C01: Computer Demo 1

NCBI Variation Portal: Databases and Tools for Genetic Variation Discovery, Analysis, and Interpretation

Lon Phan, NIH/NLM/NCBI, Bethesda, MD

The Variation Portal site (http://www.ncbi.nlm.nih.gov/variation/) at NCBI is a gateway for users to access databases and tools that can be used in the fields of genomics, genetics, and management of variation data for human and over 300 organisms. There are five NCBI databases that archive, analyze, display, and report information about germline and somatic variants and the relationship of these variants to phenotype and clinical significance. dbSNP houses short variations; dbVar houses large scale genomic variants; the database of Genotypes and Phenotypes (dbGaP) houses genotypes and phenotypes associations; ClinVar houses reported relationships between human variation and phenotypes; and the Genetic Testing Registry (GTR) provides a central location for accessing inherited and somatic genetic variations that are being tested for a specific trait or disorder. The Variation Portal site also includes tools to explore and expedite the analysis of submitted variant and NCBI annotated information. The Variation Viewer allows users to search, navigate, and view variations in genomic and gene context. The Variation Reporter accepts uploaded VCF files and generates a comprehensive report that includes molecular consequences, allele novel to NCBI, and information from NCBI's databases. NCBI Remap is a tool that allows users to project annotation data from one coordinate system to another. In addition, dbSNP, dbVar, and ClinVar provides a VCF report that can be used for filtering large sets of variation. This presentation will provide demonstrations to navigate, search, view, and retrieve the large volume of data from these public NCBI databases and tools.

C02: Computer Demo 1

Increasing Discoverability and Connectivity of Scientific Media through Annotation with iCLiKVAL

Todd D. Taylor and Naveen Kumar, RIKEN Center for Integrative Medical Sciences, Yokohama, Japan

Scientific media comes in a variety of languages and formats, including journal articles, books, images, videos, blog and database entries, etc. In the case of textual media, there is often additional information such as tables, figures and supplementary data, associated with or embedded in the text. While are there are many good resources for browsing, searching and annotating some of this media, there is not one place where you can search everything, and generalized search engines such as Google do not allow for the type of comprehensive and precise searches that researchers require. To address this, we created iCLiKVAL (http://iclikval.riken.jp/), an easy-to-use web-based tool that uses the power of crowdsourcing to accumulate annotation information for all scientific media found online (and potentially offline). Annotations in the form of key-relationship-value tuples (in any language), added by users with a variety of options, make information easier to find and allow for much richer data searches. Users can create or join common interest groups, both public and private, to annotate related media together as a community. Users can also create and edit their own controlled vocabulary lists, or import established vocabularies such as MeSH and GO. Within the user groups, vocabulary and bookmark lists can easily be shared. iCLiKVAL is an open-access online tool. While the software itself is not open-source, the database is completely searchable, and all of the collected data is freely available to the research community via our API.

C03: Computer Demo 1

DBCLS SRA: Functional Characterization of Public NGS Data

Takeru Nakazato, Tazro Ohta and Hidemasa Bono, Database Center for Life Science, Mishima, Japan

High-throughput sequencing, also called next-generation sequencing (NGS), makes it easy to perform omics analysis with non-classical model organisms. DDBJ, EBI, and NCBI capture a large amount of NGS data, and launch public database, the Sequence Read Archive (SRA). As of October 2015, 4 peta bp of reads are archived in about 60,000 projects. We categorized SRA data by study types (e. g. whole genome study, transcriptomics, and metagenomics), sequencing platforms, and species or cell lines of samples. To visualize these information, we developed a web-service called DBCLS SRA (http://sra.dbcls.jp/). By using DBCLS SRA, researchers can search NGS data of their interest including child taxonomies at once. For example, they can retrieve *Oryza sativa indica* or *japonica* data by searching *Oryza sativa*. Users also list up other *Oryza* species data as a parent taxonomy.

C04: Computer Demo 1

Publicly Available Resources for Plant Genomics Research at the Arabidopsis Information Resource (TAIR).

Leonore Reiser, Phoenix Bioinformatics, Redwood City, CA

The Arabidopsis Information Resource (TAIR, https://www.arabidopsis.org) is a continuously updated, manually curated genome database for the model plant *Arabidopsis thaliana*. Since losing federal funding in 2013, TAIR has found alternative funding to continue gathering new gene function information and updating the 'gold standard annotation' for this important reference plant genome. Although TAIR now requires subscriptions to access recently curated data, it remains a nonprofit project and many essential functions remain freely accessible without a subscription. In this demonstration, we will show you how to use public resources in TAIR including registration, data searches, stock ordering, data submission and downloading large datasets. Registered users who are affiliated with labs can order seed, DNA and protein chip stocks from our partner, the Arabidopsis Biological Resource Center (ABRC). Registered users can also submit gene functional annotations, add community comments to TAIR detail pages and reserve gene names using our online tools. Curated data that have been in TAIR for one year are released on a quarterly basis and are openly accessible from the website. The exception is the set of GO annotations, which is updated on a monthly basis and made available to the community at the GO Consortium website.

C05: Computer Demo 1

TENOR: Database for Comprehensive mRNA-Seq Experiments in Rice

Yoshihiro Kawahara[1], Youko Oono[1], Hironobu Wakimoto[1,2], Jun Ogata[1], Hiroyuki Kanamori[3], Harumi Sasaki[1], Satomi Mori[3], Takashi Matsumoto[3] and Takeshi Itoh[1], (1)National Institute of Agrobiological Sciences, Tsukuba, Japan, (2)BITS. Co., Ltd., Tokyo, Japan, (3)National Institute of Agrobiological Sciences, Tsukuba, Ibaraki, Japan

Here we present TENOR (Transcriptome ENcyclopedia Of Rice, http://tenor.dna.affrc.go.jp), a database that encompasses large-scale mRNA-Seq data obtained from rice under a wide variety of conditions. Since the elucidation of the ability of plants to adapt to various growing conditions is a key issue in plant sciences, it is of great interest to understand the regulatory networks of genes responsible for environmental

changes. We used mRNA-Seq and performed a time-course transcriptome analysis of rice, *Oryza sativa* L. (cv. Nipponbare), under ten abiotic stress conditions (high salinity; high and low phosphate; high, low and extremely low cadmium; drought; osmotic; cold; and flood) and two plant hormone treatment conditions (abscisic acid and jasmonic acid). A large number of genes that were responsive to abiotic stresses and plant hormones were detected by differential expression analysis. Furthermore, several responsive genes were found to encode transcription factors that could control the transcriptional network of stress responses, but the timing of the induction of these genes was not uniform across conditions. A significant number of cis-regulatory elements were enriched in the promoter regions of the responsive genes and were shared among conditions. These data suggest that some key components of gene regulation networks are shared between different stress signaling pathways. All the resources (novel genes identified from mRNA-Seq data, expression profiles, co-expressed genes and cis-regulatory elements) can be searched for and are available in TENOR.

C06: Computer Demo 1

Use of KitBase to Facilitate Forward and Reverse Genetics Research in Rice

Rashmi Jain[1], Guotian Li[2], Mawsheng Chern[1], Nhan T. Pham[3], Joel Martin[4], Wendy Schackwitz[4], Anna Lipzen[4], Jeremy Schmutz[5], Kerrie W. Barry[4] and Pamela Ronald[6], (1)UC Davis/JBEI, Davis, CA, (2)UC-Davis, davis, CA, (3)UC-Davis, Davis, CA, (4)DOE Joint Genome Institute, Walnut Creek, CA, (5)HudsonAlpha Institute for Biotechnology, Huntsville, AL, (6)Joint BioEnergy Institute (JBEI), Emeryville, CA

Mutagenized populations are useful for studying gene function in plants. We have generated a mutant population in the model rice cultivar Kitaake using fast-neutron mutagenesis. Kitaake is an early flowering, short-statured, short life cycle rice that is easy to transform, compared with other *Japonica* and *Indica* rice varieties. In collaboration with Joint Genome Institute, we are sequencing 4,000 mutants. Genomic analysis of more than 1,000 mutants has been done, revealed that 13,469 genes are affected. Mutation types include single base substitutions, deletions, insertions, inversions, translocations, tandem duplications and complex events. Single base substitutions predominate, but deletions affect the greatest number of genes, accounting for 57.2% of all affected genes. Using the same FN population we have carried out a forward genetic screen for mutants that suppress XA21-mediated immunity. We identified ten mutants from this screen and in several cases, have isolated the affected gene. KitBase, a comprehensive repository for rice mutant information is publically available to the scientific community. KitBase integrates JBrowse and BLAST to facilitate identification of mutations and searching of the database. The first phase of KitBase includes genomic data, phenotypic data, and seed information for each of the mutant lines.

C07: Computer Demo 1

TreeGenes and CartograTree: Community Resources for Forest Tree Genomics

Steven A Demurjian Jr, University of Connecticut, Storrs, CT, Emily Grau, Department of Plant Sciences, University of California, Davis, Davis, CA, Hans Vasquez-Gross, University of California Davis, Davis, CA, Damian Gessler, University of Arizona, Tucson, AZ, David Neale, Dept. Plant Sciences University of California Davis, Davis, CA and Jill Wegrzyn, Department of Ecology and Evolutionary Biology, University of Connecticut, Storrs, CT

Changes in the frequency, duration, or severity of drought and heat stress associated with climate change could modify the composition, structure, and biogeography of forests in many regions. It is critical to identify vulnerable populations as well as individuals that are resistant to the pests, pathogens, and fluctuating temperatures that result from climate change. These identifications rely on reliable integration of high resolution genomic, phenotypic, and environmental data. We will discuss recent development in the TreeGenes project, and specifically the CartograTree interface which aims to deliver this level of integration to the research community. TreeGenes is a web-based database that provides interfaces and tools to collect, curate, visualize, and analyze data from over 1700 forest tree species. The map-based CartograTree interface connects georeferenced trees with curated phenotypic and genetic data as well as environmental metrics from sources such as Ameriflux and WorldClim. Through the use of SSWAP (Simple Semantic Web Architecture and Protocol) web services, researchers are able to directly access high performance computing resources to run specific applications such as MUSCLE for sequence alignment or TASSEL for genetic association analyses. Current development has been focused on improving the integration of robust environmental layers by migrating from the Google Maps API to a vector-tile map delivery method and switching to a NoSQL database to improve performance. Future plans include improvements to data filtering and visualization prior to association analysis and full implementation of trait and environmental ontologies to extend the utility of the search interface.

C08: Computer Demo 2

Mango: An Environment for Combining Massive Heterogeneous Networks

Jennifer Chang, Hui-Hsien Chou and Hyejin Cho, Iowa State University, Ames, IA

Biological data include diverse types such as sequences, gene expressions, protein-protein interactions, genome-wise associations, biochemical and genetic pathways, etc. Heterogeneous data can be merged within graphs, or networks, associating data to one another. However, analyzing modern heterogeneous biological data can easily overwhelm existing software. The complexity and polymorphic nature of modern biological data sets requires a more standardized method for integration, exploration and inference.

The Mango network analysis and visualization software enables the integration and analysis of large heterogeneous networks. Mango is highly scalable; it can load and merge several large networks containing up to 4 million links on a desktop computer. Mango facilitates network analyses with its integrated Graph Exploration Language and real-time visualization of intermediate analysis results. The Graph Exploration Language automates graph attribute merging and promotion among many graph types and contains rich expressive syntax for graph analyses. Mango provides 3D visualization of multiple graphs, allowing users to map data attributes to visual effects. Many sophisticated graph analysis functions such as force-directed layouts, traversal and subsetting of graphs, direct online network database queries, and import and export options, are built into Mango. To sum up, Mango provides a standard, fast, flexible and powerful method for integrating and analyzing massive heterogeneous biological networks, which can significantly speed up serendipitous discoveries in complex data.

Mango is a C++ program that runs on 32 and 64 bit versions of Mac, Windows and Linux. Mango provides a standalone environment for graph analysis and is freely available from http://www.complex.iastate.edu/download/Mango.

C09: Computer Demo 2

Web Portal for Next Generation RNA-seq Sequence Computation and Analysis for Agricultural Animal Species

Weizhong Li[1], Robert W. Li[2] and Alexander Richter[1], (1)J Craig Venter Institute, La Jolla, CA, (2)ARS, USDA, Beltsville, MD

In order to address the computational challenges in RNA-seq data analysis, especially for researchers in the field of agriculturally important animal studies, we developed a bioinformatics web portal, which offers integrated workflows for RNA-seq NGS data computation and analysis. These workflows can perform end-to-end computational analysis, including sequence QC, reads mapping, transcriptome assembly and reconstruction and quantification, differential analysis, visualization and so on. One workflow mainly utilizes Tophat2, Cufflink, Cuffmerge and Cuffdiff suite of tools. Another workflow deploys Trinity for *de novo* assembly and uses RSEM for transcript quantification. Another workflow uses STAR and RSEM combination for computation. For all the workflows, the server supports analysis for multiple samples and multiple groups of samples and performs differential analysis between samples or groups, all in a single job submission. The server currently supports chicken, cow, duck, goat, guinea pig, horse, rabbit, sheep, turkey, as wells as c. elegan, fruit fly, human and yeast. The calculated results are available for download and post analysis.

The portal is implemented with several state-of-the-art HPC, workflow and web development software tools including Galaxy, StarCluster, OGS/GE utilizing modern scalable cloud compute and storage sources from AWS.

This demo will cover several end-to-end RNA-seq analysis projects. The demo includes uploading data files, choosing reference genomes, defining groups (e.g. case vs control), running the full workflow, downloading and visualizing results and post analysis.

The RNA-seq portal is freely available from http://weizhongli-lab.org/RNA-seq.

C10: Computer Demo 2

Trimmomatic: A Flexible Trimmer for Illumina Sequence Data

Anthony Bolger, RWTH Aachen, Aachen, Germany and Bjoern Usadel, Forschungzentrum Juelich & RWTH Aachen, Aachen, Germany

Read pre-processing is an important first step to optimize the analysis of next generation sequencing data by removing technical artifacts which may interfere with downstream analysis. It is particularly important in de-novo genome assembly projects, since technical sequences like adapters can easily be incorporated as part of the target genome.

Trimmomatic is primarily targetted at data from Illumina sequencers, and offers a range of sequence processing steps which can be flexibly combined as need to suit user requirements.

Paired end data is not only supported, but actually helps ensure highly accurate identification of the most common source of contamination: 'read-through' of short DNA fragments into the adapters. The tool also tracks the paired state of processed data to produces separate paired and unpaired files, ensuring that downstream tools which require paired data can be used without problems.

Two main approaches to quality trimming are provided, a standard 'sliding window' approach, and a 'maximum information' approach which balances the cost/benefits of additional, potentially erroneous data.

Additional processing steps for read cropping, filtering of entire reads by length or quality are also included.

The tool was originally developed as an in-house tool for the S. pennellii genome sequencing project, and proved both effective and efficient on the large datasets needed for this project.

C11: Computer Demo 2

GACD: Integrated Software for Genetic Analysis in Clonal F1 and Double Cross Populations

Luyan Zhang[1], Jiankang Wang[2], Lei Meng[2] and Wencheng Wu[2], (1)Instititue of Crop Sciences, Chinese Academy of Agricultural Sciences, Beijing, China, (2)Chinese Academy of Agricultural Sciences, Beijing, China

Clonal species are common among plants. Clonal F_1 progenies are derived from the hybridization between two heterozygous clones. In self- and cross-pollinated species, double crosses can be made from four inbred lines. A clonal F_1 population can be viewed as a double cross population when the linkage phase is determined. The software package GACD (Genetic Analysis of Clonal F_1 and Double cross) is freely-available public software, capable of building high-density linkage maps and mapping quantitative trait loci (QTL) in clonal F_1 and double cross populations. Three functionalities are integrated in GACD version 1.0: binning of redundant markers (BIN); linkage map construction (CDM); and QTL mapping (CDQ). Output of BIN can be directly used as input of CDM. After adding the phenotypic data, the output of CDM can be used as input of CDQ. Thus, GACD acts as a pipeline for genetic analysis. GACD and example datasets are freely available from www.isbreeding.net.

C12: Computer Demo 2

Integrated, Accurate and Multi-Environment Structural Variation Discovery from Whole Genome Sequencing Data with NGSEP

Jorge A. Duitama Castellanos and Juan Fernando De la Hoz, International Center for Tropical Agriculture (CIAT), Cali, Colombia

Structural Variants (SVs) are genomic differences between two or more samples that encompass large regions of their DNA. Whole genome sequencing (WGS) data can be effectively used to discover SVs, combining different algorithms specialized on different types of variants. However, integration of software tools for SV discovery is currently difficult because different tools are implemented in different languages, have different sets of requirements, and differ greatly in their parameter settings. We previously released NGSEP as a software tool that tightly integrates state-of-the-art algorithms for simultaneous discovery of SNVs, indels, and copy number variation (CNVs). Continuing this effort, we recently implemented in NGSEP three new algorithms for SV discovery: CNV-seq, for read-depth (RD) comparison between two samples; RDXplorer for detection of small CNVs; and a novel implementation of the read-pair (RP) and split-read (SR) strategies for detection of medium to large indels, including breakpoint resolution. Comparisons using benchmark datasets of yeast and rice shows that the combination of RD, RP and SR strategies integrated in NGSEP achieves superior accuracy and efficiency compared to individual approaches implemented in other currently available software packages. To facilitate the use of NGSEP we improved its documentation for command line usage and its integration

with the Galaxy environment. We also made NGSEP available within the iPlant collaborative platform and the DNAnexus cloud platform for SV discovery in the 3000 rice genomes project. We expect that our continuous development efforts facilitate the use of NGSEP for a growing number of researchers in plant and animal genomics.

C13: Computer Demo 2

Exploring Wheat Physical Maps and Genomic Data Using URGI Browsers

Thomas Letellier[1], Jane Rogers[2], Kellye Eversole[3], Frederic Choulet[4], Etienne Paux[4], Michael Alaux[1] and Hadi Quesneville[1], (1)INRA - URGI, Versailles, France, (2)International Wheat Genome Sequencing Consortium, Cambridge, United Kingdom, (3)IWGSC, Bethesda, MD, (4)INRA GDEC, Clermont-Ferrand, France

The wheat genome is complex and its analysis is a key challenge for agronomy and modern bioinformatics. To enable analysis of this important crop, the URGI (INRA research unit in genomics and bioinformatics dedicated to plants and crop parasites) developed customized browsers (GMOD tools) to display physical maps and reference sequence annotation data.

The IWGSC (International Wheat Genome Sequencing Consortium) sequences (survey sequences and 3B reference sequence) and physical maps data of each chromosome are hosted by the URGI Sequence Repository (http://wheat-urgi.versailles.inra.fr/Seq-Repository).

Using our physical maps browser (https://urgi.versailles.inra.fr/gb2/gbrowse/wheat_phys_pub/), the scientific community can browse contigs, BACs, markers, deletion bins, and WGP tag sequences as they become available.

We also host the wheat survey sequence annotation browser with polymorphism data and gene models (https://urgi.versailles.inra.fr/gb2/gbrowse/wheat_survey_sequence_annotation/). One can reach the wheat survey sequence annotation from our BLAST page (https://urgi.versailles.inra.fr/blast/).

Moreover, we setup a JBrowse to display 3B reference sequence annotation data such as mRNA, transposable elements, markers, QTLs, etc., and will be incorporating similar data for for reference sequences of other chromosomes as they are completed.

We provide a strong interoperability between data sources by linking the 3B annotation JBrowse to the 3B physical map. There are also many external links to complete searches: toward GnpIS (https://urgi.versailles.inra.fr/gnpis/, Steinbach *et al*. 2013 doi: 10.1093/database/bat058), a multi-species integrative information system that we developed at URGI or toward Wheat3BMine (http://urgi.versailles.inra.fr/Wheat3BMine/), a data warehouse dedicated to the wheat 3B chromosome.

C14: Computer Demo 3

The Pathway Tools Software for Metabolic Reconstruction and Modeling

Peter D Karp, SRI International, Menlo Park, CA

Pathway Tools [1] provides an array of capabilities for storing and analyzing integrated collections of genomic, metabolic, and regulatory information, and for deriving metabolic models from sequenced genomes. It supports development of organism-specific Pathway/Genome Databases (PGDBs) that integrate many bioinformatics datatypes. It supports querying, analysis, scientific visualization, and web publishing of those organism-specific DBs. It provides tools for analysis of omics datasets, comparative analysis, and metabolic engineering.

The MetaFlux module of Pathway Tools generates steady-state metabolic flux models from sequenced genomes using flux-balance analysis (FBA). MetaFlux contains a number of tools for accelerating model development for individual organisms and for organism communities, including a gap filler tool. Models can be executed using both FBA and Dynamic FBA, with gene knockouts. MetaFlux has been used to generate a very accurate *E. coli* metabolic flux model from EcoCyc.

Pathway Tools now includes the ability to integrate gene-knockout data and Phenotype Microarray data within a PGDB. A new RouteSearch capability enables users to search for optimal routes through the metabolic network that connect two metabolites, with the option of adding additional reactions from MetaCyc to support metabolic engineering. New web services allow metabolomics data to be painted onto pathway diagrams. A new pathway collage tool allows users to create diagrams of interconnected pathways painted with omics data.

[1] P.D. Karp et al, Pathway Tools version 19.0 update: Software for Pathway/Genome Informatics and Systems Biology, *Briefings in Bioinformatics* 2015 doi:10.1093/bib/bbv079.

C15: Computer Demo 3

Apollo: Improving Collaborative Genome Curation.

Monica C. Munoz-Torres[1], Nathan A Dunn[1], Deepak Unni[2], Colin Diesh[2], Christine G. Elsik[2], Ian Holmes[3] and Suzanna Lewis[1], (1)Lawrence Berkeley National Laboratory, Berkeley, CA, (2)Division of Animal Sciences, University of Missouri, Columbia, MO, (3)Department of Bioengineering, Berkeley, CA

Apollo enables collaborative, real-time curation (akin to Google Docs) of genomic elements using both structural and experimental information. Built on top of the JBrowse framework, Apollo is composed of a web-based client, an annotation-editing engine, and a server-side data service. Users can visualize gene models, protein alignments, and expression and variant data to conduct structural and/or functional annotations. In our most recent release, version 2.0.x, the improved architecture allows users to more easily query data and build extensions, supports multiple organisms per server, and allows additional types of sequence annotations based on the Sequence Ontology. The new, removable side-dock offers detailed view of annotations, sequences, and organisms, a new reporting structure, and WebSocket support to improve real-time communication. The new Grails framework (Spring / Hibernate / Groovy) in the server more robustly scales a single server over multiple organisms while better supporting additional curators. Apollo's entire secure REST API is exposed, allowing genomic features to be injected into Apollo from an automated curation process or organization-specific metadata to be extracted directly from Apollo using a SQL query or REST. The new version offers improved features, including the ability to bring together 2 or more scaffolds in order to annotate genes split across them, and increases the ability to customize and integrate Apollo into modern curation pipelines. During this demonstration we will introduce the new architecture, highlight advantages for users, and detail our future plans.

Project Website: http://genomearchitect.org/ **Source Code**: https://github.com/GMOD/Apollo **License**: Berkeley Software Distribution (BSD) License at https://github.com/GMOD/Apollo/blob/master/LICENSE.md

C16: Computer Demo 3

plantGPS: A Whole-Genome Modeling Framework to Accurately Predict Quantitative Traits

Yuri V. Nikolsky, George Mason University, Fairfax, VA and Tatiana Tatarinova, University of Southern California, Los Angeles, CA

Understanding the relationship between genomic variation and variation in phenotypes for quantitative traits such as physiology, yield, fitness or behavior, will provide important insights for predicting adaptive evolution and for breeding schemes. A particular question is whether the genetic variation that influences quantitative phenotypes is typically the result of one or two mutations of large effect, or many more mutations of small effect.

In the past, we developed GPS method to infer provenance for human subjects. Recently, we expended the approach to the wild model legume *Medicago truncatula*. We show that phenotypes, such as quantitative disease resistance, can be well-predicted using genome-wide patterns of admixture, from which it follows that there must be many mutations of small effect.

Our findings prove the potential of our novel "whole-genome modeling" – method and experimentally validate, for the first time, the infinitesimal model as a mechanism for adaptation of quantitative phenotypes in plants. This insight can accelerate breeding and biomedicine research programs.

C17: Computer Demo 3

Cool Season Food Legume Genome Database: an Up-to-Date Resource Enabling Genetics, Genomics and Breeding Research in Pea, Lentil, Faba Bean and Chickpea.

Jodi L. Humann[1], Sook Jung[1], Ping Zheng[1], Chun-Huai Cheng[1], Taein Lee[1], Morgan Frank[1], Deah McGaughey[1], Kristin Scott[1], Jing Yu[1], Stephen P. Ficklin[1], Marwa N.M.E. Sanad[2], Heidi Hough[1], Clare Coyne[3], Rebecca McGee[1] and Dorrie Main[1], (1)Washington State University, Pullman, WA, (2)National Research Center- Egypt, Pullman, WA, (3)USDA ARS, Pullman, WA

The new, mobile-friendly version of the Cool Season Food Legume Genome Database (CSFL, www.coolseasonfoodlegume.org) has been redesigned to allow for more efficient access to data, tools, and resources by users. The database has been updated with all current genetic maps, molecular markers, and QTL data in addition to the most current genome data for pea, lentil, chickpea, and faba bean. The new interface allows users to quickly search and retrieve data from the database. Quick access to popular tools which allow users to use BLAST for searches with current genome sequences and transcripts, view genomes in the JBrowse genome browser, make comparisons of genetic and physical map data with CMap, and view metabolic PlantCyc maps are also easily found from the website header. The ultimate goal of CSFL is to provide a single website where researchers can view/query/download all current genetics, genomics and breeding data for pea, lentil, chickpea and faba bean as well as have access to analysis tools that are useful for research. This project is supported by USDA NRSP10, the USA Dry Pea and Lentil Council, Northern Pulse Growers Association, USDA-ARS and Washington State University.

C18: Computer Demo 3

The Legume Information System and The Legume Federation: Working Together for the Legume-Fed World

Andrew D. Farmer[1], Alan Cleary[2], Alex G. Rice[1], Pooja E. Umale[1], Sam Hokin[3], Sudhansu Dash[1], Jacqueline D. Campbell[4], Wei Huang[4], Nathan T. Weeks[5], Andrew Wilkey[4], David Grant[5], Rex Nelson[5], Kevin H. Feeley[5], Vivek Krishnakumar[6], Akshay Yadav[4], Jeremy D. DeBarry[7], David Fernandez-Baca[4], Ethalinda Cannon[4], Christopher D. Town[6] and Steven B. Cannon[5], (1)National Center for Genome Resources (NCGR), Santa Fe, NM, (2)Montana State University, Bozeman, MT, (3)National Center for Genome Resources, Santa Fe, NM, (4)Iowa State University, Ames, IA, (5)USDA-ARS-CICGRU, Ames, IA, (6)J. Craig Venter Institute, Rockville, MD, (7)University of Arizona, Tucson, AZ

The significant increase in genomic and genetic data within the legume family has created tremendous opportunities for making use of cross-species comparative techniques to address concrete problems faced in modern agricultural breeding programs (e.g. increasing yield or introducing disease resistance). Conversely, the explosion of available data characterizing the diversity present within each crop species and its near relatives makes it increasingly difficult for any one group to adequately manage it all. We will present ongoing work aimed at addressing these challenges through a growing federation of groups serving different areas of the legume research community but united through: the adoption of common data and metadata standards, use of common protocols and cyber-infrastructure resources for data sharing, and a firm commitment to shared development efforts. We will demonstrate how we are approaching problems of integration across this diverse and economically important plant family with new tools in development for cross-species comparisons which make use of shared data sets such as the Phytozome gene family models, open source tools such as the Chado/Tripal framework, and common infrastructural resources provided through the iPlant collaborative. Our intent in featuring specific use cases will be to provide an incentive to other groups working within the legumes to consider aligning their efforts with ours, as well as to showcase tools available for adoption by groups working in other clade-based systems.

C19: Computer Demo 3

Updates to CottonGen: The Community Database for Genomics, Genetics and Breeding Research in Cotton

Jing Yu[1], Sook Jung[1], Chun-Huai Cheng[1], Taein Lee[1], Katheryn Buble[1], Ping Zheng[1], Jodi L. Humann[1], Deah McGaughey[1], Heidi Hough[1], Stephen P. Ficklin[1], B. Todd Campbell[2], Richard G. Percy[3], Don C. Jones[4] and **Dorrie Main**[1], (1)Washington State University, Pullman, WA, (2)USDA-ARS, Florence, SC, (3)USDA-ARS, Southern Plains Agricultural Research Center, College Station, TX, (4)Cotton Incorporated, Cary, NC

CottonGen (http://www.cottongen.org) is the community database for basic, translational and applied research in cotton. Developed using Tripal, an open-source, resource-efficient, standardized platform for biological database construction, it provides an online portal of curated and integrated genomics, genetics and breeding data, combined with a suite of tools facilitating intuitive data mining and analysis. We highlight new data and functionality in CottonGen, with a particular emphasis on application in breeding, and present future plans for development over the next 5 years for this industry, USDA, USDA-ARS funded resource.

C20: Computer Demo 3

GDR, the Genome Database for Rosaceae: New Data and Functionality

Sook Jung[1], Taein Lee[1], Chun-Huai Cheng[1], Stephen P. Ficklin[1], Anna Blenda[2], Ksenija Gasic[3], Jing Yu[1], Kristin Scott[1], Michael Byrd[2], Sushan Ru[1], Katherine M. Evans[4], Cameron Peace[1], Lisa DeVetter[1], Nnadozie Oraguzie[5], Albert G. Abbott[6], Mercy Olmstead[7] and Dorrie Main[1], (1)Washington State University, Pullman, WA, (2)Erskine College, Due West, SC, (3)Clemson University, Clemson, SC, (4)Washington State University, Wenatchee, WA, (5)Washington State University, Prosser, WA, (6)Forest Health Research and Education Center, University of Kentucky, Lexington, KY, (7)University of Florida, Gainesville, FL

The Genome Database for Rosaceae (GDR, http:/ www.rosaceae.org) is the central repository and data-mining resource for genomics, genetics, and breeding data of Rosaceae, an economically important crop family that includes almond, apple, blackberry, cherry, peach, pear, plum, raspberry, rose, and strawberry. GDR contains whole genome sequences and annotation, reference transcriptomes, gene sequences from NCBI, genetic maps, trait loci, germplasm, marker diversity, breeding and publication data. The predicted genes of the whole genome sequence, reference transcriptomes and NCBI genes have been further annotated by homology to genes in other species, InterPro protein domains, GO terms and KEGG pathway terms. In this computer demo, we update the users with the new data including new whole genome, genetic maps, trait loci and genotypic and phenotypic data from breeding projects. We also report our effort toward data standardization on gene names, QTL metadata, and trait ontology.

C21: Computer Demo 4

BovineMine: A Data Mining Warehouse for the *Bos taurus* Genome

Christine G. Elsik[1,2], Colin M. Diesh[1], Deepak R. Unni[1], Aditi Tayal[1], Hung N. Nguyen[2] and Darren E. Hagen[1], (1)Division of Animal Sciences, University of Missouri, Columbia, MO, (2)MU Informatics Institute, University of Missouri, Columbia, MO

BovineMine (http://BovineGenome.org/bovinemine) is a data mining and warehousing system based on InterMine. Datasets include bovine gene annotations (NCBI, Ensembl, bovine OGS), protein annotations (UniProt), protein families and domains (InterPro), homologs and orthologs (OrthoDB, TreeFam, EnsemblCompara, HomoloGene) pathways (Reactome), gene-gene interactions (BioGRID), Gene Ontology (GO), QTL (AnimalQTLdb), SNP (dbSNP), and tissue-specific gene expression (SRA). BovineMine provides reports for various entities, such as genes, transcripts, proteins and ontology terms, along with tools that allow users to analyze and download genome-wide datasets. A central feature of BovineMine is the QueryBuilder tool, which allows users to explore the data and construct custom queries that integrate the BovineMine datasets. Researchers can load their own lists of identifiers using the List tool (which accepts various types of identifiers or symbols) or the Genomic Regions tool (which accepts chromosome identifiers and coordinates), allowing them to connect their data to information in the BovineMine database. Results are provided as tables that can be filtered, reorganized and downloaded in various formats. We have created pre-defined query templates that provide starting points for data exploration. A strong point of BovineMine is the ability to mine tissue specific gene expression levels together with genomic variation data, a function not previously available to bovine researchers. BovineMine enables researchers to leverage the curated gene pathways of model organisms (e.g. human, mouse and rat) based on orthology, and is especially useful for GO and pathway analyses in conjunction with GWAS and QTL studies.

C22: Computer Demo 4

Desktop BioLegato Applications for Easy NCBI Keyword Queries and BLAST Searches

Brian Fristensky and Graham Alvare, University of Manitoba, Winnipeg, MB, Canada

BioLegato is a programmable graphic interface, leveraging object-oriented concepts to rapidly create desktop tools that are intuitive because they resemble real-world objects. We now present an integrated set of BioLegato tools for searching NCBI databases that overcome many of the limitations inherent in web-based interfaces. BioLegato promotes exploration of data through point-and-click pipelining. Output from each step, such as sequences or search results, appears in a new BioLegato tool, with methods appropriate for the output. To illustrate, complex NCBI keyword queries are done using a form-based query builder, and query results returned in a BioLegato spreadsheet making it easier to sift through large numbers of hits and to refine the query. Sequences that match search criteria are retrieved directly from the spreadsheet to BioLegato sequence tools. BLAST searches can be launched from BioLegato sequence tools, and results returned to BioLegato spreadsheets for direct retrieval of sequences. We also include automated tools for downloading and configuring local copies of NCBI BLAST databases. Even on modest computer systems, local BLAST searches return results quickly by avoiding the waiting time in NCBI BLAST queues. BioLegato is part of the Open-Source BIRCH Bioinformatics system, found at http://home.cc.umanitoba.ca/~psgendb.

C23: Computer Demo 4

The Vigna Genome Server, 'VigGS': A Genomic Knowledge Base of the Genus Vigna

Hiroaki Sakai, Ken Naito, Yu Takahashi, Takeshi Itoh and Norihiko Tomooka, National Institute of Agrobiological Sciences, Tsukuba, Japan

The genus *Vigna* includes legume crops such as cowpea, mungbean, and azuki bean as well as over 100 wild species. A number of the wild species are highly tolerant to severe environmental conditions including high salinity, acid or alkaline soil; drought; flooding; and pests and diseases, making it a good target for investigation of genetic diversity in adaptation to stressful environments. However, a lack of genomic information has hindered such research in this genus. Here, we present a genome database of the genus *Vigna*, Vigna Genome Server '*Vig*GS', based on the recently sequenced azuki bean genome, which incorporates annotated exon-intron structures, along with evidence for transcripts and proteins, visualized in GBrowse. *Vig*GS also facilitates user construction of multiple alignments between azuki bean genes and those of six related dicots. In addition, the database displays sequence polymorphisms between azuki bean and its wild relatives and enables users to design primer sequences targeting any variant site. To incorporate up-to-date genomic information, *Vig*GS automatically receives newly deposited mRNA sequences of preset species from the public database once a week. Users can refer to not only gene structures mapped on the azuki bean genome on GBrowse but also relevant literature of the genes. Currently genome sequencing of 13 wild *Vigna* species is in process. Comparative

protein mapping data based on each species genome and genome alignment for each pair of closely related species will be implemented in *Vig*GS, which will shed light on the genetic consequences of the adaptive evolution of the *Vigna* species.

C24: Computer Demo 4

Using Wheat BLAST Database to Search for Mutations and Expression

Hans Vasquez-Gross[1], Ksenia V Krasileva[2], Tyson R. Howell[1], Paul C. Bailey[3], Sarah Ayling[3], Cristobal Uauy[4], Stephen Pearce[1] and Jorge Dubcovsky[5], (1)University of California Davis, Davis, CA, (2)The Genome Analysis Centre, The Sainsbury Laboratory, Norwich, United Kingdom, (3)The Genome Analysis Centre, Norwich, United Kingdom of Great Britain and Northern Ireland, (4)John Innes Centre, Norwich, England, (5)University of California, Davis, CA

The exome sequencing of an EMS mutagenized tetraploid wheat population comprised of 1,500 mutant lines enables a valuable *in silico* targeting of induced local legions in genomes (TILLING) reverse genetics tool for wheat researchers and breeders. The database includes 3.7 million mutations in the coding regions of ~82,000 genes, resulting in a high probability of deleterious mutations or truncations in the two homoeologs of each gene. This computer demonstration will show how to use this new tool to rapidly identify wheat genes of interest, detect mutations, and request corresponding seed stocks. During the session, we will cover performing a BLAST search for gene of interest, finding potential high quality hits, and visualizing the results on a reference genome browser (JBrowse). For each annotated gene, we have identified silent synonymous and intergenic changes as well as non-synonymous and truncation mutations (premature stop codons and splice site mutations). This resource is available upon request from the UCDavis - Dubcovsky Lab at: http://dubcovskylab.ucdavis.edu/wheat-tilling

We have also developed a complementary tool that allows the user to visualize the expression profiles of the different homoeologs of the wheat gene of interest in published RNAseq experiments. Using BLAST users can identify a gene of interest, select one or more resulting genes and compare expression profiles through multiple different tissue types. For more information visit: http://dubcovskylab.ucdavis.edu/wheat-expression-database

The combination of expression and mutant databases has the potential to accelerate functional studies of most of the wheat genes.

C25: Computer Demo 4

GenSAS v4.0: A Web-Based Platform for Structural and Functional Genome Annotation and Curation

Jodi L. Humann[1], Stephen P. Ficklin[1], Taein Lee[1], Chun-Huai Cheng[1], Sook Jung[1], Jill Wegrzyn[2], David Neale[3] and Dorrie Main[1], (1)Washington State University, Pullman, WA, (2)Department of Ecology and Evolutionary Biology, University of Connecticut, Storrs, CT, (3)Dept. Plant Sciences University of California Davis, Davis, CA

The Genome Sequence Annotation Server (GenSAS) is a web-based annotation platform (gensas.bioinfo.wsu.edu) that packages multiple command line annotation tools under one, easy-to-use interface. GenSAS walks researchers through each step of the annotation process and provides instructions and hints as each step progresses. These steps include: uploading of genomic sequence; uploading of supportive repeat library files, species-specific (or species-related) protein and transcript files; execution of repeat masking tools (including *de novo* repeat identification); execution of gene prediction tools; consensus gene prediction; functional annotation; manual structural annotation editing; and assistance with organization of files for final publication. GenSAS is integrated with the popular JBrowse genome browser for feature visualization, and manual curation of structure and function is provided by an integrated Web Apollo interface. GenSAS was designed with researchers as the target user group and based on feedback, GenSAS is an easy-to-use, customizable, online DNA annotation platform that allows users with little computer science knowledge to create a custom DNA annotation for their sequence(s) of interest.

C26: Computer Demo 4

Learn to Use Integrated Genome Browser to Explore, Analyze and Share Data for your Newly Sequenced Genome - an Example from Blueberry

Ann Loraine, David Norris, John Eckstein, Tarun Mall, Nowlan Freese and Mason Meyer, University of North Carolina Charlotte, Kannapolis, NC

Integrated Genome Browser is a high-performance, desktop genome browser freely available for download from BioViz.org. Easy-to-use installers are available for every platform. Originally developed (at Affymetrix) to display genome tiling array data for the ENCODE project, IGB is now a versatile tool for visualizing RNA-Seq, ChIP-Seq, Bisulfite-Seq, and other *Seq datasets. IGB can also be used to run BLAST searches, search InterPro, view data from Galaxy, design PCR primers, map restriction sites, show plus and minus strand features in the same or different tracks, view paired-end read data, copy and paste genomic sequence, and much more. IGB is also extensible. Using a new services-based API, developers can add new functions and features as IGB Apps. For communities working with a new genome assemblies, one the most useful features is IGB Quickload. IGB Quickload is a simple system of files and folders for sharing genome data on the Web. Using IGB Quickload, research groups can easily set up their own genome browser system. In this demo, I'll describe this simple system and how we used it to make an IGB Quickload site for blueberry.

C27: Computer Demo 4

ProtAnnot: Visualizing Protein Function and Effects of Alternative Transcription

Nowlan Freese, Tarun Mall, John Eckstein, David Norris and Ann Loraine, University of North Carolina Charlotte, Kannapolis, NC

ProtAnnot is a new optional plug-in App for the Integrated Genome Browser (IGB) that displays protein annotations in the context of genomic sequence. To run ProtAnnot, users select one or more gene models in IGB and choose Tools > Open ProtAnnot. This opens ProtAnnot in a separate window, showing a new view of the selected gene models. In ProtAnnot, exons are color-coded by the frame of translation, making it easy to notice when splicing differences cause frameshifts in the gene models. Below the gene models is an exon summary graphic, where differences in block heights indicate difference regions, segments of genomic sequence that are included in one gene model and not the other. To identify regions within gene models likely to be important for gene function, users can search the InterPro database from within ProtAnnot. To run a search, users select “Run InterProScan”. This opens a menu with options to select and then search any of the sixteen profile databases

available from InterPro. Once the search completes, protein annotations are mapped onto the genomic sequence alongside gene models. Selecting protein annotations displays more information, including name, description and links to more information. Viewing and interacting with protein annotations displayed in ProtAnnot helps with identifying conserved domains within a protein and overall gene function. An easy-to-use Image Export tool generates high-resolution ProtAnnot images for use in printed publications or slide presentations. ProtAnnot and IGB are freely available from http://www.bioviz.org.