increasing genotype and phenotype data every 2–3 mo.

Key Words: genomics, beef cattle, multibreed

FUNCTIONAL ANNOTATION OF ANIMAL GENOMES (FAANG) ASAS-ISAG JOINT SYMPOSIUM

0411 Important lessons from complex genomes.

T. R. Gingeras*, *Cold Spring Harbor Laboratory, Functional Genomics, NY.*

The approximately three billion base pairs of the human DNA represent a storage devise encoding information for hundreds of thousands of processes that can go on within and outside of a human cell. This information is revealed in the RNAs that are composed of 12 billion nucleotides considering the strandedness and the allelic content each of the diploid copies of the genome. Results stemming from the efforts to catalog and analyze the RNA products made by cells in the human (ENCODE), fly-worm (modENCODE) and mouse ENCODE projects have shed light on both the functional content and how this information is organized by various genomes. In human cells, a total of ~161,000 transcripts present within ~50,000 genic regions represent our previously best manually-curated annotation (based on v 7 Gencode) of the transcriptome. The results from the ENCODE project point to considerable supplementation of these data. Analyses of these transcriptome data sets have resulted in important and under appreciated lessons such as: (1) pervasive genome-wide transcription prompts a need to redefine the definition of a gene, (2) expression ranges follow transcript types and subcellular localization, (3) expression of isoforms of a gene by a cell do not follow a minimalistic strategy, and (4) genomic characteristics of potential *trans*-acting enhancer regions are distinguishable from other types of *cis*-acting regulatory regions. These and other lessons drawn from the landscape of both coding and non-coding RNAs present in eukaryotic cells have been used to assist in understanding and organizing what is often seen as dauntingly complex genomes.

Key Words: annotation, ENCODE, transcriptome

0412 Causal inference of molecular networks integrating multi-omics data.

F. Peñagaricano*, *University of Florida, Gainesville.*

Recent developments of massively parallel technologies allow assaying different biological molecules at very high throughput rates, including sequencing and genotyping of DNA, quantifying whole-genome gene expression, including measuring mRNA and microRNA abundance, identifying genome-wide epigenetic modifications, such as DNA methylation, and measuring different proteins and cellular metabolites. These advancements provide unprecedented opportunities to uncover the genetic architecture underlying phenotypic variation. In this context, the main challenge is to decipher the flow of biological information that lies between the genotypes and the phenotypes under study; in other words, the new challenge is to integrate multiple sources of molecular information, i.e., multiple layers of omics data, to reveal the causal biological networks that underlie complex traits. It is important to note that knowledge regarding causal relationships among genes and phenotypes can be used to predict the behavior of complex systems, as well as to optimize management practices and selection strategies. Here, we describe a multistep procedure for inferring causal gene-phenotype networks underlying complex phenotypes integrating multi-omics data. We initially assess marginal associations between genotypes and either intermediate phenotypes (such as gene expression) and endpoint phenotypes (such as carcass fat deposition and muscularity), and then, in those genomic regions where multiple significant hits co-localize, we attempt to reconstruct molecular networks using causal structural learning algorithms. These algorithms attempt to infer networks assuming that the pattern of conditional independencies observed in the joint probability distribution of these set of correlated variables are compatible with the unknown causal model. As a proof of principle of the significance of this integrative approach, we show the construction of causal molecular networks underlying economically relevant meat quality traits in pigs using multi-omics data obtained from an F2 Duroc x Pietrain resource population. Interestingly, our findings shed light on the mechanisms underlying some known antagonist relationships between important phenotypes, for instance, carcass fat deposition and meat lean content. More generally, the proposed methodology allows further learning regarding phenotypic and molecular causal structures underlying complex traits in farm species.

Key Words: causal inference, graphical models, systems biology

0413 Genotypes to phenotypes: Lessons from functional variation in the human genome and transcriptome.

B. E. Stranger*, *Section of Genetic Medicine, Department of Medicine, Institute of Genomics and Systems Biology, Center for Data Intensive Sciences, University of Chicago, IL.*

Complex trait association mapping in humans has successfully identified genetic loci influencing trait variation for hundreds of different phenotypes, including disease. The vast majority of associated loci localize to non-coding regions of the genome, suggesting possible effects on gene regulatory mechanisms. Without a clear understanding of the regulatory code of the human genome, deep characterization of the molecular function(s) of genetic variants in the human genome has become increasingly important for defining that code and for understanding genetic associations to complex traits. Studies of the human

transcriptome, its complexity, and its relation to genetic variation in a variety of contexts have proven highly informative for understanding genome function and for suggesting testable hypotheses involving candidate genes for complex traits and the functional mechanisms though which they may act. These approaches are increasingly leading to successful functional characterization of trait-associated variants, in some cases, suggesting possible targets for trait manipulation. Finally, these characterizations are being used to build models predicting variant function, further extending possible applications.

Key Words: genome function, non-coding variants, regulatory mechanisms

---

0414 Recurrent chimeric transcripts in human and mouse.

S. Djebali[*1,2,3], B. Rodríguez Martín[2,3], E. Palumbo[2,3], D. D. Pervouchine[2,3], A. Breschi[2,3], C. Davis[4], A. Dobin[4], G. Alonso[5], A. Rastrojo[5], B. Aguado[5], T. R. Gingeras[4], and R. Guigó[2,3], *[1]GenPhySE, INRA, Castanet-Tolosan, France, [2]Universitat Pompeu Fabra (UPF), Barcelona, Spain, [3]Bioinformatics and Genomics Program, Centre for Genomic Regulation (CRG), Barcelona, Spain, [4]Cold Spring Harbor Laboratory, Functional Genomics, NY, [5]Centro de Biología Molecular Severo Ochoa (CSIC- UAM), Madrid, Spain.*

The formation of chimeric transcripts (chimeras) has been widely reported. Some of them reflect underlying chromosomal rearrangements, or are the results of the propensity of reverse transcriptase to engage in template switching, however, a proportion of cases genuinely appear to correspond to *trans*-splicing of RNAs, as has previously been described.

Here we use ENCODE and mouse ENCODE deeply sequenced and bio-replicated RNaseq data from 18 human and 30 mouse samples, and the ChimPipe program, to identify chimeras occurring in multiple biological samples (recurrent), and between the same pairs of genes in human and mouse, since they are more likely to be transcriptionally induced and functional.

Recurrent common chimeras tend to connect gene pairs located on the same chromosome and relatively near to each other (< 100kb), therefore pointing to polymerase read-through, however interchromosomal chimeras are also observed, pointing to *trans*-splicing. Importantly, these recurrent chimeras tend to maintain an open reading frame, and could therefore generate chimeric proteins. We also observe that not only the gene-to-gene connection is conserved, but strikingly so are specific junction sites. The genes connected in common chimeras tend to be involved in morphogenesis and body plan formation, and consistently tend to be detected in cell lines of embryonic origin.

Validation of human chimeras by RT-PCR yielded a success rate of 50%, and subsequent cloning and sequencing revealed novel transcript structures, of which some preserve the domains from the two parent genes. Applying this method to multiple animal species and breeds will help us understanding chimera evolution as well as reveal some links between genotype and phenotype.

Key Words: chimeras, transcripts, *trans*-splicing

0415 Improving genomic selection across breeds and across generations with functional annotation.

B. Hayes[*1], A. J. Chamberlain[2], H. Daetwyler[3], C. J. Vander Jagt[2], and M. E. Goddard[4], *[1]Department of Economic Development, Melbourne, Australia, [2]Dairy Futures Cooperative Research Centre, Bundoora, Australia, [3]Department of Economic Development, Jobs, Transport and Resources, Bundoora, Australia, [4]Department of Primary Industries, Melbourne, Australia.*

Identification of causal mutations which affect complex traits in livestock (including production, health and fertility) could accelerate genetic gains for these traits by improving the accuracy of genomic estimated breeding values, particularly across breeds and with greater persistency of accuracy across time. Identification of these causal mutations could also reveal facets of the biology underlying such traits. A significant proportion of the genomic variation in cattle, for *Bos taurus* breeds at least, has been identified. The 1000 bull genomes project now includes whole genome sequences from 1682 cattle of 55 breeds, from which 67.3 million variants (64.8 million SNP, 2.5 million indel) have been identified. The challenge is now to determine which subset of these variants affect complex traits. This challenge is magnified by the fact that the size of effects of the causal mutations are likely to be small, given the large number of mutations typically affecting complex traits. We propose that an approach that includes (1) a multi-breed reference population (necessary to break down the extensive linkage disequilibrium that exists within many livestock breeds), (2) intermediate phenotypes, such as gene expression and protein abundance, where mutation effect is much larger than on the complex trait phenotype, (3) genome annotation information, to identify which classes of variants are more likely to affect complex traits, and (4) a genomic prediction algorithm that uses all this information simultaneously, will lead to identification of causal mutations on a genome-wide scale. Several examples identifying potential causal mutations affecting milk composition from dairy cattle are given. The results highlight the need for better annotation of the bovine genome—many of the most significant mutations are in poorly annotated genomic regions, likely regions regulating gene expression. The functional annotation of animal genomes (FAANG) consortium will greatly improve this situation.

Key Words: genomic selection, functional annotation

0416 Integrating dynamic omics responses for universal personalized medicine. G. I. Mias*, *Michigan State University, East Lansing.*

The advent of readily available omics technologies, and the recent Precision Medicine Initiative announced by the White House and National Institutes of Health are guiding our efforts to make advances in the implementation of personalized medicine. High quality genomes are now complemented with other dynamic omics data (e.g., transcriptomes, proteomes, metabolomes), that may be used to profile temporal patterns of thousands of molecular components in individuals. We are pursuing the profiling of multiple such omics in parallel $n = 1$ studies that extend the pilot integrative Personal Omics Profiling (iPOP) approach to diseases affecting the immune system. In particular, we will describe our investigations that follow longitudinally healthy and asthmatic individuals, and the integration of multiple omics obtained from peripheral blood cells, that we believe may provide novel medical insights. Concurrently, we are developing the necessary statistical and computational methodology for integrating the different omics platforms toward a medical interpretation, including our MathIOmica framework. Our approach enables us to query RNA sequencing, mass spectrometry (proteomics/metabolomics) and any longitudinal omics data, starting from lab samples to raw data, and including downstream quantitation methods for each analysis. We will present a clinically relevant classification scheme of longitudinal patterns, integration that accounts for missing data and uneven time sampling, and ultimately a biological interpretation and dynamic visualization of an integrated profile. Additionally, we are developing the necessary experiments and data sets for future iPOP investigations, with dense profiling of cell-drug treatment responses utilizing Rituximab and other interventions. Our combined transcriptome-proteome profiles enable us to reconstruct dynamic pathways of Rituximab's action on B-cells on a global scale. In summary, our clinical, laboratory and computational investigations are providing the next steps in the development of omics data generation and integration, toward a universal personalized medicine implementation.

G.I.M. and research reported in this presentation are supported by grants from MSU and the National Human Genome Research Institute of the National Institutes of Health under Award Number 4R00HG007065. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Key Words: disease, personal omics profiling, transcriptome-proteome profiles

0417 A review of sequencing and assembly methods that enhance computational use. W. C. Warren*, *McDonnell Genome Institute, Washington University School of Medicine, St Louis, MO.*

In essence high quality genome references are proven to be a necessity to enable research on so many levels of biological investigation including disease etiology, small molecule drug screening and interactions, canonical disease pathway manifestation, and so many others. To date very few genomes can be classified as near finished, defined as only missing small regions that are recalcitrant to known molecular biology methods. Ultimately our goal is to produce contiguous chromosomes for genomes de novo at the lowest cost. So far most published de novo genome assemblies are derived from deep coverage Illumina only sequencing, most often utilizing two popular but independent assembly algorithms, yet all are documented to be inadequate for numerous types of genetic investigation. During this surge of short reads genome assembly new long read sequencing technology arrived, albeit at considerable cost, ~6-fold higher than pure Illumina de novo assembly approaches. However, long reads, now averaging ~14 kb in length, have transformed our ability to capture most chromosomes that compel us to fund these approaches to obtain higher quality. Our lab and others now routinely assemble human genomes with N50 contig lengths of 10 Mb and up to 53 Mb size contigs, contigs defined as uninterrupted consensus sequence. In our studies we have seen how an incomplete genome sequence was hindering studies designed to detect signatures of selection in the poultry industry, such as missing microchromosome sequence assignments and partial or completely missing gene models in the chicken. In the chicken, despite the use of older long read sequencing technology (average read length of 8 kb), we observed an increase of ~180 Mb in assembled size, added 1920 new gene models and reduced gaps by sevenfold among ordered chromosomes. Given the intense interest in better genome reference models, I will review the generally compartmentalized phases for producing high quality genome references and provide examples of analysis outcome.

Key Words: assembly, genome reference, long reads