primate and rodent protein-coding and associated non-coding nucleotide sequences**

Sudhindra R Gadagkar1, Luke A Rond2

1Biomedical Sciences Program, Midwestern University, Glendale, AZ 85308, USA; 2Arizona College of Osteopathic Medicine, Midwestern University, Glendale, AZ 85308, USA

**BMC Proceedings 2012, 6(Suppl 6):P9**

**Background:** Several mammalian species have been characterized by means of genome-wide analysis of the protein-coding sequences, but this has not been done in conjunction with the associated non-coding sequences, including regulatory regions.

**Materials and methods:** We obtained gene data (coding sequences, 5’ and 3’ UTRs, intron sequences, and 5,000 bases of the 5’ and 3’ flanking regions) from Ensembl (http://wwwensembl.org) after determining the Ensembl IDs from the online database InParanoid7 (http://inparanoid.sbc.su.se) for all known orthologs among four mammalian species (two primate and two rodent): human (Homo sapiens), chimpanzee (Pan troglodytes), mouse (Mus musculus) and rat (Rattus norvegicus).

**Evolutionary analyses** were done using in-house computer programs or by means of the program MEGA-CC [1]. Homogeneity of the nucleotide substitution pattern between species was tested using the Disparity Index test [2], and selection tests were done using the z-test for coding sequences and Tajima’s D [3] for non-coding sequences.

**Results:** There was a total of 16,511 error-free sets of orthologs containing human genes, of which 7,244 were orthologous among all four species. A very small number (23, approximately 0.32%) of these four-way orthologs were determined to be undergoing adaptive evolution in the primate lineage. A majority of them (approximately 71%) were found to be evolving neutrally, with the rest (approximately 29%) were determined to be under purifying selection. All of the 23 genes under positive selection in the primate lineage are under strong purifying selection when compared with the orthologs of both of the rodent species. On average, these genes show a lower G+C content (compared with the A+T content) in all four species, but especially in the primates. In contrast, the genes under negative or neutral selection show a high G+C content. Interestingly, while 327 genes were found to be evolving with a heterogeneous nucleotide substitution pattern between human and chimpanzee, only two of them are under positive selection, while 140 are under purifying selection, and 185 are evolving neutrally. Furthermore, as many as 21 genes are under positive selection, even though they are evolving with a homogeneous substitution pattern. We discuss these results and others, and compare them with those from the non-coding regions.

**Conclusions:** Our work compares the evolution of coding sequences across four mammalian genomes (two primate and two rodent), and adds perspective to the results by means of comparisons with the associated non-coding sequences.

**References**

**P10**

**Exome sequencing combined with semantic discovery identifies strong disease-associated candidates in a single case of relapsing remitting multiple sclerosis**

Mahjoubeh Jalali Seidh Dashti1, Maritha Kotte2, Susan Janse van Rensburg1, Alan Chrostoffels1, Juraid Gamjatjien1

1South African National Bioinformatics Institute, University of the Western Cape, Bellville, 7535, South Africa; 2Department of Pathology, University of Stellenbosch, Tygerberg, 7505, South Africa

**BMC Proceedings 2012, 6(Suppl 6):P10**

**Background:** As known disease-associated variants identified through large cohort-based studies often explain only a small percentage of genetic risk in multifactorial disorders such as multiple sclerosis (MS), alternative methods for identification and prioritization of variants that directly and/or indirectly play a role in disease development have become increasingly important. We were tasked with identifying possible genetic causes in a case of atypical relapsing remitting MS (RRMS) that also presented with porphyria-like symptoms and where demyelination was halted in the patient upon iron supplementation. As the patient had no parents or siblings that could be used as references for filtering exome variants, we aimed to develop a new prioritization strategy based on the combination of a predicted deleterious effect on the protein and existing knowledge of the biological roles of the genes and their contribution to relevant phenotypes.

**Materials and methods:** Exome sequencing was performed and functional SNP analysis and frameshift indel detection were carried out using a combination of prediction tools. Predicted deleterious variants were further assessed with respect to their possible involvement in MS using our internal biomedical semantic database, the B.O.R.G. (BioOntological Relationship Graph), which integrates existing biomedical knowledge and uses path-based graph theoretic querying to discover links between biological concepts. The semantic model used in this study incorporated known human, mouse and rat gene, gene-to-disease, gene-to-phenotype, gene-to-function and gene-to-pathway and orthology relationships in conjunction with ‘surrogate’ phenotype and biological function links to the relevant disease term.

**Results:** The exome sequence resulted in identification of 64,890 variants based on the human reference genome, of which 4,847 missense variants were predicted to have a damaging effect. These were used to interrogate the semantic network, which simplified exploring the network for transitive relationships that may explain the biological contribution of identified mutations to the development of disease, either directly from human evidence or transitively via model organism evidence such as knockout phenotypes. 750 variants were found to be potentially involved in MS based on known gene-to-disease associations or via surrogate-phenotype, -function or -pathway links to MS. Examples of strong candidate MS genes identified were: BACE1, HEXB and NRCAM (involved in myelination), CNTN2 (involved in axon development), and CCM2 (implicated in immune response and central nervous system inflammation). In addition, we identified deleterious variants in IREBP and CYBBD1, which are involved in iron regulation and homeostasis, and may contribute to the iron deficiency condition of the patient. Several other strong candidates are being evaluated further.

**Conclusions:** While the list of candidates obtained from this study is quite large, a number of them may in fact represent a part of
the large number of variants proposed to be associated with the ‘missing’ (approximately 80%) MS genetic risk, with each adding a tiny percentage to the overall risk of developing the disease. As these variants fulfill many criteria, our knowledge-driven prioritization strategy may appear to have the potential to improve discovery of causative variants in non-Mendelian diseases and also rare diseases where large cohorts are almost impossible to build.

P12

Abstract not submitted for online publication.

P13

Discovering epigenetic changes in response to tungsten-alloy treatment using next-generation sequencing technologies

Ranjana Verma, Deepak Grover, Zygmunt Galdzicki

Department of Anatomy, Physiology and Genetics, School of Medicine, Uniformed Services University of the Health Sciences, 4301 Jones Bridge Road, Bethesda, MD 20814, USA

BMC Proceedings 2012, 6(Suppl 6):P13

Background: Heavy metals used in industrial and household applications can pose harmful health effects. Tungsten alloys (WAs) have been widely used in many engineering, automotive and marine applications. In military operations, WAs containing tungsten (W; 91% w/w), nickel (Ni; 6% w/w) and cobalt (Co; 3% w/w) have been deployed in armor-penetrating munitions as substitutes for depleted uranium. Despite widespread use, a clear understanding of the potential effects of WAs on physiological processes and gene expression is not available. WAs have carcinogenic potential as demonstrated by cancer development in rats with intramuscular implanted WA pellets. This suggests a potential involvement of epigenetic events previously implicated as environmental triggers of cancer. In the present study, we have attempted to unravel WA-mediated alterations in gene expression and gene-specific epigenetic modifications at the genome-wide level.

Materials and methods: C2C12 (mouse myoblast) cell cultures were exposed to 1,000 μM WA for 24 h. ChIP assays were performed with cross-linked cells from WA-treated and untreated C2C12 cultures using anti-RNA Pol II, anti-phospho histone H3S10, anti-trimethyl histone H3K4 and anti-trimethyl histone H3K27 antibodies. ChIP assays were followed by sequencing where single-ended 50 bp reads were generated using a combination of the Applied Biosystems SOLiD™ and Illumina Hi-Seq™ systems. Data analysis was done using in-house developed pipelines (PERL/Unix) as well as several open source bioinformatics software from R/Bioconductor project and other sources.

Results: Using different in vitro models, we examined metal-induced cytotoxicity and epigenetic modifications where WA showed cytotoxicity at concentrations >50 μg/ml, with C2C12 being relatively resistant to WA-mediated toxic impact. Using ChIP-Seq, we found several histone modifications up- and downregulated in the promoter regions of genes related to learning and memory mechanisms, with maximum impact observed for H3S10 phosphorylation. A total of 101 regions in the mouse genome were found to be most significantly depleted of H3S10 phosphorylation after WA treatment (fold change >3), these targets included several genes with neurological functions including voltage-dependent calcium channel (CACNB1), phosphatidylinositol 4-kinase type 2 (PHKA2a) and Kinesin 5A (KIF5A). In addition, pathway analysis of these 101 regions revealed 15 genes as part of regulatory networks that are responsible for developmental disorders, hereditary disorders and neurological diseases. ChIP-Seq analysis of other histone modifications and Pol II binding patterns are still ongoing. We plan to correlate expression levels investigated by RNA-Seq and epigenetic profiles impacted by WA exposure.

Conclusions: Our results reveal epigenetic modifications triggered by WA exposure in C2C12 cells for the first time at a genome-wide level. In addition to epigenetic changes observed for specific genes, ChIP-Seq analysis confirmed our previous report on the gross genomic depletion of H3S10 phosphorylation [1]. Future investigations on genes identified in this study will help unravel the mechanisms involved in WA toxicity that may lead towards the development of therapeutics.

References


P14

An integrative genomics approach for mapping the emerging genetic susceptibility landscape of triple-negative breast cancer

Ching-Chih Huang1, Antonio Pannuti1, Karolin Bakucz1, Lucia Mile1,2
1Cancer Institute, University of Mississippi Medical Center, 2500 N. State Street, Jackson, MS 30921, USA
2BMC Proceedings 2012, 6(Suppl 6):P14

Background: Genome-wide association studies (GWAS) have achieved great success in identifying common variants associated with increased risk for developing breast cancer. More recently, advances in next-generation sequencing (NGS) have made possible identification of mutations associated with breast cancer. However, to date, the information generated by GWAS and NGS has not been maximally leveraged and integrated with gene expression data to identify biomarkers associated with the most aggressive subset of breast cancer: the triple-negative breast cancer (TNBC). Here we present results from an integrative genomics approach that combines GWAS and sequence information with gene expression data to identify functionally related genes and biological pathways enriched for expression-associated genetic loci and mutations associated with TNBC using publicly available data.

Materials and methods: We used publicly available data derived from 60 GWAS involving over 400,000 cases and over 400,000 cancer-free controls to identify SNPs and associated genes with increased risk of developing breast cancer. Specifically, we first identified SNPs in population-based human cohorts that are associated with the expression of genes in TNBC. Mutations and associated genes were identified by mining publicly available RNA-Seq and whole exome/-genome sequencing data derived from 104 TNBC patients. Gene expression data were from 124 TNBC tumors and 142 cancer-free controls. We performed supervised and unsupervised analysis on gene expression data from genes containing genetic variants and mutations to identify functionally related genes. Additionally, we performed pathway prediction and network modeling using Ingenuity. For each predicted pathway and network, we counted the number of SNPs and mutation events by direct enumeration.

Results: We identified 600 SNPs mapped to 205 genes, and 250 genes with mutations that included single nucleotide polymorphisms (SNPs) and copy number variants. Hierarchical clustering revealed functional relationships and similarity in patterns of expression profiles between SNP-containing genes and genes containing mutations. We identified multi-gene biological pathways enriched for SNPs and mutations. Many of the pathways identified have been proposed as important candidate pathways for TNBC, including the p53, NFκB, apoptosis, BRCA, DNA repair and DNA mismatch repair pathways.

Conclusion: The results provide convincing evidence that integrating GWAS and sequence information with gene expression data provides a unified and powerful approach for biomarker discovery in TNBC. Furthermore, the results provide insights about the broader context in which genetic variants and mutations operate in TNBC.

P15

Bernoulli mixture models in application to the evaluation of algorithms estimating functionality of missense mutations

Stephanie Hicks1, Sharon E Plon2, Marek Kimmel1
1Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan; 2Institute of Biotechnology in Medicine, National Yang-Ming University, Taipei, Taiwan
BMC Proceedings 2012, 6(Suppl 6):P15

Background: Whole genome and whole exome sequencing projects yield a large number of missense mutations, many of which are likely to affect protein function. Direct estimation of the sensitivity and specificity of bioinformatic algorithms predicting the impact of missense mutations on protein function requires a ‘gold standard’ or set of mutations with known functionality. In the absence of a gold standard, additional statistical methods are needed to estimate the accuracy of these algorithms. It has been shown informative predictions depend on the algorithm and sequence alignment employed and often algorithms disagree as to which mutations are predicted deleterious or neutral [1].

Materials and methods: To investigate the level of agreement, disjoint categories of sets of mutations are defined depending on which algorithms predict which mutations to be deleterious or neutral. We have developed two statistical models called Bernoulli mixture (BM) and augmented Bernoulli mixture (ABM) based on the capture-recapture technique that employs these disjoint categories. Application of these models allows us to jointly estimate the sensitivities and specificities of each algorithm considered without the use of a gold standard and to estimate the proportion of deleterious mutations in a given set. These estimates may then be used to calculate the posterior probability of a given variant being deleterious. When considering n algorithms, there are 2n disjoint categories employed by the ABM model, which includes 2n + 3 parameters, and the BM model is a special case of the ABM model that includes 2n + 1 parameters. We use the expectation-maximization algorithm for parameter estimation.

Results: We apply the models to two types of predictions of functionality: simulated and real predictions. Using simulated predictions, we accurately recover the true sensitivity and specificity values and report confidence regions. We show example posterior probabilities of a given variant being deleterious. When a gold standard is available, we show the sensitivity and specificity estimates reported the BM and ABM models closely match the sensitivity and specificity estimated directly using the true functionality status. To test our models on mutations without known functionality, we apply the models to mutations obtained from the exomes of four individuals which were sequenced at the Human Genome Sequencing Center at Baylor College of Medicine to identify cancer susceptibility genes for acute lymphocytic leukemia and lymphoma in children. Within each individual, we estimate posterior probabilities for each variant being deleterious and apply an intersection filter to look for deleterious mutations shared by the three affected individuals, but not in the unaffected individual.

Conclusions: The BM and ABM models may be used to estimate the sensitivity and specificity of algorithms predicting the functionality of mutations without the use of a gold standard and to calculate posterior probabilities of a given variant being deleterious which may be used downstream in application of finding causal variants in next-generation sequencing.

Acknowledgements: Supported by CPRIT grant R83940, NCI grant CA155767 and NCI T32 training grant CA096520.

Reference

P16

In silico drug screening and potential target identification for hepatocellular carcinoma using support vector machine

Wu-Lung R Yang1, Yu-En Lee2, Ming-Huang Chen3, Yu-Wen Liu4, Pei-Ying Lee5, Kun-Mao Chao4,5, Chi-Ying F Huang6
1Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan; 2Institute of Biotechnology in Medicine, National Yang-Ming University, Taipei, Taiwan; 3Institute of Biopharmaceutical Sciences, National Yang-Ming University, Taipei, Taiwan; 4Division of Hematology and Oncology, Department of Medicine, Taipei Veterans General Hospital, Taipei, Taiwan; 5Graduate Institute of Biomedical Electronic and Bioinformatics, National Taiwan University, Taipei, Taiwan; 6Institute of Biopharmaceutical Sciences, National Yang-Ming University, Taipei, Taiwan
BMC Proceedings 2012, 6(Suppl 6):P16

Hepatocellular carcinoma (HCC) is a severe liver malignancy with few drug treatment options. Drug screening using FDA-approved drugs will provide a fast track in clinical trials if drugs are found effective against HCC. The Connectivity Map (cmap), a large repository of chemical-induced gene expression profiles, provides the opportunity of analyzing drug property with the expression. A support vector machine (SVM) was utilized to classify the effectiveness of drugs against HCC using gene expression profiles in cmap. The classification results will help us to identify significant chemical-sensitivity genes.
and to predict the effectiveness of remaining chemicals in cmap, with a prioritized listing for biological verification. The cell viability of four HCC cell lines treated with 146 chemicals was conducted. The SVM successfully classified the effectiveness of chemicals with an average area under the receiver operating curve of 0.9. Chemical sensitivity genes which are possible HCC therapeutic targets, such as MT1E, MYC and GADD45B, were identified with opposite signs of gene differential changes compared with reported HCC patient samples. Several known HCC inhibitors, such as geldanamycin, alvespimycin (histone deacetylase inhibitors) and doxorubicin (chemotherapy drug), were predicted to be effective. Seven out of 23 predicted drugs were cardiac glycosides, suggesting a close link of these drugs to the inhibition of HCC. The study demonstrates a strategy of in silico drug screening using a large repository of microarrays based on initial in vitro drug screening results. The biological verification result can serve as a feedback into the process for the development of a more accurate chemical sensitivity model.

In addition to somatic mutation data, we have integrated the data from the Genomics of Drug Sensitivity in Cancer Project [http://www.cancerrxgene.org], which is screening a wide range of anticancer drugs against over 1,000 genotypically characterized human cancer cell lines. Data analysis is becoming increasingly challenging due to the rapid expansion in cancer genome sequencing capacity. COSMIC is a major cancer genetics resource aiming to help such investigations, providing a centralized somatic mutations database with a wide suite of tools for its examination.

P17 Mining cancer genomes in COSMIC

Mingming Jia, Simon A Forbes, David Beare, Nidhi Bindal, Prasad Burasekaran, Kenric Leung, Chai Yin Kok, Sally Bamford, Charlotte Cole, Sari Ward, Jon Teague, Michael R Stratton, Peter Campbell, Andrew Futreal

Cancer Genome Project, Wellcome Trust Sanger Institute, Hinxton, Cambridge, CB10 1SA, UK

BMC Proceedings 2012, 6(Suppl 6):P17

COSMIC, The Catalogue Of Somatic Mutations In Cancer [http://cancer.sanger.ac.uk], is one of the largest repositories for somatic mutational events in human cancer. Data in COSMIC are curated from multiple sources, including from over 14,310 scientific publications, and are along-side data from the Cancer Genome Project at the Sanger Institute and global international consortia, such as The Cancer Genome Atlas and the International Cancer Genome Consortium. The COSMIC database currently accommodates over 300,000 mutations across 750,000 analyzed samples from 21,850 genes (COSMIC v60, July 2012). The Cancer Gene Census [http://cancer.sanger.ac.uk/cancergenome/projects/census/] is a list of almost 500 known cancer genes for which mutations have been identified as causally implicated in cancer. These genes are prioritized for full literature curation.

The collection of whole exome and genome sequencing data in COSMIC continues to grow at a rapid pace. There are: 17,614 coding mutations, 84,747 non-coding variants in 396 whole genome screens; 121,619 coding mutations and 12,949 non-coding variants as result of full exome sequencing; 3,512 structural mutations derived from 77 rearrangement screens. The data overview for each whole genome screen is presented using Circos, for example, the NCI-H209 Circos summary [http://cancer.sanger.ac.uk/cosmic/sample/overview?id=688013].

Analyzing information from whole genome sequencing can greatly enhance the chance of discovering novel genes implicated in human cancer. Unlike hot spot screening of gene regions where somatic mutations are most frequent, the use of whole genome data can identify all mutations in all genes, providing much more expansive annotations to recurrence analysis as used to discover new cancer genes. For instance, there are recurrent somatic mutations identified in genes, for example: SPOC in 19 prostate samples; SDK1 in 20 large-intestine samples.

There are several ways to access and analyze the data in COSMIC. The website allows data viewing in a genomic context supported by GBrowse while maintaining our gene-centric perspective. New additional features include a filter for excluding identified SNPs from the 1000 Genomes Project, and displaying Pfam domains and links to biological pathways for selected genes.

For mining a large dataset, COSMICmart (an instance of BioMart) is a tool for downloading user-customized datasets federated with external databases such as Ensembl and Uniprot. Moreover, we provide data export in multiple formats and Oracle database export through the FTP site [ftp://ftp.sanger.ac.uk/pub/CGP/cosmic].

P18 Deploying next-generation sequencing in a hospital setting

Lars G T Jorgensen, Katherine Siminovitch

Mount Sinai Hospital, Toronto, ON, Canada MSG 1X5

BMC Proceedings 2012, 6(Suppl 6):P18

Personalized medicine is the ability to tailor healthcare decisions based on an individual's unique characteristics (genetics, demographic information, healthcare experience, environment and social factors) to more accurately diagnose the individual's disease, predict its outcomes, and select treatments that increase the chances of a successful outcome and reduce possible adverse reactions. Moreover, it is the ability to predict an individual's susceptibility to diseases with the goal of taking measures to prevent or mitigate the extent to which an individual will experience a disease.

The promise of personalized medicine in healthcare delivery includes:

- avoiding the cost of a drug or therapy with little chance of success for each patient
- avoiding exposing patients to side-effects from the wrong treatment
- avoiding increased disease burden from delays in finding the right treatment
- being able to act preventatively in cases where we can predict disease susceptibility, to delay onset and/or progression.

These changes in healthcare delivery promise improved patient outcomes as well as reduced healthcare and societal costs of managing disease.

A key component in personalized medicine is the emergence of whole-genome-scale sequencing as a platform to identify gene variants. Many challenges exist, however, to the use of large-scale sequencing in disease gene discovery and the efficacious application of such knowledge to health benefit. Leveraging upon the Ontario Government’s CAD$7 million committed investment to build a clinical genomics exome sequencing platform for personalized healthcare at Mount Sinai Hospital, our project is developing a portfolio of computational tools and methods to enable acquisition, analysis, interpretation and reporting of high-throughput sequence data and the ethical, economic, psychosocial and knowledge synthesis support needed to move deep sequencing datasets and their findings into clinical practice.

The challenges to moving next-generation sequencing (NGS) into the hospital setting are both technical and regulatory. Since introduced into the research community in 2007, broad use of NGS has been greatly enabled by development of sophisticated informatics and analytic tools. While such tools have reduced many of the barriers to clinical application of NGS, the investment needed to bring NGS into medical practice remains significant, the scale of IT required being unprecedented at most hospitals. In this context, hospitals are, to some extent, at a similar place to that of research institutes when NGS first emerged. Some of the IT barriers to clinical use of NGS have been bypassed by the expertise in NGS data management developed at genome centers and across the research community. However, even this wealth of experience does not fully address the roadblocks inherent to integrating NGS information into existing health workflows and IT systems, and the added challenges posed by the regulatory controls in place across the health system.

This talk will describe strategies our group is using to overcome these hurdles and thereby expedite clinical integration of NGS capability and its translation to delivery of personalized healthcare.
P19
Investigation of rare variation in one thousand cancer genes and its association with risk of breast and prostate cancer
Daniel Klevebring1, Johan Lindberg1, Simon Sundling2, Märten Neiman1, Julia Sandberg1, Fredrik Wiklund1, Per Hall1, Henrik Gronberg2, Kamila Czene1
1Department of Medical Epidemiology and Biostatistics, Science for Life Laboratory, Karolinska Institutet, SE-171 77 Stockholm, Sweden; 2Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, SE-171 77 Stockholm, Sweden
BMC Proceedings 2012, 6(Suppl 6):P19

Genome-wide associated studies have provided a first understanding of the genetics behind breast and prostate cancer risk. However, this methodology has limitations in that only previously detected variants can be interrogated, and thus it also is limited to variants with high frequency in a population. Genetic variants detected by genome-wide associated studies each commonly have a low impact on the risk of developing disease. Additionally, a handful of genes have been detected to confer a moderate risk of developing disease.

With the advent of exome sequencing of the genome of various tumor types, it is becoming clear that very few somatically mutated genes (SMGs) are detected to confer a moderate risk of developing disease. However, most SMGs are found in several tumor types, albeit in different frequencies. Here we outline a project aimed at investigating germline variation in 1,242 genes of interest.

Based on previously collected sample sets where extensive questionnaire data are available for lifestyle factors, family history of cancer et cetera, we selected 480 breast cancer cases and 480 prostate cancer cases with a positive family history and/or young age of onset. All lung tissue samples used in this study, as well as additional LGRC lung tissue samples, were run on Agilent V2 human whole genome arrays and Agilent V3 human miRNA microarrays.

Results: Using a mixture model, expression of 6,359 genes was detected, 9,338 genes were not detected, and 8,397 genes were variably detected across all lung tissue samples. Using a subset of 58 samples from subjects with IPF (n = 19), emphysema (n = 19) or control samples (n = 20), differential gene expression was determined using a t test to compare each disease state with control. The expression levels of 1,770 genes differed between IPF and control, and 220 genes between emphysema and control (P < 0.001). Genes that go up in both emphysema and IPF relative to control were enriched for the p53/hypoxia pathway (KEGG, Biocarta) by GSEA. Gene expression estimates were highly correlated between the mRNA-Seq and array datasets across the same samples; however, we identified additional gene expression changes by RNA-Seq (either not significant by or not assayed by microarray) and validated these by quantitative PCR. Array-based gene expression estimates from additional lung tissue samples not sequenced, together with immunohistochemistry, confirmed the upregulation of the p53/hypoxia pathway in emphysema and IPF. Using reads that aligned across known splice junctions, we identified 5 emphysema-associated and 19 IPF-associated alternative splicing events. These events included the loss of exons and changes to the 3' UTR. Using qPCR and Nanostring, we validated two examples of differential splicing, one of which is shared by both chronic obstructive pulmonary disease (COPD) and IPF and involved with the p53 pathway. Finally, using miRconnX, miRNA microarray data and mRNA-Seq, data were integrated with a prior network of computationally predicted and experimentally validated miRNA-mRNA interactions. miRNAs that co-vary with differentially expressed p53/hypoxia genes in IPF and emphysema were identified and validated using in vitro miRNA perturbation studies.

Conclusions: Our data indicate that the lung transcriptome is altered in emphysema and IPF, and suggest that these changes may include alterations in gene expression that are regulated by miRNA as well as disease-associated alternative splicing events that are shared between COPD and IPF. Specifically we observed a shared upregulation of the p53/hypoxia pathway and decreased expression of the miRNAs that may regulate this pathway. Our data also reveal disease-associated changes in known splice junctions that may affect gene regulation, or protein function. This unprecedented high-resolution view of the lung transcriptome associated with IPF and COPD may ultimately provide biomarkers of risk and response to therapy as well as potential therapeutic targets.

P20
Abstract not submitted for online publication.

P21
Comprehensive genomic profiling of the lung transcriptome in emphysema and idiopathic pulmonary fibrosis using RNA-Seq
Rebecca L Kusko1, John Brothers II*2, Gang Liu3, Lingqi Luo3, Julia Sandberg1, Fredrik Wiklund1, Per Hall1, Henrik Gronberg2, Kamila Czene1, Mark Geraci2, John Quackenbush1, Frank Schurba5, Marc Lenburg1,2,5, David A Schwartz2, Jennifer Beane3, Naftali Kaminski4, Avrum Spira2,3
1Genetics and Genomics Graduate Program, Boston University School of Medicine, Boston, MA, USA; 2Bioinformatics Program, Boston University, Boston, MA, USA; 3Division of Computational Biomedicine, Department of Medicine Boston University School of Medicine, Boston, MA, USA; 4Simmons Center for Interstitial Lung Disease and Department of Medicine, University of Pittsburgh Medical Center, Pittsburgh, PA, USA; 5Department of Pathology and Laboratory Medicine, Boston University School of Medicine, Boston, MA, USA; 6Center for Genes, Environment and Health and Department of Medicine, National Jewish Health, Denver, CO, USA; 7Dana-Farber Cancer Institute and Harvard School of Public Health, Boston, MA, USA; 8Department of Medicine, University of Colorado School of Medicine, Aurora, CO, USA
*Equal contributors
BMC Proceedings 2012, 6(Suppl 6):P21

Background: As part of the Lung Genomics Research Consortium (LGRC), we sought to characterize transcriptomic alterations underlying the molecular pathogenesis of emphysema and idiopathic pulmonary fibrosis (IPF) using mRNA-Seq, and comparing to miRNA and microRNA microarray data obtained from the same samples.

Methods: 67 LGRC lung tissue samples were sequenced on the Illumina GAIIx, generating 75 nt paired-end reads and approximately 30-40 million reads per sample. Using gapped aligner Tophat, an average of 85% of reads aligned to hg19. Gene expression was quantified using Cufflinks and Ensembl59 known gene annotation (n = 24,249 genes). All lung tissue samples used in this study, as well as additional LGRC lung tissue samples, were run on Agilent V2 human whole genome arrays and Agilent V3 human miRNA microarrays.

Results: Using a mixture model, expression of 6,359 genes was detected, 9,338 genes were not detected, and 8,397 genes were variably detected across all lung tissue samples. Using a subset of 58 samples from subjects with IPF (n = 19), emphysema (n = 19) or control samples (n = 20), differential gene expression was determined using a t test to compare each disease state with control. The expression levels of 1,770 genes differed between IPF and control, and 220 genes between emphysema and control (P < 0.001). Genes that go up in both emphysema and IPF relative to control were enriched for the p53/hypoxia pathway (KEGG, Biocarta) by GSEA. Gene expression estimates were highly correlated between the mRNA-Seq and array datasets across the same samples; however, we identified additional gene expression changes by RNA-Seq (either not significant by or not assayed by microarray) and validated these by quantitative PCR. Array-based gene expression estimates from additional lung tissue samples not sequenced, together with immunohistochemistry, confirmed the upregulation of the p53/hypoxia pathway in emphysema and IPF. Using reads that aligned across known splice junctions, we identified 5 emphysema-associated and 19 IPF-associated alternative splicing events. These events included the loss of exons and changes to the 3' UTR. Using qPCR and Nanostring, we validated two examples of differential splicing, one of which is shared by both chronic obstructive pulmonary disease (COPD) and IPF and involved with the p53 pathway. Finally, using miRconnX, miRNA microarray data and mRNA-Seq, data were integrated with a prior network of computationally predicted and experimentally validated miRNA-mRNA interactions. miRNAs that co-vary with differentially expressed p53/hypoxia genes in IPF and emphysema were identified and validated using in vitro miRNA perturbation studies.

Conclusions: Our data indicate that the lung transcriptome is altered in emphysema and IPF, and suggest that these changes may include alterations in gene expression that are regulated by miRNA as well as disease-associated alternative splicing events that are shared between COPD and IPF. Specifically we observed a shared upregulation of the p53/hypoxia pathway and decreased expression of the miRNAs that may regulate this pathway. Our data also reveal disease-associated changes in known splice junctions that may affect gene regulation, or protein function. This unprecedented high-resolution view of the lung transcriptome associated with IPF and COPD may ultimately provide biomarkers of risk and response to therapy as well as potential therapeutic targets.

P22
Investigation of retroviral integration preference via high-throughput sequencing
Matthew C LaFave, Gaurav K Varshney, Shawn M Burgess
Genome Technology Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD 20892, USA
BMC Proceedings 2012, 6(Suppl 6):P22

The integration of retroviral DNA into the host genome is a key step of the retrovirus lifecycle. The positions of such integrations are non-random and are characteristic to the virus in question. Retroviral vectors have been useful in gene therapy, but the integration carries with it the danger of host gene misregulation by viral integration, potentially contributing to cancer. An understanding of how integration sites are determined is critical for any approach seeking to avoid such consequences. To better understand the factors influencing retroviral integration site selection, we are generating integrations of murine leukemia virus (MLV) in human K562 cells. We use massively
parallel Illumina technology to sequence the cellular genome and have developed a bioinformatic pipeline to efficiently identify sites of integration. Our aim is to create a high-resolution, high-density integration map with 1 million unique integrations mapped to single-nucleotide resolution. This map will identify regions prone or refractory to MLV integration. We plan to compare this map with data from the ENCODE Consortium, looking for significant associations with characteristics of the local genomic environment, to enhance our understanding of MLV integration site selection.

Results: The best finding was obtained at the HMLGCL1 gene (P = 3 × 10⁻²⁰, odds ratio 0.15). The HMLGCL1 gene has been found to be highly expressed in some brain regions. Other genes harbored in regions enriched with EHHs associated with autism include ZFP91, CNTF, NAPL, and TLE4. We also used the webtool Panther to assess if this set of genes is over-represented in any pathways. The results suggest that these genes collectively may be involved in the Wnt signaling pathway and neurodevelopment process. No remarkable evidence for recent positive selection was obtained for most of these loci, except the TLE4 gene (P = 0.008).

Conclusions: Taken together, these genes involved in brain development may harbor variants that exert a recessive effect on the risk of autism. Further work is warranted to replicate these findings in other populations.

P25
Multiple de novo copy number variations in two subjects with developmental problems and multiple congenital anomalies
Penfée Liu1, Klaudia Walter2, Karin Würz2, Violet Gelovani3, Sarah Lindsay3, Claudia M B Carvalho4, Marjorie Withers5, Joanna Wisiowska3, Anikta Patel6, Bernd Rautenstrauss4, Matthew E Hurles6, James R Lupski1,2,3

Background: de novo copy number variation (CNV) can occur constitutionally in gametogenesis or in early development leading to sporadic genomic disorders. Such de novo CNVs appear to also be important in somatic mutagenesis relevant to cancer and population events important to species evolution. Since large pathological CNVs are rarely observed at more than one locus in a single patient, and are often de novo, current efforts in understanding their molecular features and underlying mechanisms have relied on comparing CNVs from different individuals. Therefore, knowledge regarding size, mechanism and spatial distribution of de novo genomic rearrangements in a single genetic background is lacking.

Materials and methods: Two subjects with developmental problems and multiple congenital anomalies were identified by clinical array comparative genomic hybridization (aCGH) to have more than five de novo CNVs larger than 500 kb. Custom-designed aCGH and whole genome SNP arrays were used to fine map large de novo CNVs in both of the subjects and determine parental origins, whereas breakpoint sequencing was performed to identify mechanisms. In order to characterize smaller sized CNVs not efficiently interrogated by aCGH, Illumina whole genome sequencing was performed on both subjects to obtain a sequence coverage of more than 30×.

Results: We report two subjects with a constitutional ‘CNV mutator’ phenotype, in whom multiple de novo rearrangements were observed on different chromosomes. Such observations are distinguishable from the phenomenon of chromothripsis in which the multiple CNV changes concentrate on one chromosome [1]. Subject no. 1 carried 8 large (>100 kb) copy number gains, ranging from 104 kb to 6.4 Mb. Subject no. 2 carried 11 large copy number gains, ranging from 211 kb to 4.7 Mb. Breakpoint sequencing analysis showed that microhomologies and breakpoint complexities are the prevailing features left at rearrangement tracts, suggesting that the rearrangements were likely produced by replication mechanisms such as fork stalling and template switching and/or microhomology-mediated break-induced replication (FoSTeS/MbMR). Haplotyping analysis in subject no. 1 revealed that the duplicated or triplicated materials were derived from both the paternal chromosome and the maternal chromosome, suggesting a postzygotic timing of the mutations.

Conclusions: Our results document a genome-wide spectrum of de novo CNVs in a ‘CNV mutator’ phenotype background, and we suggest that errors in the cellular DNA replication machinery could lead to multiple independent de novo rearrangements. Our findings have important implications for genomic disorders, cancer and evolution.
Next-generation sequencing (NGS) has significantly impacted human genetics, enabling a comprehensive characterization of human genome as well as better understanding of many genomic abnormalities. By delivering massive DNA sequences at unprecedented speed and cost, NGS promises to make personalized medicine a reality in the foreseeable future. To date, library construction with clinical samples has been a challenge, primarily due to the limited quantities of sample DNA available. To overcome this challenge, we have developed a fast library preparation method using novel NEBNext reagents and adaptors, including a new DNA polymerase that has been optimized to minimize GC bias. This method enables library construction from an amount of DNA as low as 5 ng, and can be used for both intact and fragmented DNA. Moreover, the workflow is compatible with multiple NGS platforms.

P27

A functional role of the SNP -592 of human IL10 gene regulatory region is associated with an increased IL-10 expression and risk for human papillomavirus cervical lesion and cervical cancer development

K Torres-Poveda1, A Burguete-García1, M Bahena-Román1, F Torres-García1, J Delgado-Romero1, D Cantu1, A García-Carrancá1, V Madrid-Marina1

Background: An immunosuppressive state had been identified in women with persistent infection by the human papillomavirus (HPV), characterized by high levels of interleukin (IL)10 at cervix level [1]. The present study assessed the association of SNP -592 in the IL10 gene, with the risk of developing squamous intraepithelial lesions of the cervix (SIL) and cervical cancer (CC), and evaluated the level of mRNA expression of IL10 (mRNA-IL10) systemically and in the cervix and the IL10 protein level in serum.

Materials and methods: Using a cross-sectional design, samples of peripheral blood of patients with SIL (n = 204), patients with CC (n = 80) and patients without SIL (n = 166) were used to evaluate SNPs at loci -592A/C (rs1800872), -1082C/T (rs1800871), -1082A/G (rs1800896) and -1352A/G (rs1800893) by allele discrimination with Taqman probes and evaluating the mRNA-IL10. Cervical swabs in women without SIL and cervical biopsies in women with SIL and CC were used for HPV typing and evaluating mRNA-IL10. Gene expression levels were evaluated by real-time PCR. The genotype and allele frequencies of polymorphisms were analyzed using logistic regression, adjusting for age and genotype of HPV, to determine the association with SIL and CC.

Results: No significant differences were found between the frequencies of genotypes at loci -819, -1082 and -1352. Individuals who carried at least one copy of the risk allele A of SNP -592, showed an odds ratio (OR) of 2.02 (95% CI 1.26 to 3.25, P =< 0.003) for SIL, an OR of 1.70 (95% CI 1.06 to 2.71, P =< 0.02) for CC, and higher levels of IL10 systemically and in the cervix. Thus, a copy of the risk allele A is sufficient to increase the risk of SIL and CC. We found a significant difference in mRNA-IL10 in both systemic and the cervical levels in women with SIL and CC compared with women without SIL, and in the serum IL10 protein (P < 0.0001), being higher in patients carrying the risk allele A of SNP -592. The levels of both mRNA-IL10 and IL10 protein were progressively higher with increasing degree of malignancy of the lesion, so that the presence of IL10 can be regarded as a relevant factor for viral persistence and progression of disease.

Conclusions: The SNP -592 C/A in the IL10 promoter is associated with increased risk of SIL and CC and can serve as a biomarker predictive of risk for the development of SIL in the cervix and CC in Mexican women, and is associated with the regulation of this cytokine expression in both cervix and systemic level. HPV 16 E6/E7 proteins bind to Sp1 transcription factor and upregulate IL10 gene expression through GGCGCG consensus sequence located at -569 to -603 of the human IL10 gene, and the C/A exchange in the SNP -592 results in increased IL10 gene promoter activity in HPV-transformed cells.

Reference
siRNAs targeting the ERK2 signaling pathway and AML1/MTG8 fusion gene attenuate the differentiation, proliferation and growth arrest in (t(8;21) leukemia

Salem Bashanfer1, Faisal Mutee Al-Hassan1, Rosline Hassan2, Olaf Heidenreich1, Mohamed Safuuljan Mohamed Said4, Narazah Mohd Yusof3
1Advanced Medical and Dental Institute (AMDI), Universiti Sains Malaysia, Bandar Putra Bertam, 13200, Kepala Batas, Penang, Malaysia; 2School of Medical Sciences, Universiti Sains Malaysia, 16150, Kota Bharu, Kelantan, Malaysia; 3University of Newcastle, Northern Institute for Cancer Research, Framlington Place, Newcastle upon Tyne NE2 4HH, UK; 4Faculty of Applied Sciences, Universiti Teknologi MARA (UiTM) 40450, Shah Alam, Selangor, Malaysia

Background: The t(8;21) translocation is one of the most frequent chromosome abnormalities associated with acute myeloid leukemia (AML). This translocation generates the (AML1/MTG8) fusion protein causing blockage of the differentiation process. Moreover, constitutive activation of the mitogen-activated protein kinase (MAPK) pathway as a consequence of this translocation results in an increase in the proliferation rate of leukemic cells.

Materials and methods: To ensure subsequent experiments are clinically relevant, we carried out microarray on clinical samples from patients who manifested the t(8;21) translocation together with the corresponding cell lines (Kasumi-1 and SKNO-1). We carried out microarray, qRT-PCR, gene knockdown by siRNA, flowcytometry and various cell assays for apoptosis and cellular proliferation testing.

Results: Analyses revealed a number of overlapping differentially expressed genes in the AML clinical samples and their corresponding cell lines. These genes were uploaded to the KEGG database through the DAVID software (v6.7) to carry out pathway analyses of these leukemic cells. Furthermore, gene expression profiles of both clinical samples and their relevant cell lines showed increased expression of AML1/MTG8 and ERK2 (MAPK1). Using 100 nM and 200 nM of ERK2-siRNA and siRNA-AML1/MTG8, respectively, siRNAs were transfected by electroporation to knockdown ERK2 and the fusion gene AML1/MTG8 individually and in combination in the t(8;21) cell lines. Gene knockdowns were validated by qRT-PCR, and demonstrated successful mRNA suppression by approximately 80% to 90%. In addition, a slight increase of ERK2 mRNA expression was observed upon AML1/MTG8 suppression.

FACS experiments showed reduced expression of CD34 cell surface marker when AML1/MTG8 was knocked down, which indicated onset of the differentiation process whereas the ERK2 suppression displayed higher CD34 expression when compared with mock and siRNA controls, indicating an antiedifferentiation effect. Cell cycle analyses demonstrated increased growth arrest where G0/G1 phase increased by 10% to 15% individually and 25% to 30% in a combination, which correlated with the MTS assay results. An apoptosis assay revealed anti-apoptotic effect of ERK2 and AML1/MTG8 suppression, which was increased when silenced together in a combination silencing.

Conclusions: These results suggest possible roles of ERK2 activation in differentiation induction, while ERK2 inhibition arrested cell growth and reduced proliferating cells in (t(8;21) leukemia when combined with AML1/MTG8 knockdown.

Comparative analysis of strand-specific RNA sequencing approaches

Daniela Munro, Ping Liu, Christine Sumner, Erbay Yigit, Landon Merrill, Lynne Apone, Brad Langhorst, Fiona Stewart, Eileen T Dimalanta, Theodore Davis
New England Biolabs, Inc., 240 County Road. Ipswich, MA 01938, USA

Background: Standard RNA sequencing approaches generally require double-stranded cDNA synthesis, which erases RNA strand information. Synthesis of a randomly primed double-stranded cDNA followed by addition of adaptors for next-generation sequencing leads to the loss of information about which strand was present in the original mRNA template. The polarity of the transcript is important for correct annotation of novel genes, identification of antisense transcripts with potential regulatory roles, and for correct determination of gene expression levels in the presence of antisense transcripts. Different strand-specific RNA-seq approaches have been developed to preserve information about strand polarity with different levels of performances.

Materials and methods: Using our custom Deep Sequencing Technologies, this work investigates the performance of two different directional RNA-Seq (strand-specific RNA-seq) strategies. One is based on direct ligation of adaptors to the RNA ends and the other is based on the labeling and excision of the second strand cDNA. The RNA-seq workflows present in this work have been improved over current more laborious RNA-seq methods. Their low RNA input and streamlined workflows make them compatible with high throughput and automation. We also analyze the effect of different RNA fragmentation methods (divalent cations plus heat versus enzymatic fragmentation).

Results: We will provide a comparative full data analysis of different strand-specific RNA methods (library performance, complexity, continuity of gene coverage, strand specificity, RNA background).

Conclusions: Our results show improved methods for high-quality strand-specific RNA-seq library construction amenable to large-scale library construction and automation.
endothelium constitutes a critical step in the initiation of atherosclerosis.

**Table 1. mtDNA mutations found in infertile patients**

<table>
<thead>
<tr>
<th>Type of Infertility</th>
<th>Gene</th>
<th>At nucleotide position</th>
<th>At amino acid position</th>
<th>Nature of mutation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Asthenozoospermia</td>
<td>ND2</td>
<td>T4823C</td>
<td>V118V</td>
<td>Silent</td>
</tr>
<tr>
<td>Asthenozoospermia</td>
<td>ND2</td>
<td>T4993C</td>
<td>L175P</td>
<td>Novel</td>
</tr>
<tr>
<td>Asthenoteratozoospermia</td>
<td>ND2</td>
<td>C4730T</td>
<td>T87T</td>
<td>Silent</td>
</tr>
<tr>
<td>Asthenoteratozoospermia</td>
<td>ND2</td>
<td>T5250G</td>
<td>L261V</td>
<td>Novel</td>
</tr>
<tr>
<td>Asthenoteratozoospermia</td>
<td>ATPase6</td>
<td>G8557C</td>
<td>L64F</td>
<td>Novel</td>
</tr>
<tr>
<td>Oligoastheneteratozoospermia</td>
<td>ATPase6</td>
<td>T8614G</td>
<td>L30V</td>
<td>Novel</td>
</tr>
<tr>
<td>Oligoastheneteratozoospermia</td>
<td>ATPase6</td>
<td>A8925G</td>
<td>T133T</td>
<td>Silent</td>
</tr>
<tr>
<td>Oligoastheneteratozoospermia</td>
<td>ATPase6</td>
<td>G9064A</td>
<td>A180T</td>
<td>Novel</td>
</tr>
<tr>
<td>Oligoastheneteratozoospermia</td>
<td>ND4</td>
<td>A10978G</td>
<td>L73L</td>
<td>Silent</td>
</tr>
</tbody>
</table>

Conclusions: The study reveals that the frequency of Yq microdeletion is higher in semen samples than blood samples, possibly because most of the genes studied are testis-specific in nature. This is in accordance with the previous studies [3, 4]. In both types of samples A2Zr regions have highest frequency of Yq microdeletion. The mtDNA mutations common in fertile and infertile patients have also been reported by other Indian researchers [5], but doubted to play any role in infertility [6]. The mutation G9064A in ATPase6 has also been reported to play a role in female infertility [7]. Beside these, the other mtDNA mutations observed in the present study have not been reported previously.

Acknowledgements: This work was supported by DST Grant under the research project to Dr. Rita Mahanta and also DST grant under Women Scientist-A (WOS-A) scheme to PNBC. Authors are also thankful to the doctors and technicians of the infertility clinics from where the samples have been collected for their support.

References


**P32**

The human microbiome project [11] is the first systematic and large-scale survey of microbial communities present in a eukaryotic metagenome. The availability of low-cost high-throughput next-generation sequencing technologies that enable us to sequence eukaryotes to a high depth of coverage presents an opportunity for in silico discovery of endosymbionts. We describe a method for mining a whole genome shotgun metagenome from an insect to identify members of the endosymbiont community, followed by reconstruction and validation of a high-quality draft microbial genome. The Asian citrus psyllid (Diaspphinae citri Kuwayama or ACP) is host to 7+ bacterial endosymbionts [2] and is the insect vector of Candidatus Liberibacter asiaticus (Las), which is the causal agent of citrus greening, a disease costing the Florida citrus industry US$3.63 billion and 6,611 jobs since 2006. Citrus greening is a complex patho-system that involves interactions between the psyllid vector, the citrus host, and Candidatus Liberibacter asiaticus.

Materials and methods: To gain a better understanding of the ACP endosymbiont community, DNA from ACP was sequenced to 100X coverage to produce paired-end and mate-pair libraries. Initial analyses focused on Wolbachia, an alpha-proteobacterial primary endosymbiont typically found in the reproductive tissues of ACP and other arthropods. The metagenomic sequences were mined for Wolbachia (wACP) reads using four sequenced Wolbachia genomes as bait. Putative wACP reads were then assembled using Velvet and MiRA3 assemblers. The resulting wACP contigs were annotated using the RAST pipeline and compared with the closest sequenced Wolbachia from an insect genome: Wolbachia endosymbiont of Culex quinquefasciatus (wPip). MiRA3 was able to reconstruct a majority of the wPip coding sequence regions and was therefore selected for scaffolding using large insert mate-pair libraries. The wACP scaffolds were further improved using Abacas and Mauve contig mover with wPip as reference genome to orient and order the contigs.

Results and conclusions: We validated the final wACP scaffold by comparing all wACP proteins against the four sequenced Wolbachia genomes. We ran OrthoMCL and selected core and shared Wolbachia proteins where the classification was highly conserved (>80%) across all runs. In order to determine the presence of the core Wolbachia
proteins in our wACP scaffold, we compared wACP proteins with labeled Wolbachia proteins. 1,164/1,213 wACP proteins had matches, of which 669 were to core proteins. This number compares favorably to the number of core proteins (670) found in sequenced Wolbachia. The scope of endosymbiont characterization was expanded beyond wACP using 16S rDNA and partial 23S rDNA analysis [2] as a guide. Reads from the ACP metagenome were mapped to reference genomes of candidate endosymbionts and mapped regions were analyzed manually. We are currently screening out rRNA regions and using taxonomic classification in order to determine the specificity of these regions to the genome. Results will be presented regarding targeted assembly methods, validation metrics and effectiveness of taxonomic classification strategies for metagenomics.

References

P34
Dissecting the genetic architecture of coronary artery disease by genome engineering
David J Segal1, Mital S Bhakta1, Kumita Theva Das1, Chongxiu Sun1, Natalie M Grace1, Jan A Nolta1, Anne A Knowlton1, David M Rocke1, Scott I Simon1
1Genome Center and Department of Molecular Medicine, University of California - Davis, CA, 95616, USA; 2Biomedical Engineering, University of California - Davis, CA, 95616, USA; 3Institute for Regenerative Cures, University of California - Davis, CA, 95616, USA; 4Cardiovascular Medicine, University of California - Davis, CA, 95616, USA; 5Division of Biostatistics, University of California - Davis, CA, 95616, USA

BMC Proceedings 2012, 6(Suppl 6):P34

One of the greatest challenges facing biomedical research since the sequencing of the human genome has been to understand the role of genetic variation in human disease. Many genetic variants have been associated with common diseases. However, determining the functional consequences of these variants has been hard. Several variants are often inherited together in tightly linked blocks, making it difficult to determine the causative variant. People have millions of other genetic differences, making it difficult to correlate cellular phenotypes with a particular variant. Different gene sets are expressed in different cells, but it is difficult to extract disease-relevant cells from large numbers of patients. We describe a method with the potential to revolutionize the functional analysis of genetic variation, using custom nucleic acids to genetically modify individual variants in induced pluripotent stem cells. This process would provide unprecedented analytical power, present the first general method to determine if a variant is causative, and analyze function disease-relevant cell types. We will focus on variants at the 9p21 region of the genome that have been associated with coronary artery disease (CAD). The methods should provide a new way to unlock the wealth of data from genome-wide association studies, and to probe the genetic architecture of common diseases. We will describe our improved methods for inexpensive and rapid construction of highly active zinc fi nger and TALE nucleases to examine variations within the genomes of both model organisms and human populations.

Acknowledgement: This work is generously supported by the W M Keck Foundation.

P35
Exome sequencing identifies somatic point mutations associated with acquired endocrine resistance in breast cancer cell lines
Natasia S Ehlers1, Zhu Shi Da1, Daniel Elais1, Xin Lin1, Jian Lin2, Christina Bjørre1, Nils Brunner1, Lars Bolund1,3,4, Wang Jun1,3, Ramneek Gupta1, Henrik J Ditzel1,4
1Sino-Danish Breast Cancer Research Centre; 2Center for Biological Sequence Analysis, Department of Systems Biology, Technical University of Denmark, Building 208, DK-2800 Kongens Lyngby, Denmark; 3BGI Shenzhen, Beishan Industrial Zone, Yantian District, Shenzhen 518083, China; 4Department of Cancer and Inflammation Research, Institute of Molecular Medicine, University of Southern Denmark, Wilslevvej 25, DK-5000 Odense, Denmark; 5Department of Human Genetics, Aarhus University, Wilhelm Meyers Allé 4, DK-8000 Aarhus C, Denmark; 6Section for Pathobiology, Department of Veterinary Science, College of Life Sciences, University of Copenhagen, Dyrlægevej 68, DK-1870 Frederiksberg C, Denmark

BMC Proceedings 2012, 6(Suppl 6):P35

Background: Endocrine therapy is an effective treatment of estrogen receptor-positive (ER+) breast tumors, significantly reducing mortality. However, approximately 30% of patients receiving adjuvant endocrine therapy will experience recurrence within a 15-year period. The mechanisms of endocrine resistance are poorly understood. Understanding the underlying genetic diversity of breast cancers responding differently to endocrine therapy is important for the development of more optimal and individualized treatments strategies.

Materials and methods: In the current study, a panel of isogenic MCF-7-derived human breast cancer cell lines [1-3] that are resistant to tamoxifen only, or to both tamoxifen and fulvestrant, respectively, were analyzed for mutations through exome sequencing and compared with the exome of the parental cell line. In addition, global gene expression levels for the same panel of cell lines were generated. Detected variations were integrated with gene expression profiles and analyzed in the context of prior knowledge of drug action and genes associated with resistance to endocrine therapies as identified by extensive literature curation.

Results and conclusion: A small panel of somatic point mutations potentially associated with acquired endocrine resistance were identified. Future experimental validation will reveal which of the detected mutations that are causatively involved in resistance to endocrine therapy.

References

P36
Next-generation sequencing at Merck-Boston
Richard Stevens
Affiliation: tbc
BMC Proceedings 2012, 6(Suppl 6):P36

Next-generation sequencing (NGS) is quickly replacing other methods for determining expression profiling of RNA as well as single-nucleotide variations within the genomes of both model organisms and human samples. At the Boston facility of Merck Research Laboratories, we concentrate on preparation and sequencing of samples where the research needs cannot be met by commercial vendors. These unmet needs may be due to either availability of up-to-date protocols or to deadline constraints. In addition to validating and developing new library-construction and sequencing protocols, we also evaluate commercial vendors in consideration for future outsourcing. One of these current projects is an evaluation of transcriptome sequencing and profiling of formalin fixed paraffin embedded (FFPE) samples from Merck’s current and past clinical studies. FFPE samples are readily available and easily stored but create difficulties in NGS analysis due to the low quality of the purified nucleic acids. Here, we compare techniques for transcriptome sequencing to microarray profiling of RNA purified from FFPE tissues as well as their mimicked fresh frozen counterparts.
P37
AVIA: an interactive web-server for annotation, visualization and impact analysis of genomic variations
Hue Vuong, Robert M Stephens, Natalia Volfovsy
Advanced Biomedical Computing Center (ABCC), Information Systems Program, SACL-Frederick Inc., Frederick National Laboratory for Cancer Research, Frederick, MD 21702, USA
BMC Proceedings 2012, 6(Suppl 6):P37

Background: A plethora of information that emerges from large-scale genome characterization studies has triggered the development of computational frameworks and tools for efficient analysis, interpretation and visualization of genomic data. Functional annotation of genomic variations and the ability to visualize the data in the context of whole genome and/or multiple genomes has remained a challenging task. We have developed an interactive web-based tool, AVIA (http://avia.abcc.ncbi.gov), to explore and interpret large sets of genomic variations (single nucleotide variations and insertion/deletions) to help guide and summarize genomic experiments.

Material and methods: Our tool is based on coupling a comprehensive annotation pipeline with a flexible visualization method. We leveraged the ANNOVAR [1] framework for assigning functional impact to genomic variations by expanding its list of reference annotation databases (RefSeq, UCSC, SIFT, PolyPhen, etc) with additional in-house developed sources (Non-B DB, PolyBrowse). Further, because many users also have their own annotation sources, we have added the ability to supply their own files as well. The results can be obtained in tabular format or as tracks in whole genome circular views generated by the Circos application [2]. Users can select different sets of pre-computed tracks, including whole genome distributions of different genomic features (genes, exons, repeats), as well as variations analysis tracks for the 69 CGI public genomes for reference.

Results and Conclusions: This version of AVIA is focused on gene-related impact assessment. Tracks showing the distribution of genes with variations of specific functional effects, such as non-synonymous variations, frame shifts, variable miRNA target sites or variations in G-quadruplexes in 5’ UTRs, can be produced. Users can screen all RefSeq genes or other publicly available datasets or provide their own list of genes. Another option of the interactive analysis with AVIA is the ability to compare two input genomes whereby we can generate views of shared or distinct variations. This allows for comparison of tumor/normal data sets to illustrate tumor-specific mutations. By exploring their work with AVIA, users can browse different tracks with their data and then re-generate signature plots to summarize the project.

Acknowledgements: The authors wish to thank members of Bioinformatics Support and Scientific Web Programming Groups at ABCC and ABCC’s in Silico Research Center of Excellence. This work was supported with federal funds from the National Cancer Institute, National Institutes of Health, under contract HHSN26120080001E.

References

P38
Deep sequencing of the microRNA transcriptome in current, former, and never smokers with lung adenocarcinoma
Teresa Wu Wang1,2, Joshua D Campbell1,2, Lingqi Luo1,2, Gang Liu1, Ji Xiaoz1,2, Marc E Lemberg1,2,3, Steven A Belinsky1,2,3,5, J Soikkeli, M Yin, E Hölttä5
1Bioinformatics Program, Boston University, Boston, MA 02215, USA; 2Section of Computational Biomedicine, Boston University School of Medicine, Boston, MA 02118, USA; 3Department of Pathology and Laboratory Medicine, Boston University School of Medicine, Boston, MA 02118, USA; 4Lung Cancer Program, Lovelace Respiratory Research Institute, Albuquerque, NM 87108, USA; 5Section of Pulmonary, Critical Care and Allergy, Boston University School of Medicine, Boston, MA 02118, USA
BMC Proceedings 2012, 6(Suppl 6):P38

Background: Lung cancer is the leading cause of cancer-related death in the USA. While smoking is a major risk factor, about 10% to 15% of cases arise in never smokers. Small non-coding RNAs such as microRNAs (miRNAs) often act as oncogenes or tumor suppressors to regulate gene expression during disease, and represent attractive targets for lung cancer risk assessment and therapeutic intervention. We therefore sought to characterize the tumor-associated miRNA transcriptome of current, former and never smokers.

Materials and methods: Total RNA was isolated from the tumor and adjacent-normal pairs (tumor purity >70% normal tissue is collected most distant from the tumor at time of resection) of current (n = 8), former (n = 11) and never (n = 14) smokers with adenocarcinoma. Small RNA libraries were generated and then multiplexed 7 to 8 per lane for sequencing on the Illumina HiSeq 2000. Through a custom miRNA sequencing analysis pipeline, reads were trimmed, size-selected, and mapped to an assembly of the human genome reference (hg19) using Bowtie. Counts per mature miRNA from aligned reads were computed using Bedtools and a list of genomic features retrieved from miRBase (version 17). Finally, differential expression analysis was conducted using Student’s t test and ANOVA.

Results: Small RNA sequencing generated an average of 10 million high-quality miRNA reads per sample. Among the 1,906 mature miRNAs examined, 354 miRNAs had at least an average of 20 counts across all 44 samples. 94 miRNAs (false discovery rate q<0.25) are differentially expressed in tumors of ever smokers (n = 19) compared with never smokers (n = 14). This includes miRNA-320b, which appears to be highly activated in current and former smokers, but unchanged in never smokers. Several miRNAs, including miRNA-21 and miRNA-182, were significantly upregulated in tumors compared with the adjacent-normal, irrespective of smoking status.

Conclusions: Using small RNA sequencing, we are comprehensively profiling the miRNA transcriptome in current, former and never smoker lung adenocarcinoma. We have demonstrated that subsets of miRNAs are markedly dysregulated in tumors when compared to their adjacent-normal counterparts. Some of these profiles may be unique to lung cancers cases that arise in current and former smokers. These results will supplement our analysis of large RNA libraries prepared from the same set of tissues to extend our knowledge of the lung cancer transcriptome, to identify key microRNA-mRNA interactions underlying lung carcinogenesis, and to better understand general mechanisms of lung carcinogenesis as well as those that are specific to carcinogenesis in the presence or absence of tobacco smoke exposure.

P39
Metastatic outgrowth encompasses COL-I, FN1 and POSTN upregulation and assembly to fibrillar networks regulating cell adhesion, migration and growth
J Soikkeli, M Yin, E Hölttä
Department of Pathology, Haartman Institute, University of Helsinki, Helsinki, Finland F-0014
BMC Proceedings 2012, 6(Suppl 6):P39

Background: Metastasis is a complex event in tumor development and accounts for most mortality in cancer patients. During this multistep process, cancer cells leave the primary site, intravasate into the blood or lymphatic vessels to disseminate, extravasate and finally outgrow at a distant site. Unfortunately, the outgrowth of metastatic cancer cells still remains poorly understood. It would therefore be important to unveil the molecular mechanisms involved in the development of micrometastases to overt metastases.

Materials and methods: In the present study, we performed comparative DNA microarray (Affymetrix) analysis of melanoma lymph node micrometastases and macrometastases. Microarray data were analyzed by different methods and tools, including Significance Analysis of Microarrays (SAM) 3.0, fold change-ranking combined with t statistics (Volcano plot), GeneSpring GX 7.3, Gene Set Enrichment Analysis (Molecular Signature Database, v2.5), GenMAPP 2.1 and Ingenuity Pathway Analysis (IPA), to identify the genes and signaling pathways related to the outgrowth of metastases. The identified signaling pathways and genes were further validated by immunohistochemical and functional analyses.

Results: We found that the transforming growth factor-beta (TGF-β) signaling pathway and TGF-β-induced extracellular matrix (ECM)
molecules, peristin (POSTN), fibronectin 1 (FN1), collagen type I (COL-I) and versican (VCAN), were associated with the outgrowth of melanoma lymph node micrometastases to macrometastases. Immunohistochemical analyses confirmed the activation of TGF-β signaling pathway, as assessed by phospho-SMAD2 staining. The ECM proteins POSTN, FN1, COL-I and VCAN, together, were found to form fibrillar networks that regulated the adhesion, migration and growth of melanoma cells, fibroblasts and endothelial cells.

Conclusions: The results suggest an import role for the TGF-β signaling pathway, as well as POSTN, FN1, COL-I and VCAN, in the regulation of metastatic outgrowth. Thus, TGF-β receptors and the metastasis-related matrix proteins FN1 and POSTN, in particular, may offer good therapeutic targets for the prevention of melanoma metastasis.

P40 Massively parallel sequencing of formalin-fixed paraffin-embedded tissues
Rachel L ErlICH, Emanuele Palescandolo, Robert T Jones, Ashwini Sunkavalli, Alina Raza, Matthew Ducar, Charles Hatton, Christina Go, Christine Roden, Megan Hanna, Adin Mills, Ravalì Adusumilli, Prateek Kumar, Priscilla K Brastianos, Matthew L Meyerson, William C. Hahn, Laura E. MacConaill, Paul Van Hummelen, Dana-Farber Cancer Institute, Boston, Massachusetts 02115, USA

Formalin fixation followed by paraffin embedding (FFPE) is the most common method of preserving resected tissues. FFPEs can be easily stored, retrieved and processed for further analysis when compared to the logistical complications of processing fresh frozen (FF) material. Conversely, they may not be ideal for sequencing because of the nucleic acid fragmentation and artifacts introduced by fixation. Here we report our efforts in performing next-generation sequencing on more than 70 specimens comprising ~40 FFPE samples, from low to high quality, and compared the performance of FFPE, FF and commercially available cell lines.

P41 Metabolomic and exome sequence analysis reveal novel molecular signatures associated with colorectal cancer relapse
Rebinder Gauba, Thamemozhi G Natarajan, Lei Song, Krithika Bhuvaneswar, Subha Madhavan, Yuriy Gusev
1Innovation Center for Biomedical Informatics, Georgetown University Medical Center, Washington DC, 20007, USA

Background: Colorectal cancer (CRC) is the third leading cancer killer in the United States, affecting both men and women. Standard treatment for stage II colon cancer is surgical removal of the cancer and an area surrounding the cancer. Adjuvant chemotherapy may also be given as a precaution against cancer recurrence, usually recommended to high-risk patients. Recent studies have shown that for stage II patients, adjuvant therapy does not increase overall survival, but there is a significant increase in progression-free survival after therapy [1]. Therefore, it is important to identify a subgroup of patients with the highest risk of relapse who can potentially benefit from adjuvant chemotherapy. The informatics team at Georgetown University has generated and analyzed multi-omics profiling datasets in stage II CRC patients with or without relapse to identify molecular signatures in CRC that may serve both as prognostic markers of recurrence, and also allow for identification of the subgroup of patients who might benefit from adjuvant chemotherapy.

Methods and results: Mass spectrum data from serum and urine samples of 40 CRC patients (20 relapse and 20 relapse-free samples) were collected and pre-processed using a standard pre-processing pipeline. Statistical group comparison was performed using t-test (Bio-conductor). A total of 77 differential metabolites in serum and 47 in urine were identified. Nearly 60% of differential metabolites in both urine and serum were identified using the in-house developed Metabolomics - Data Analysis and Annotation Pipeline (M-DAAP). Enrichment analysis revealed top enriched pathways, including 'Nicotinate and Nicotinamide Metabolism' (P = 0.007) in urine and 'Methionine Metabolism' (P = 0.002) in serum. A comprehensive catalogue of genomic alterations in the coding regions of 40 tumor samples was also generated by whole-exome sequencing to decipher the mutational landscape associated with CRC relapse. Enrichment analysis of the statistically significant genes from a gene-based association test revealed several pro-inflammatory and chemokine remodeling pathways, including 'BRADYKININR → STAT3 signaling' (P = 0.005); 'NG080 Chromatin Remodeling' (P = 0.013); and 'TNFR → AP-1/ATF53 signaling' (P = 0.03).

Conclusions: Preliminary exome data analysis revealed pro-inflammatory signaling as the most significant pathway. Bradykinin receptor (BK) is a potent pro-inflammatory mediator and, together with TNF-α (another pro-inflammatory cytokine implicated in CRC), has been observed in colon myofibroblasts, leading to enhanced signaling of several downstream kinases involved in oxidative stress, tumor proliferation and metastasis [2]. Urinary metabolites revealed Nicotinate and Nicotinamide Metabolism as the top enriched pathway. Genes in this pathway have been shown to be differentially expressed in colon cancer cells and are also closely related to oxidative stress and inflammatory response [3] and thus support its potential biological role in colon cancer etiology.

Acknowledgements: The authors thank the Lombardi Shared Resources for generating the metabolomics datasets. They also thank Dr Louis Weiner and Dr Steve Byers for their support. This effort was partly funded by NCI’s in silico research centers for excellence contract.

References

P42 Optimizing genotype quality metrics for individual exomes and cohort analysis
Paul MK Gordon,1 Leo Dimnik,2 Ryan Lamont1,3, Michele Innes1, Francois Bernier1, Jillian Parboosingh1,2
1Alberta Children’s Hospital Research Institute (ACHRI) Genomics Platform, University of Calgary, Calgary, Alberta, Canada, T2N 4N1; 2Genetic Laboratory Services, Alberta Health Services, Calgary, Alberta, Canada, T3B 6A8; 3Department of Medical Genetics, University of Calgary, Calgary, Alberta, Canada, T2N 4N1

Background: Few evidence-based best practice bioinformatics guidelines exist for genotyping using next-generation sequencing data, especially colorspace data produced by Life Technologies sequencers. Dozens of software packages can perform the various steps required, and genome features such as pseudogenes or large paralogous gene families are problematic. High false positive and negative rates can compound the difficulty of cohort analysis.

Materials and methods: Using a Sanger-validated set of 32 BRCA gene regions from 16 patients, high-throughput colorspace (Life Technologies) sequencing performance was optimized by comparing various combinations of sequence aligners, re-aligners, de-duplicators, quality re-calibrators and genotype callers. Independently, six exomes were captured using the Agilent SureSelect v3 kit. The optimized pipeline was applied, and results were compared to microarray genotyping to characterize false positives and negatives. A further four exomes were pair-end sequenced on both the Life Technologies 5500xl and Illumina HiSeq sequencers to check platform concordance. Variant metrics for each exome were compared to the literature. In the clinic, individual exomes are manually triaged by a medical geneticist, and salient variants are confirmed by Sanger sequencing.

For disease cohorts, software was developed to isolate variants possibly causing monogenic rare diseases, taking likely false positives into account.
Results: Using results from Life Technologies' reference genome aligner, the intersection of single nucleotide polymorphism (SNP) calls from FreeBayes [1] (with Samtools [2] de-duplication) and Life Technologies' diBayes (with Picard de-duplication) was optimal. Using reads realigned by the Broad Institute Genome Analysis Toolkit (GATK) [3], the intersection of insertion and deletion calls from FreeBayes and Atlas2 [4] was optimal. A threshold of 14% variant reads for true heterozygous calls was observed.

For bases with 10x coverage, variant calls are on average 98.9% concordant with SNP microarrays (versus 99.2% microarray technical reproducibility [5]). False positive and negative variant rates are each approximately 0.5%, with all false positives called heterozygous. Concordance with Illumina variant calls from a standard GATK pipeline was 95.2%. GATK produced more novel variants, especially in non-unique genomic regions: such variants are flagged with caveats in the colorspace pipeline. In a dominant heterozygous model analysis of five Nager syndrome patients, our cohort analysis software excluded 19 candidate genes, based mainly on a preponderance of genotype caveats.

Many published metrics for SNP quality control are based on a small number of cases, including those in Table 1, which includes a high proportion of non-coding, non-protein-coding genes.

Table 1. Quality metrics reported in the literature, and the optimized colorspace genotyping results.

<table>
<thead>
<tr>
<th>Quality Type</th>
<th>Ideal</th>
<th>Colorspace genotype</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Protein coding</td>
<td>0.048% [6]</td>
<td>0.052</td>
<td></td>
</tr>
<tr>
<td>Non-coding</td>
<td>&gt;coding</td>
<td>0.056</td>
<td></td>
</tr>
<tr>
<td>Non-synonymous</td>
<td>45% [7]</td>
<td>46.2</td>
<td></td>
</tr>
<tr>
<td>Homozygous</td>
<td>37.40% [8]</td>
<td>38.7</td>
<td></td>
</tr>
<tr>
<td>Coding SNP transitions</td>
<td>2.8-3.0 :1 [9]</td>
<td>3.2:1</td>
<td></td>
</tr>
<tr>
<td>Non-coding SNP transitions</td>
<td>2.0-2.21 :1 [9]</td>
<td>2.3:1</td>
<td></td>
</tr>
<tr>
<td>CDS novel (versus dbSNP135)</td>
<td>N/A</td>
<td>0.58</td>
<td></td>
</tr>
</tbody>
</table>

Conclusions: Low false positive and negative rates using colorspace data can be achieved by: first, reporting only concurrent variants from multiple methods; and second, reporting caveats where the reference sequence is not unique. Accurate calls and caveats enable major cohort gene triage when modeling diseases caused by monogenic rare variants.

Acknowledgements: We thank Dr Richard Pon’s laboratory for producing the high-quality colorspace data. We also thank the FORGE Consortium for the HiSeq-derived genotypes.

References

P43
ICGC PedBrain – dissecting the genomic complexity underlying medulloblastoma using whole-genome sequencing
Natalie Jager1, David TW Jones1, Marcel Kool2, Thomas Zichner1, Barbara Hutter1, Marc Sultan1, Yoon-Jae Cho3, Trevor Pugh4, Volker Hovestadt4, Adrian M Stutz5, Tobias Rausch6, Hans-Jörg Wannitz6, Benedikt Brors1, Paul A Northcott4, Michael D Taylor5, Matthew Meyerson6, Scott L, Pomery1, Marie-Laure Yaross7, Jan O Korbel1, Andy Krushunov1,11,13, Roland Eis1,15, Stefan M Pfister1,14, Peter Lichter1
1Division of Theoretical Bioinformatics, German Cancer Research Center (DKFZ), Im Neuenheimer Feld 280, Heidelberg, 69120, Germany; 2Division of Pediatric Neurooncology, German Cancer Research Center (DKFZ), Im Neuenheimer Feld 280, Heidelberg, 69120, Germany; 3European Molecular Biology Laboratory (EMBL), Meyerhofstrasse 1, Heidelberg, 69117, Germany; 4Max Planck Institute for Molecular Genetics, Innesstrasse 63-73, Berlin, 14195, Germany; 5Division of Child Neurology, Stanford University, 750 Welch Road, Palo Alto, CA, 94304, USA; 6Broad Institute of MIT and Harvard, Cambridge, MA, 02142, USA; 7Division of Genetics, German Cancer Research Center (DKFZ), Im Neuenheimer Feld 280, Heidelberg, 69120, Germany; 8Division of Neurosurgery and The Arthur and Sonia Labatt Brain Tumour Research Centre, Hospital for Sick Children, 555 University Avenue, Toronto, Ontario, Canada, M5G 1X8; 9Dana Farber Cancer Institute, 450 Brookline Avenue, Boston, MA, 02215, USA; 10Children's Hospital Boston, 300 Longwood Avenue, Boston, MA, 02115, USA; 11Department of Neuropathology, University of Heidelberg, Im Neuenheimer Feld 220, Heidelberg, 69120, Germany; 12Institute of Pharmacy and Molecular Biotechnology, and Bioquant Center, University of Heidelberg, Im Neuenheimer Feld 280, Heidelberg, 69120, Germany; 13Institute of Medical and Molecular Biotechnology, and Bioquant Center, University of Heidelberg, Im Neuenheimer Feld 267, Heidelberg, 69120, Germany; 14Department of Pediatric Oncology, Hematology & Immunology, Heidelberg University Hospital, Im Neuenheimer Feld 430, Heidelberg, 69120, Germany.

Background: Medulloblastoma is the most common malignant brain tumor in childhood. It arises in the cerebellum or medulla/brainstem, and shows tremendous biological and clinical heterogeneity. Despite advances in treatment for medulloblastoma over the past few decades, approximately 40% of children who develop this aggressively growing malignancy will experience tumor recurrence, and 30% will die from the disease. Importantly, recent work has shown that medulloblastoma is not a single disease, but comprises at least four distinct molecular subgroups [1]. WNT tumors, displaying activated wingless pathway signaling, carry a favorable prognosis. SHH tumors show hedgehog pathway activation, and have an intermediate prognosis. Group 3 and 4 tumors are molecularly less well characterized, and present the greatest clinical challenges. The full repertoire of genetic events driving this distinction, however, remains unclear.

We have recently described an integrative deep-sequencing analysis of 125 tumor-normal pairs, conducted as part of the International Cancer Genome Consortium (ICGC) PedBrain Tumor Project [2]. Here, we focus on genome-wide somatic mutations in medulloblastoma, how they are distributed throughout the genome, how they are correlated with patient age at diagnosis, the influence of subgroup affiliation on mutation rate, and how the mutation allele frequencies can be utilized to predict ploidy and infer temporal evolution of the tumor.

Materials and methods: We sequenced the complete genomes of 39 primary medulloblastoma and matched normal DNAs from the same
individuals, aged from 0 to 17 years, using Illumina technology. Cancer and normal DNAs were sequenced to an average of 35-fold coverage and analyzed to identify somatic base substitutions, small insertions, deletions, and copy number changes.

**Results:** Tetraploidy was identified as a frequent event in clinically challenging group 3 and 4 tumors. The extremely low fraction of mutations at approximately 50% allele frequency indicates that genome duplication occurred very early during tumorigenesis. For non-tetraploid tumors, a clear positive correlation of patients' age and mutation number was observed. The average somatic mutation rate was 0.52 per megabase (Mb), with an average of 10.3 non-synonymous coding single nucleotide variants, amounts considerably lower than in deep-sequenced adult malignancies. SHH tumors harboring TP53 mutations showed a significantly higher mutation rate both genome wide and for non-synonymous changes. Interestingly, the WNT subgroup, which typically shows a good prognosis, had the next highest mutation rate. Several recurrent coding mutations were identified, both in known medulloblastoma-related genes and in genes not previously linked to this tumor, often in subgroup-specific patterns. Chromatin modifiers were frequently altered across all subgroups.

**Conclusions:** Next-generation sequencing of this large cohort has provided a detailed insight into new mechanisms contributing to medulloblastoma tumorigenesis, enhancing our understanding of the genomic complexity and heterogeneity underlying medulloblastoma, and providing several potential targets for new therapeutics.

**References**

---

**P44**

**Modeling influenza sequence evolution for vaccination**
Tajiao Jiang1
Institute of Biophysics, Chinese Academy of Sciences, Beijing, China

BMC Proceedings 2012, 6(Suppl 6):P44

The rapid evolution of the influenza A virus poses a global challenge to public health. Recent events, such as the spread of highly pathogenic H5N1 influenza virus, have heightened concerns of potential pandemics. Thus, there is an urgent need for a better understanding of influenza virus evolution. Due to the development of high-throughput sequencing technologies, large-scale sequencing of influenza viruses has become routine work in influenza surveillance, and analyses of these large-scale viral sequence data have significantly enhanced our understanding of influenza evolution. However, opportunities remain to extract even more useful information to inform influenza prevention and control strategy. As we know, seasonal influenza prevention and control rely largely on the availability of effective vaccines. However, timely and accurate recommendation of vaccine strains is quite challenging, as evidenced by frequent antigenic mismatches between the recommended vaccine strains and circulating strains.

Our recent work has shown that development of sequence-based computational approaches to modeling the antigenic evolution of the influenza virus holds great promise for more effective vaccine strain recommendation [1,2]. Based on whole-genome information, we have developed a network model to represent each H3N2 virus as a nucleotide-to-occurrence network [1]. The network model effectively captures the evolutionary antigenic features of H3N2 virus at the whole-genome level and accurately describes the complex evolutionary patterns between individual gene segments. Our analyses show that the co-occurring nucleotide modules apparently underpin the dynamics of human H3N2 evolution, and that amino acid substitutions corresponding to nucleotide co-changes cluster preferentially in known antigenic regions of the viral surface protein hemagglutinin (HA). Therefore, our study demonstrates that nucleotide co-occurrence networks represent a powerful method for tracking influenza A virus evolution, and that cooperative genomic interaction is a major force underlying influenza virus evolution. Based on the sequences of HA, the major antigen of influenza virus, we have developed a computational method, denoted as PREDAC, to predict antigenic clusters of influenza A (H3N2) viruses with high accuracy from viral HA sequences [2]. Application of PREDAC to large-scale HA sequence data from H3N2 viruses isolated from diverse regions of mainland China identified 17 antigenic clusters that have dominated for at least one season between 1968 and 2010. By tracking the dynamics of the dominant antigenic clusters, we not only find that dominant antigenic clusters change more frequently in China than in the United States and Europe, but also characterize the antigenic patterns of seasonal H3N2 viruses within China. Furthermore, we demonstrate that the coupling of large-scale HA sequencing with PREDAC can significantly improve vaccine strain recommendation for China.

**References**

---

**P45**

**Apoptotic gene expression in melanoma cells treated with kaurenic acid**
Maria B Rivero1, Hernán J Mendoza1, Miguel A Chiurillo1, Miriam C Sosa-Sequera1
1Laboratorio de Genética Molecular, Yuri-Turbay Departamento de Ciencias Funcionales Medicina, Universidad Centrooccidental Lisandro Alvarado, Venezuela; 2Unidad de Farmacología Experimental, Departamento de Ciencias Funcionales Medicina, Universidad Centrooccidental Lisandro Alvarado, Venezuela

BMC Proceedings 2012, 6(Suppl 6):P45

Terpenoids have been described as pharmacologically active substances. Antibacterial, antifungal, anti-inflammatory activities and cytotoxicity against cancer cell lines are some of the biological effects described for these compounds [1,2,3,4]. Kaurenic acid is a diterpene isolated from the aerial parts of *Espeletia semiglobulata* (Compositae) and its antitumor effect via apoptosis and necrosis against melanoma cells in animal models has been described [5]. The present study researches the molecular mechanism for this anti-melanoma effect in cells inoculated in mice. One hundred thousand cells of melanoma B16F1 were used for inoculation in fifteen male, 20 g weight C57BL/6 mice. There were three groups of five mice each: group A was treated for 21 days with saline solution (0.009%/day); group B was treated for 21 days with Taxol (14.5 mg/kg/week); and group C was treated with kaurenic acid for 21 days (1 mg/kg/day). Two animals in group C died during the treatment. The expression of genes for proteins Bcl-2, Bax-a, Bcl-xL, c-Myc, P53, ICE (caspase 1), ICH1 (caspase 2), CPP32 (caspase 3), Apa F1 (activator), TNF-a, eNOS, iNOS, NOS, Flice (caspase 8), MCH6 (caspase 9) involved in different apoptotic pathways, was qualitatively assessed by using Multiplex PCR Maxim Biotech Inc. Kits for Mouse Apoptotic Genes, with GAPDH and 18S as internal controls in cDNA synthesized by Trizol method [6]. Each PCR was repeated twice in the same conditions for each sample. Amplification products were visualized in agarose gels cut with ethidium bromide and illuminated with ultraviolet light. The results showed that there was no expression of Bcl-xL in any of the animals from groups B and C. Furthermore, there was no expression of iNOS, nNOS and eNOS in samples from group C. Other evaluated genes were present in all groups. The alteration in the expression of the Bcl-xL gene (anti-apoptotic protein) and nitric oxide family proteins could be crucial events for the anti-melanoma effect. Kaurenic acid could offer potential usefulness as an agent for therapy of this cancer.

**Acknowledgements:** Thanks to Alfredo Usobiaga for providing kaurenic acid and to Peter Taylor for tumor cells.

**References**
22-1127-9.

P46
Abstract not submitted for online publication.

P47
Abstract not submitted for online publication.

P48
Verification of systems biology research in the age of collaborative competition
Carine Poussin
Pathway Informatics, Philip Morris International Research & Development, Neuchatel, Switzerland
BMC Proceedings 2012, 6(Suppl 6):P48

Modern society demands greater scrutiny of the potential health risks and benefits of long-term, and sometimes lifelong, exposure to drugs, chemicals, and substances found in consumer products and the environment. Organizations such as companies and academic consortia conduct large multi-year scientific studies that entail the collection and analysis of thousands of data points. The individual experiments are often conducted over many physical sites and with internal and outsourced components. To extract maximum value, the interested parties need to verify the accuracy and reproducibility of automated collection and analysis workflows in systems biology before the initiation of large multi-year studies.

Traditional verification using the peer-review process has shortcomings, such as lack of scalability, which renders it insufficient for the assessment of high-throughput research. A team of researchers at Philip Morris International (PMI) and IBM, whose aim is to improve the effectiveness of scientific studies and the verification of scientific findings, propose a scheme called IMPROVER, for Industrial Methodology for Process Verification of Research. This methodology evaluates research programs by dividing its workflow into smaller building blocks, whereby the verification of each building block can be done internally or externally via challenge-based ‘crowd-sourcing’ to a research community.

Scientific challenges will be broadcast to potential stakeholders in the form of an open call for participation, with the intention of providing the community with the opportunity to test their computational methods on new data as well as to partake in a collaborative effort whose ultimate goal could contribute to solving a grand scientific problem.

Considering cancer as the leading cause of death worldwide, we formulate the Diagnostics Signature Challenge to evaluate novel approaches for the identification of robust and predictive signatures for this disease. The goal of a Diagnostics Signature Challenge is to verify that transcriptomics data contain enough information for the determination and prognosis of certain human disease states that could profit from better diagnostics signatures.

Here we will describe the approach, the necessary operational steps, and how we intend to engage the wider scientific community to assess the applicability of the IMPROVER approach to molecular diagnostics (that is, genomic signatures).

Acknowledgements: The SBV 3.0 IMPROVER Project (Industrial Methodology for Process Verification in Research) is a scientific collaboration between PMI and IBM’s Thomas J. Watson Research Center on a project funded by PMI.

P49
Dirichlet process model for joint haplotype inference and GWAS
Avinash Das Sahu 1,2, Sridhar Hannenhalli 1
1. University of Maryland, College Park, 20740, USA; 2. University of Maryland Institute for Advanced Computer Studies, USA
BMC Proceedings 2012, 6(Suppl 6):P49

Identification of causal genomic mutations that underlie disease phenotypes remains a key problem in the field of medical informatics. With the advent of new sequencing technologies and decreasing cost of human genotyping, it is now possible to study genotype-phenotype interactions, such as genome-wide association studies (GWAS), at the population level. However, due to large genomic variance and linkage disequilibrium, genetic diversity of a complete human population cannot be captured by a limited number of clusters. Furthermore, application of current haplotype inferencing (phasing) methods to rare genomic variance, such as disease-related alleles, is not reliable. Hence, a satisfactory method for deleterious mutation identification remains largely elusive. Here we present a non-parametric Bayesian model that jointly infers haplotypes and identifies deleterious mutations, taking into consideration genomic variance in the human population. The model is based on the Dirichlet process, which can capture genomic variance by modeling it with non-bounded numbers of clusters.

P50
Structural effect of P278A mutation conferring breast cancer susceptibility in the p53 DNA-binding core domain
Y Sureetha, C Kumaraswamy Naidu
Department of Zoology, Sri Venkateswara University, Tirupati, 517502, India
BMC Proceedings 2012, 6(Suppl 6):P50

One of the common malignancies faced by women around the world is breast cancer. Risk factors for breast cancer include both genetic and non-genetic. Variants in some of the candidate genes are a common risk factor in breast cancer. These genetic variants associated with breast cancer can be classified as high, moderate or low based on relative risk [1]. Among them, genes that predispose to high risk for breast cancer include TP53, BRCA1, BRCA2, PTEN, STK11 and CDH1. A large number of studies have assessed the prognostic and predictive role of TP53 alterations in breast cancer. It is well known that TP53 is mutated in about 30% of breast cancers [2]. We have analyzed the genetic variation that may alter the expression and function of the TP53 gene using the sequence-homology-based SIFT tool [3] and a structure-based approach using the PolyPhen-2 server [4]. These two computational approaches showed that rs17849781 (P278A) has a deleterious phenotypic effect conferring to breast cancer. Further, we have analyzed the structural effect of the P278A mutation in the p53 DNA-binding core domain by employing different computational methods.

References

P51
Simple and precise targeted editing of the human genome using rAAV-mediated homologous recombination
Christopher J Torrance
Horizon Discovery Ltd, Cambridge Research Park, Cambridge, CB25 9TL, UK
BMC Proceedings 2012, 6(Suppl 6):P51

The rate limiting step in translating normal and pathological human genetic variations into a deeper understanding of which are functionally relevant, and the development of genetically defined disease models
that can speed up and rationalize the discovery of novel targeted therapies, now lies firmly with our ability to quickly and precisely alter any sequence in the human genome, just as we have been able to do in mice and other lower organisms for many years. However, routinely altering any user-defined endogenous DNA sequence in somatic human cells has been historically challenging due to their uniformly low levels of homologous recombination.

Horizon’s GENESIS™ gene-editing platform (http://www.horizondiscovery.com) uses the unique ability of targeted recombinant adeno-associated viral (rAAV) vectors to naturally stimulate and piggyback on the high-fidelity homologous recombination DNA-repair pathway in human cells, without the requirement for eliciting double-stranded DNA breaks using exogenous nucleases, to precisely engineer any DNA variation of choice into any human cell line, including the introduction of subtle point mutations and single nucleotide polymorphisms (SNPs), just as they occur in real patients.

The ability to rapidly, accurately and stably engineer human genomes without introducing confounding off-target effects has the potential to transform the field of functional genomics and the design of future targeted, or personalized, therapies. Data will be presented highlighting the use of GENESIS to create genetically defined ‘X-MAN’ (gene X, mutant and normal) human isogenic cell-line pairs, which accurately span specific genetic variations present in a defined cancer patient population, and provide a matched normal genetic background for reference, in key application areas. First, defining the core phenotype driven by specific cancer genes; second, profiling the activity and therapeutic window of novel cancer therapies on specific target patient genotypes; third, systematically de-organizing the cancer genome with next-generation ‘synthetically lethal’ drug targets by combining with RNA interference and chemical genomics tools; and fourth, their use as gold-standard genomic reference materials for commercial diagnostics and assays.

**Results:**

None of the 37 identified indels could be confirmed, probably due to indel errors generated by the Ion PGM platform [2]. Three of the four missense alterations were not identified in any of the 99 controls, and therefore were classified as possibly associated with WT: two at APC (Ile2541Val and Met1413Val) and one at PCLG2 (Asn946Ser). No point mutation was observed for WT1 and CTNNB1 genes. In the intronic regions, the somatic substitution pattern was evaluated by classifying the tumor’s substitutions in two groups: single nucleotide polymorphisms (SNPs) (alterations also present in normal samples), and somatic substitutions (remaining alterations). A general trend in the somatic base substitution pattern was observed, with the tumors presenting an over-representation of G:C>A:T changes (P >0.001) and a reduction of A:T>G:C changes (P >0.001) when compared to the SNP variants. Regarding the frequency of the other classes of somatic substitutions, a wide variation was observed among the different tumors.

**Conclusions:** This study provides insights into the spectrum of WT substitution mutations occurring in exonic and intronic regions of five previously associated genes, revealing a low frequency of point mutations in the coding sequence of these genes and a general over-representation of somatic G:C>A:T transitions.

**References**


---

**PS2**

**Mutational spectrum of WTX, WT1, CTNNB1, APC and PCLG2 genes in Wilms tumor defined by massive parallel resequencing**

Bruna D Barros1, Giovana T Torrezan1, Elisa N Ferreira1, Mariana Maschiott1, Ana CV Krepsch1, D cute M Carraro1,1 CICE - International Center of Research and Training - A. C. Camargo Hospital, São Paulo, SP, Brazil; 1IMCB Unit, UCL Institute of Child Health, London, UK

**Background:** The identification of molecular alterations that trigger Wilms tumor (WT) development is crucial to understanding the tumorigenesis of this malignancy. Currently, it is estimated that WTX and WT1 genomic losses together with CTNNB1/WT1 loss account for about 30% of WTs. However, the majority of cases remain without any identified driver mutation. Results from a previous study by our group pointed to APC and PCLG2 as candidate genes altered in WT [1]. Given the advent of modern DNA sequencing technologies, it is now feasible to evaluate large genomic regions spanning complete genes (exons and introns), allowing the description of the mutation patterns occurring in tumor cells. Thus, the aim of this study was to identify point mutations and indels in the complete sequence of APC, CTNNB1, WTX, WT1, WTX and PCLG2 genes in order to characterize both the exonic mutational spectrum and the intronic nucleotide substitution pattern.

**Material and methods:** The complete genomic regions of the selected genes, spanning a total of 430 kb, were amplified by long-range PCR in 15 WTs and 3 non-neoplastic control samples, giving a total of 60 amplicons per sample (10 kb on average). The resulting amplicons were mixed at equimolar concentrations and, for each sample, the Ion PGM library preparation protocol was performed. The libraries of the 18 bar-coded samples were combined in four sequencing pools that were individually submitted to an Ion PGM™ Sequencer run on an Ion 316™ Chip. Point mutation and indels not present in the non-neoplastic controls were selected for capillary sequencing validation.

**Results:** For the 15 tumors and 3 normal controls, we obtained 400,000 sequences per sample on average (350× mean target coverage). At the exonic regions, four out of the four identified missense alterations were validated. None of the 37 identified indels could be confirmed, probably due to indel errors generated by the Ion PGM platform [2]. Three of the four missense alterations were not identified in any of the 99 controls, and therefore were classified as possibly associated with WT: two at APC (Ile2541Val and Met1413Val) and one at PCLG2 (Asn946Ser). No point mutation was observed for WT1 and CTNNB1 genes. In the intronic regions, the somatic substitution pattern was evaluated by classifying the tumor’s substitutions in two groups: single nucleotide polymorphisms (SNPs) (alterations also present in normal samples), and somatic substitutions (remaining alterations). A general trend in the somatic base substitution pattern was observed, with the tumors presenting an over-representation of G:C>A:T changes (P >0.001) and a reduction of A:T>G:C changes (P >0.001) when compared to the SNP variants. Regarding the frequency of the other classes of somatic substitutions, a wide variation was observed among the different tumors.

**Conclusions:** This study provides insights into the spectrum of WT substitution mutations occurring in exonic and intronic regions of five previously associated genes, revealing a low frequency of point mutations in the coding sequence of these genes and a general over-representation of somatic G:C>A:T transitions.

**References**


---

**PS3**

**RNA-Seq analysis of two brain regions vulnerable to Alzheimer’s disease**

Xinkun Wang1,2,3

1Genomics Facility, The University of Kansas, Lawrence, KS, 66047, USA
2Higuchi Biosciences Center, The University of Kansas, Lawrence, KS, 66047, USA
3Department of Pharmacology and Toxicology, The University of Kansas, Lawrence, KS, 66047, USA

**BMC Proceedings 2012, 6(Suppl 6):P53**

**Background:** Alzheimer’s disease (AD), one of the most devastating neurodegenerative diseases, does not affect different brain regions equally. The temporal and frontal lobes are among the brain regions most affected in AD. As the transcriptomic profile of neurons in a certain brain region largely affects their response to pathological conditions like AD, comparative transcriptomic analysis of these susceptible regions can be used to understand why they are particularly vulnerable to the disease. With rapid advances in next-generation sequencing technologies, RNA-Seq, or whole-transcriptome shotgun sequencing, has begun to become a mainstream approach to study brain regions that are affected by neurodegenerative diseases, including AD.

**Materials and methods:** In this study, RNA-Seq data that have recently become available [1] from both normally aged and AD brain temporal and frontal lobes (Sequence Read Archive [http://www.ncbi.nlm.nih.gov/sra]; accessions SRX035166, SRX035171, SRX035167 and SRX034874) were analyzed, in order to provide molecular insights into their common vulnerability while accounting for their regional specificities.

**Results:** Here we present transcriptomic similarities and differences between the temporal and frontal lobes as detected by RNA-Seq. Interpretation of their similarities helps understand their shared vulnerability to AD. Detection of their differences in both normal aging and AD helps elucidate the progression of this disease in the two different regions.

**Conclusion:** The study of AD from the perspective of selective regional vulnerability is the first step towards minimizing its devastating effect on patients through protecting vulnerable brain neurons.

**Acknowledgments:** The author is supported by NIH grants including RIGMS 1F20GM103638-01, NIA P30 AG035982, NIA P01 AG12993 and NICHD P30 HD02528, and the Miller-Hedwig and Wilbur Fund.
The development of next-generation sequencing (NGS) technology opens new avenues for clinical researchers to make discoveries, especially in the area of clinical diagnostics. However, combining NGS and clinical data presents two challenges: first, the accessibility to clinicians of sufficient computing power needed for the analysis of high volume of NGS data; and second, the stringent requirements of accuracy and patient information data governance in a clinical setting.

Cloud computing is a natural fit for addressing the computing power requirements, while Clinical Laboratory Improvement Amendments (CLIA) certification provides a baseline standard for meeting the demands on researchers in working with clinical data. Combining a cloud-computing environment with CLIA certification presents its own challenges due to the level of control users have over the cloud environment and CLIA's stability requirements. We have bridged this gap by creating a locked virtual machine with a pre-defined and validated set of workflows. This virtual machine is created using our Galaxy VM launcher tool to instantiate a Galaxy environment at Amazon with specific versions of the tools used in the workflow. The VM launcher tool can reliably recreate the same virtual machine on several cloud environments. Once a baseline virtual machine is created, the tool can launch any number of clones to analyze samples in parallel. We describe herein a pilot project as an example of a working clinical analysis pipeline. In order to validate the clinical diagnosis of diseases with a genetic cause using NGS data, patient samples were collected by Dr Bharat Thyagarajan and staff at the Molecular Diagnostics Laboratory, University of Minnesota medical center–Fairview. The patient samples were analyzed using customized hybrid-capture bait libraries to boost read coverage in low-coverage regions, followed by targeted enrichment sequencing at the BioMedical Genomics Center. The NGS data is imported to a tested Galaxy single nucleotide polymorphism (SNP) detection workflow in a locked Galaxy virtual machine on Amazon's Elastic Compute Cloud (EC2). This project illustrates our ability to carry out CLIA-certified NGS analysis in the cloud, and will provide valuable guidance in any future implementation of NGS analysis involving clinical diagnosis.