C01: Digital Tools and Resources Session 1

Bionano Data Analysis on the Command Line and Genome Assembly Quality Assessment using a GUI

Yuxuan Yuan[1], Philipp E. Bayer[2], Jenny HueyTyng Lee[3], Armin Scheben[1,2], Chon-Kit Kenneth Chan[2] and David Edwards[2], (1)University of Western Australia, Perth, Western Australia, Australia, (2)University of Western Australia, Perth, Australia, (3)Justus Liebig University, Giessen, Germany

Advances in optical mapping are helping to produce highly contiguous genome assemblies. However, the tools for the analysis of optical mapping data are limited in their ease of use and computing platform restrictions. In this workshop, I will present and demonstrate runBNG [1], an open-source software package which wraps Bionano genomic analysis tools within a single script that can be run on the command line. runBNG can complete analyses including quality control of single molecule maps, optical map *de novo* assembly, comparison between different optical maps, super-scaffolding, and structural variation detection. Additionally, I will also presnet BioNanoAnalyst [2] to assess genome assembly quality through a graphic user interface (GUI) using optical map data.

1. Yuan, Y., et al., *runBNG: a software package for BioNano genomic analysis on the command line.* Bioinformatics, 2017. **33**(19): p. 3107-3109.

2. Yuan, Y., et al., *BioNanoAnalyst: a visualisation tool to assess genome assembly quality using BioNano data.* BMC Bioinformatics, 2017. **18**(1): p. 323.

C02: Digital Tools and Resources Session 1

RepetDB: A Resource for Unified Transposable Element References with Classification

Joelle Amselem, Guillaume Cornut, Nathalie Choisne, Michael Alaux, Françoise Alfama-Depauw, Véronique Jamilloux, Florian Maumus, Thomas Letellier, Isabelle Luyten, Cyril Pommier, Anne-Francoise Adam-Blondon and Hadi Quesneville, URGI, INRA, Université Paris-Saclay, Versailles, France

The ability of Transposable Elements (TEs) to move and replicate throughout the genomes makes them perhaps the most important contributors to genome evolution. Their detection and annotation are considered essential, and must be undertaken in the frame of any genome sequencing project.

Only a fully automated TE *de novo* annotation process is able to face the sequence deluge rapidly increasing with the improvement of high-throughput sequencing technologies. This process generally relies on pipelines to *de novo* detect, build and classify consensuses using similarity with sequences already identified and/or according to their structure. However, as any automated procedure, TE identification and classification are intrinsically an error prone process. Consequently, there was a crucial need to provide TE consensuses with evidences able to justify their classification, and this on the up to thousands TE consensuses for one genome that such an approach may provide.

Few TE databases already exist focusing generally on different TE biology views (Species, TE families, TE masking, …) but biological information on the sequences remains globally poor.

Here, we will present RepetDB developed in the frame of GnpIS, a genetic and genomic Information System. RepetDB was designed to store and retrieve detected, classified and annotated TEs in a standardized manner. RepetDB is an implementation with extensions of InterMine, an open-source data warehouse framework used here to store, search, browse, analyze and compare all the data recorded for each TE reference sequence. InterMine can display diverse information for each sequence and allows simple to very complex queries. Finally, TE data are displayed via a worldwide data discovery system.

RepetDB is designed to be a TE knowledge base populated with full *de novo* TE annotations of complete (or near-complete) genome sequences. Indeed, the description and classification of TEs facilitates the exploration of specific TE families, superfamilies or orders across a large range of species. It also makes possible cross-species searches and comparisons of TE family content between genomes

C03: Digital Tools and Resources Session 1

DiNAR: Revealing Hidden Patterns of Plant Signalling Dynamics using Differential Network Analysis in R

Kristina Gruden, Maja Zagorscak, Andrej Blejec, Ziva Ramsak, Marko Petek and Tjasa Stare, National Institute of Biology, Department of Biotechnology and Systems Biology, Ljubljana, Slovenia

Progress in high-throughput molecular methods accompanied by more complex experimental designs demands novel data visualization solutions. To specifically answer the question which parts of the specific biological system are responding in particular perturbation, integrative approach in which experimental data are superimposed on a prior knowledge network is shown to be advantageous.

We have developed DiNAR, Differential Network Analysis in R, a user-friendly application with dynamic visualization that integrates multiple condition high-throughput data and extensive biological prior knowledge. Implemented differential network approach and embedded network analysis allow users to analyse condition-specific responses in the context of topology of interest (e.g. immune signalling network) and extract knowledge concerning patterns of signalling dynamics (i.e. rewiring in network structure between two or more biological conditions). We validated the usability of software on the *Arabidopsis thaliana* and *Solanum tuberosum* data sets and demonstrated how DiNAR facilitates detection of network-rewiring events, gene prioritization for future experimental design and allows capturing dynamics of complex biological system.

The fully cross-platform Shiny App is hosted and freely available at https://nib-si.shinyapps.io/DiNAR while the most recent version of the source code is available at https://github.com/NIB-SI/DiNAR/.

C04: Digital Tools and Resources Session 1

NCGAS Makes Robust Transcriptome Assembly even easier with Added Features to an Accessible *de novo* Transcriptome Assembly Workflow

Sheri Sanders[1], Bhavya Papudeshi[2], Carrie Ganote[2] and Tom Doak[2], (1)National Center for Genome Analysis Support, Pervasive Technology Institute, Bloomington, IN, (2)Indiana University, Bloomington, IN

The National Center for Genome Analysis Support (NCGAS) assists research groups with de novo transcriptome assembly. Following best practice for combined de novo transcriptome assemblies can put a technical burden on genomic researchers who may not be fully

computationally trained on efficient use of HPC clusters or the variety of available software packages. NCGAS has created a workflow template to move RNAseq data through 19 parallelized assemblies using four software packages (Trinity, SOAP-denovo, transABySS, and Velvet Oases) and multiple kmers. The transcripts are then combined and filtered using EviGenes to output putative transcripts and alternative forms in a replicable manner. The process is semi-automated but flexible enough to allow researchers to adjust parameters if they desire. This workflow provides a low bar for entry into robust transcriptome assembly that follows best practices, while also providing a replicable means of filtering large numbers of transcripts into a draft version of a transcriptome. We will highlight the main work flow in this demo but will concentrate on the additional features added to the workflow in the last year, including annotation via Trinotate, differential expression handling, and the automated creation of table of assembly metrics via BUSCO and Quast for each sub-assembly. As this workflow has now been adopted by several groups, we will also discuss available training and current implementations of the tool.

C05: Digital Tools and Resources Session 1

Efficient Ortholog Identification and Interactive Web-Based Whole Genome Alignment with NGSEP

Daniel Tello, Camilo Escobar-Velásquez, Mario Linares-Vásquez and **Jorge Duitama**, Universidad de los Andes, Bogotá, Colombia

Recent developments on long read high throughput sequencing technologies enabled high quality genome assemblies for an unprecedented number of species. These new genomes represent a unique data resource to elucidate complex patterns of evolutionary events through comparative genomics. A basic operation in comparative genomics is the alignment of the entire genomes of two or more species to identify at the same time conserved elements and patterns of genomic variability. Although genome alignment is a classical problem in bioinformatics, recent developments on data structures, algorithms and technologies create opportunities to develop a new generation of bioinformatic tools for this problem. Here we present a new software solution for alignment of complete genomes through the efficient identification of synteny blocks built from large chains of orthologous genes. Given two genomes, our solution builds first FM-indexes from their corresponding proteomes to enable a highly efficient identification of paralogs and orthologs within and between genomes. The classical longest common subsequence dynamic programming algorithm was reimplemented to identify sequences of collinear orthologs, building synteny blocks. In contrast with classical tools that build alignments based on raw sequence similarities, ortholog chains enable alignments between chromosomes of large genomes within minutes of computation, requiring fewer processors and less than 8Gb of RAM. We developed this solution as part of the NGSEP software tool to take advantage of the usability features and the common structures already implemented in this solution. The NGSEP genomes aligner was benchmarked against some of the most common tools for comparative genomics, namely Mauve, MUMmer, Satsuma and Symap, and also to ProteinOrtho for ortholog identification. Our solution takes significantly less time and resources for the processing of large genomes. Comparing the annotated genomes of common bean (*Phaseolus vulgaris*) and soy bean (*Glycine max*) with our genomes aligner we were able to reconstruct the whole genome duplication event that relates these two species and identify the correspondence between the chromosomes of the two genomes. Using state-of-the-art data visualization technologies implemented in the javascript D3 package we provide novel interactive views of the alignments provided by our software. Our genomes aligner is already available as open source software as part of the NGSEP distribution. We expect that this development represents a significant contribution to the field of comparative genomics facilitating further discoveries in evolution, functional genomics and related fields.

C06: Digital Tools and Resources Session 1

GEMmaker, a Nextflow Workflow for Large Scale RNA-Seq Processing

John A. Hadish[1,2], Tyler D. Biggs[1], Connor Wytko[3], Ben Shealy[4], Melissa C. Smith[4], F. Alex Feltus[5] and Stephen P. Ficklin[1,2], (1)Dept of Horticulture, Washington State University, Pullman, WA, (2)Molecular Plant Sciences Program, Washington State University, Pullman, WA, (3)Dept of Electrical Engineering & Computer Science, Washington State University, Pullman, WA, (4)Dept of Electrical and Computer Engineering, Clemson University, Clemson, SC, (5)Dept of Genetics and Biochemistry, Clemson University, Clemson, SC

Processing thousands of RNA-seq data sets can be cumbersome due to data storage challenges, execution on high performance computing (HPC), installation of a multitude of software and organization of data through the workflow. GEMmaker is an open-source, freely available Nextflow workflow that simplifies gene expression-level quantification and construction of gene expression matrices (GEMs) for Illumina RNA-seq data intended for differential gene expression (DGE) analysis or construction of gene co-expression networks. GEMmaker can automatically download a set of RNA-seq data from NCBI's Sequence Read Archive (SRA), or use local data files. The workflow supports several popular tools including trimmomatic, fastqp, FastQC, hisat2, Salmon, Kallisto, and MultiQC, samtools and includes custom scripts for data preparation. A Docker image accompanies GEMmaker, easing the burden for software installation, on HPC systems. In addition, GEMmaker generates metadata files mapping common annotation terms to controlled vocabulary.

C07: Digital Tools and Resources Session 2

CerealsDB Version 4.0 - A Review of Tools and Data for Wheat Breeders and Research Scientists

Paul A. Wilkinson, Mark O. Winfield, Gary L. Barker, Alexandra M. Przewieslik-Allen, **Amanda J Burridge**, Jane A. Coghill, Christy Waterfall, Alexander Coulton, Daniel Shaw, Lucy Hyde and Keith J. Edwards, University of Bristol, Bristol, United Kingdom

CerealsDB is an open access website that hosts information on wheat SNPs considered useful for both plant breeders and research scientists. The database currently contains in excess of a million putative varietal SNPs, of which 820,000 have been experimentally validated. Genotype data may be accessed for a range of genotyping platforms including KASPar, TaqMan, iSelect, Axiom and Targeted GbyS. The SNP data is supported by probe sequences, marker location and functionality where known.

Originally created in 2003 to host wheat EST sequences generated at the University of Bristol, Cereals DB has since expanded to include sequence and SNP databases plus numerous webtools. In November 2018, the website will undergo a major update to version 4.0 which will see the number of genotyped accessions increased to 7,500.

CerealsDB has maintained a policy of making its data freely available to the public without restrictions or intellectual property rights. The site does not require registration, access is not password protected and data is provided to the wheat breeding and research communities as quickly as possible to maximise the potential utility of the data. Users will find the website simple to use and the data to be in a format that is easy to understand.
While CerealsDB is used globally, this is the first time that it has been demonstrated at PAG. The session will include the ease of data access and an overview of the tools available to make the most of the data available.
http://www.cerealsdb.uk.net

C08: Digital Tools and Resources Session 2

GenSAS v6.0: A Web-Based Platform for Structural and Functional Annotation of Model and Non-Model Organisms

Jodi L. Humann[1], Taein Lee[2], Stephen P. Ficklin[3], Chun-Huai Cheng[2], Heidi Hough[1], Sook Jung[1], Jill L. Wegrzyn[4], David B. Neale[5] and Dorrie Main[1], (1)Washington State University, Pullman, WA, (2)Washington State University, Pullman, Pullman, WA, (3)Dept of Horticulture, Washington State University, Pullman, WA, (4)Department of Ecology and Evolutionary Biology, University of Connecticut, Storrs, CT, (5)Department of Plant Sciences, University of California, Davis, Davis, CA

The Genome Sequence Annotation Server v6.0 (GenSAS, www.gensas.org) is a web-based annotation platform that combines several common annotation tools into one easy-to-use, integrated resource. The annotation process is carried out in user-friendly interfaces with embedded instructions and only requires a user account and internet access. GenSAS has annotation tools for eukaryotes and prokaryotes and supports model and non-model organisms. GenSAS allows users to upload Illumina RNA-Seq reads (or specify datasets from the NCBI SRA database), align the reads to the genome using HISAT2 or TopHat2, and use the aligned data to train the gene model prediction programs AUGUSTUS and BRAKER2, which allows for more accurate gene models for eukaryotic genomes, especially non-model organisms. JBrowse and Apollo are integrated into GenSAS allowing for structural annotation results to be easily viewed and manual annotation curation to be performed. Users can share GenSAS projects with other users and manual curation can easily be a collaborative project. GenSAS also has a functional annotation step to assign protein functions and identify functional domains to the official gene set. After the annotation process is complete, the final step of the GenSAS pipeline is to generate the required files for publication and allows the user to run BUSCO on the predicted proteins to assess he completeness of the annotation.

C09: Digital Tools and Resources Session 2

Shinotate: An R-Based Shiny Server for Annotation and Analysis of RNA-Seq Transcriptome Assemblies

Ido Bar, Grifith University, Brisbane, QLD, Australia

Assembly of transcriptome data in non-model species has become common practice in the last decade thanks to the advent of high-throughput RNA-sequencing platforms and accompanying bioinformatics tools. **Trinity** is one of the most commonly used tools for transcriptome assembly from Illumina RNA-Seq data and its accompanying functional annotation framework, **Trinotate**, offers a pipeline for running the various annotation tools and consolidating the results into a single database. Trinotate also includes a web-based graphical user interface for querying the annotations and provide basic visualisation, but its Perl implementation makes it difficult to customise and deploy. **Shinotate** was developed to provide a modern graphical interface for the analysis of transcriptome annotations, utilising the Trinotate annotation framework to deliver summarised results and insights to users of all skill levels. Shinotate is written in R and uses the `tidyverse` approach to summarise and visualise the data stored in Trinotate and thus can be easily adapted to accommodate custom annotation tables. It serves interactive annotation tables and plots, with search, selection and data export functions.

C10: Digital Tools and Resources Session 2

Exploring Your Data in Ensembl

Erin Haskell, EMBL-EBI, Hinxton, United Kingdom

The Ensembl (www.ensembl.org) and Ensembl Genomes (www.ensemblgenomes.org) projects provide accessible genomic data across the tree of life including vertebrates, bacteria, fungi, metazoa, protists, and plants. We provide annotated genome assemblies, and draw in a range of data from other external databases (e.g. Gene Ontology, Expression Atlas and Reactome) to create a unified database for genome and gene information. In this session we will show you how you can upload and configure your own experimental data in Ensembl, to aid with analysis. We support upload of a range of file types, including BAM, BED, BigWig and many others. We will demonstrate direct file upload and introduce resources for uploading files by URL, which is necessary for larger file sizes. We will also cover how to search the Track Hub Registry for additional datasets available for display in the Ensembl browsers. We will provide an annotated document with step-by-step details on the processes and sample files that you can access online, and which will remain permanently available after the course.

C11: Digital Tools and Resources Session 2

TreeGenes and CartograTree: Tools for Forest Tree Genomics

Emily Grau[1], Sean Buehler[2], Nic Herndon[2], Peter Richter[2], Taylor Falk[2], Ronald Santos[1], Margaret Staton[3], Stephen P. Ficklin[4] and Jill L. Wegrzyn[2], (1)University of Connecticut, Storrs, CT, (2)Department of Ecology and Evolutionary Biology, University of Connecticut, Storrs, CT, (3)University of Tennessee, Knoxville, TN, (4)Dept of Horticulture, Washington State University, Pullman, WA

TreeGenes is a web-based information resource designed to serve the diverse needs of the forest tree genomics research community by uniting information resources with tools for visualization and analysis. TreeGenes is built with Tripal, an open source tool to create and manage genomic database websites. Tripal allows developers the flexibility to create, share, and reuse tools with other data repositories.
TreeGenes hosts a range of information resources including genetic, genomic, and phenotypic data along with literature and community resources for over 1700 species. The Tripal Galaxy module allows users to execute analytical workflows with next generation sequencing datasets on high performance computing resources with the click of a button. Other tools include the Tripal Plant PopGen Submission (TPPS)

module for collection of population genetic data and metadata, TSeq for rapid sequence similarity search, OrthoQuery for visualizing orthologous gene sets in a phylogenetic context, and CartograTree for landscape and association analysis.
CartograTree is a map-based framework that provides an efficient tool for researchers to display, select, and analyze model and non-model trees in conjunction with their associated genotypic and phenotypic metrics. CartograTree allows users to query, filter, and visualize data sourced from TreeGenes as well as external databases including Dryad and TreeSnap. The integration of environmental layers with genomic data associated with these georeferenced trees allows for analyses such as association mapping and landscape genomics. The Tripal Galaxy tool is integrated into CartograTree, allowing users to seamlessly submit datasets for analysis.

C12: Digital Tools and Resources Session 2

MtSSPdb: The *Medicago truncatula* Small Signaling Peptide Database

Clarissa Boschiero[1], Xinbin Dai[1], Peter K. Lundquist[1], Sonali Roy[1], Thomas Christian de Bang[2], Shulan Zhang[1], Zhaohong Zhuang[1], Ivone Torres-Jerez[1], Michael K. Udvardi[1], Wolf R. Scheible[1] and **Patrick X. Zhao**[1], (1)Noble Research Institute, Ardmore, OK, (2)University of Copenhagen, Frederiksberg C, Denmark

Small Signaling Peptides (SSPs) are emerging as an important class of regulatory molecules in plants, playing important roles in plant developmental processes, nutrient acquisition, and stress responses. *Medicago truncatula* is a premier model legume species, which is closely related to the most important forage species, alfalfa and to food legumes such as soybean and common bean.
We developed MtSSPdb database (http://mtsspdb.noble.org/) which hosts comprehensive genomics and transcriptomics data, with a focus on SSP-encoding genes, in *M. truncatula*. The MtSSPdb consists of three important components: (1) re-annotated gene models and annotations of *M. truncatula* genes, including our newly reported 4,425 putative and known SSP genes and related family information. It is worth mentioning that we also included a peptide library, currently consisting of 155 synthetic peptides representing 101 SSP candidates. These peptides were tested for 58 root-related traits on *M. truncatula*, *Arabidopsis thaliana* and *Panicum virgatum*. (2) SSP Gene Expression Atlas (SSP-GEA), which hosts 13 RNA-seq experiments including macronutrient deficiencies, nodule development, hormone, salt stress, symbiotic interactions, and plant tissues. (3) A series of analytic tools, such as differential gene expression, co-expression, GO enrichment, and KEGG enrichment pathway analysis, SSP prediction tool and BLAST.
The MtSSPdb is the first database in plants to integrate comprehensive SSP information and large-scale gene expression data, and to provide a detailed synthetic peptide library tested for multiple root-related traits. In near future, we aim to expand this database into other important plant models, including (but are not limited to) *A. thaliana* and *Brachypodium distachyon*.

C13: Digital Tools and Resources Session 3

An Atlas of *Sus scrofa* RNA Editome in Developmental Skeletal Muscle

Zhonglin Tang, Agricultural genomics Institute, CAAS, Shenzhen, China

A-to-I RNA editing is a widespreadpost-transcriptional event in mammals. However, both the variations and regulatory functions of A-to-I editing remain unknown in skeletal muscle. Here, for the first time, by integrating strand-specific RNA-seq, whole genome bisulfite sequencing and whole genome sequencing data, we have comprehensively profiled the A-to-I editome and analyzed its functional properties in developing skeletal muscles across 27 prenatal and postnatal stages in pig, an important farm animal and biomedical model. A total of 236,569 editing sites were identified and 84% of them were canonical A-to-I editing types. The majority of A-to-I editing sites were clusters and overlapped with the SINE/tRNA repetitive elements in non-coding regions. The RNA editing sites occurred more frequently in prenatal skeletal muscles. Both the editing level and frequency decreased during developmentand positively correlated with the expression of ADARs enzymes. The developmentally differential editing sites were significantly overrepresented in untranslated regions, functionally enriched in genes associated with muscle development, highly correlated with expression of their host mRNAs, and potentially influenced the gain/loss of miRNA binding to mRNA. Moreover, we observed the positive relationship between DNA methylation and RNA editing during skeletal muscle development. Finally, we developed a user-friendly database which we named DREP to visualize the Sus scrofa RNA editome. Taken together, our study not only provides a valuable resource for studying RNA editing in mammals, but also furthers our understanding of RNA editing regulation in skeletal muscle and human muscle-related diseases.

C14: Digital Tools and Resources Session 3

SHiNeMaS : A Database and its Web Interfaces dedicated to Seed Lots History, Phenotyping and Cultural Practices

Yannick De Oliveira, Laura Burlot, Isabelle Goldringer, Darkawi Madi, Pierre Rivière, Delphine Steinbach, Gaelle Van Frank and Mathieu Thomas, INRA, Gif sur Yvette, France

In 2005, a collaboration started between the French National Institute for Agricultural Research (INRA) and the french farmers' seed network Réseau Semences Paysannes (RSP) on bread wheat.
The aims are: (1) to study on-farm management of crop diversity; (2) to develop population-varieties adapted to organic and low inputs agricultures in the context of a participatory plant breeding program involving farmers, farmers' organisation faciltators and researchers.
In this project, researchers and farmers' organizations needed to map the history and life cycle of the population-varieties using the network formalism. All this information needed to be centralized and stored in a database.
Thus, we developed SHiNeMaS (Seeds History and Network Management System) a database with its web interface, dedicated to the management of the history of seed lots and related data like phenotyping, environment and cultural practices. SHiNeMaS is freely available and distributed under GNU Affero GPL Licence : https://sourcesup.renater.fr/projects/shinemas/

C15: Digital Tools and Resources Session 3

psRNATarget V2: A High-Performance Plant Small RNA Target Analysis Server

Xinbin Dai, Zhaohong Zhuang and Patrick X. Zhao, Noble Research Institute, Ardmore, OK

The psRNATarget (http://plantgrn.noble.org/psRNATarget/) was developed to identify the targets of Plant regulatory small RNAs (sRNAs), which include most microRNAs (miRNAs) and a subset of small interfering RNAs (siRNAs), such as the phased siRNAs (phasiRNAs) that share same mechanisms in post-translational gene silencing and translational inhibition. It predicts the targets of a given sRNA by (i) analyzing the

complementary matching between the sRNA sequence and target mRNA sequence using a predefined scoring schema, and (ii) by evaluating target site accessibility.
In the psRNATarget V2, we developed a new scoring schema that is capable of discovering miRNA-mRNA interactions at higher 'recall rates' without significantly increasing total prediction output, thus significantly enhanced its analytical performance. We also provided interfaces to enable users to further customize the scoring schema to search both canonical and non-canonical targets. The psRNATarget V2 enables transmitting and analyzing 'big' data through (a) the implementation of multi-threading chunked file uploading, which can be paused and resumed, using HTML5 APIs and (b) the allocation of significantly more computing nodes to its back-end Linux cluster. The psRNATarget V2 has clear, compelling and user-friendly interfaces that enhance user experiences and present data clearly and concisely.

C16: Digital Tools and Resources Session 3

Integrated Genomic Selection Galaxy Analysis Pipeline and Workflows

Star Yanxin Gao[1], Angel Villahoz-Baleta[1], Venice Margarette Juanillas[2], Kelly Robbins[3], Umesh Rosyara[4], Victor J. Ulat[4], Paulino Perez-Rodriguez[5], Jose Crossa[4], Xuecai Zhang[4], Juan D. Arbelaez[2], Fernando Toledo[4], Juan Burgueño[4], Alexis Dereeper[6], Sivasubramani Selvanayagam[7], Isaak Y. Tecle[8], Dmytro Chebotarov[2], Manish Roorkiwal[7], Elizabeth Jones[1], Yaw A. Nti-Addae[1], Abhishek Rathore[7], Michael Olsen[9], Kate A. Dreher[4] and Ramil Mauleon[2], (1)Genomic Open-Source Breeding Informatics Initiative Project, Ithaca, NY, (2)International Rice Research Institute, Los Baños, Philippines, (3)Cornell University, Ithaca, NY, (4)International Maize and Wheat Improvement Center (CIMMYT), Texcoco, Mexico, (5)Colegio de Postgraduado, Texcoco, Mexico, (6)Excellence in Breeding, F-34394 Montpellier, France, (7)ICRISAT, Hyderabad, India, (8)Boyce Thompson Institute, Ithaca, NY, (9)CIMMYT, Nairobi, Kenya

As sequencing and genotyping costs decrease, high throughput genotyping is becoming more feasible for routine plant breeding application. However, to benefit from this plant breeders need intuitive and scalable genomic analysis workflow and selection decision tools.
The Genomic Open-Source Breeding Bioinformatics Initiative (GOBII, http://gobiiproject.org), in collaboration with breeders and statisticians, has developed genomic prediction tools in the open-source Galaxy platform for crops such as maize, chickpea and wheat, as single-click workflows. We will give a live demo of key component tools and workflows, including importing source input files, converting file formats and types, imputing missing genotypic data, matrix based-diversity analysis, GS model selection and training, GS k-fold cross validation, predicting performance of untested populations, and customer-defined within and cross group prediction.
This first integrated genomic selection analysis is free to the public at http://galaxydemo.excellenceinbreeding.org/. A tutorial in progress can be found at: http://cbsugobii05.tc.cornell.edu:6084/x/zwHIAQ. This analysis pipeline will enable genomic selection to be used routinely in crop breeding programs.

C17: Digital Tools and Resources Session 3

A Tissue and Host-Specific Expression Atlas for *Citrus sinensis* and *Diaphorina citri* for Exploring the Citrusgreening Disease Complex

Mirella Flores-Gonzalez[1], Noe Fernandez-Pozo[1], Marina Mann[2], Prashant S Hosmani[1], Susan J. Brown[3], Michelle Cilia[2], Lukas A. Mueller[1] and Surya Saha[1], (1)Boyce Thompson Institute, Ithaca, NY, (2)Cornell University, Ithaca, NY, (3)Division of Biology - Kansas State University, Manhattan, KS

Citrusgreening or Huanglongbing (HLB) is a tritrophic disease complex involving citrus host, the Asian citrus psyllid (ACP, *Diaphorina citri*) vector and a phloem restricted bacterial pathogen *Ca.* Liberibacter asiaticus (CLas). HLB is the most devastating of all citrus diseases, and there is currently no adequate control strategy. Gene transcription and translation patterns are critical for understanding how the underlying genome sequence translates into specific phenotypes at key developmental and infection stages. We have created open-access, interactive and user friendly web tools: for Psyllid Expression Network (PEN) and Citrus Expression Network (CEN). PEN contains high-resolution proteomics and transcriptomics expression data from CLas infected and healthy individuals across multiple life stages, hosts and tissues. CEN contains public expression data from NCBI Sequence Read Archive for *C. sinensis* and *C. clementina*. It includes data from various life stages, tissues and infection states. The expression is quantified using transcripts per million reads (TPM) and only genes with more than 1 TPM in at least one tissue were retained. Heatmaps and scatter plots are also available for exploring transcription profiles. Users can filter expression data on subsets of hosts, treatments, organs and developmental stages of interest. PEN and CEN facilitate effective data analysis by enabling simultaneous visualization of correlated genes to develop novel hypothesis in addition to candidate gene identification. These tools are a useful resource not only for citrusgreening research but also other tritrophic disease systems. These resources are available at our open access and public web portal (https://citrusgreening.org).

C18: Digital Tools and Resources Session 3

GeneHummus: A Computational Pipeline to Define Gene Families and their Expression in Legumes and other Plant Species

Jose V. Die, Department of Genetics ETSIAM, University of Córdoba, Córdoba, Spain, Moamen M Elmassry, Department of Biological Sciences, Lubbock, TX, Kimberly H LeBlanc, National Institute for Drug Abuse, Bethesda, MD, Olaitan I Awe, University of Ibadan, Ibadan, Nigeria, Allissa Dillman, NCBI - NIH, Bethesda, MD and **Ben Busby**, NIH/NCBI, Bethesda, MD

During the last decade, plant biotechnological laboratories have sparked a monumental revolution with the rapid development of next sequencing technologies and the dramatic drop in sequencing costs. This momentous leap represents exciting opportunities for the plant research community, enabling genome-wide investigations into fundamental, evolutionary and physiological questions. Understanding which genes are activated in which conditions will help selective breeding programs around the world. Insights into the role of gene expression during particular conditions can be derived from classification of these genes into gene families and further comprehensive analysis, such as structure, variation and functional studies. The recent release of several plant genome sequences provides an excellent opportunity for genome-wide identification of gene families using bioinformatic tools. Manual identification of gene families is highly time-consuming and laborious, requiring an iterative process of manual and computational analysis to identify members of a given family, typically combining numerous BLAST searches and

manually cleaning data. Due to the increasing abundance of genome sequences and the agronomical interest in plant gene families, the field needs a clear, automated annotation tool. Here, we present the GeneHummus pipeline for the identification, characterization and expression analysis of plant gene families. As a case study we focused on the auxin receptor factors (ARF) gene family in *Cicer arietinum* (chickpea) and other legumes. Our pipeline was specific and sensitive, only extracting manually identified chickpea ARF proteins from RefSeq database and providing the conserved domain architectures based on SPARCLE for all legume members of this gene family. GeneHummus can be customized to be suitable for other agronomically important taxonomic families beyond legumes. We anticipate that our pipeline should be suitable for any plant gene family, vastly improving the speed and ease of genomic data processing.

C19: Digital Tools and Resources Session 3

Genomic Data Manager - a Scalable Open-Source Genomics Database for Breeding Applications

Yaw A. Nti-Addae, Elizabeth Jones and Star Yanxin Gao, Genomic Open-Source Breeding Informatics Initiative Project, Ithaca, NY

Technological advances have greatly increased the utilization of genomic information in breeding programs. However, the capacity to analyze genomic data in a meaningful way lags behind the ability to generate and organize such data. To help address these obstacles, we have developed the Genomic Data Manager (GDM) (http://gobiiproject.org/index.php/gdm-genotype-data-manager/), a searchable and scalable open-source database designed to enable breeding analyses such as genomic selection and marker assisted selection. GDM currently includes a data loading utility, data QC engine and an easy to use web interface for extracting data. To aid analyses and optimize new data incorporation, GDM also includes a comprehensive list of carefully curated web services (RESTful API) that make it possible to search for makers and germplasms across multiple genotyping projects. It also allows for storage of multiple versions of analyzed genotype datasets (e.g. unimputed, imputed, cleaned, phased, etc.) enabling breeding programs to build a genetic knowledge-base to support decision-making, and allow avoidance of unnecessary testing. GDM is actively maintained under an MIT license by the Genomic Open-source Breeding informatics initiative, and is freely available on Bitbucket to download.

C20: Digital Tools and Resources Session 4

CLfinder-OrthNet: Encoding Evolutionary History of a Gene Locus as Network Topology

Dong-Ha Oh and Maheshi Dassanayake, Louisiana State University, Baton Rouge, LA

The CLfinder-OrthNet pipeline (1) detects co-linearity among multiple closely-related genomes, (2) finds orthologous gene groups, and (3) encodes the evolutionary history of each ortholog group into an Ortholog Network (OrthNet). OrthNets connect orthologs with edges representing either the presence or absence of co-linearity between them. Each OrthNet encodes in its network topology the evolutionary history of an orthologous gene group, including different modes of gene duplication, deletion, transposition, and combinations of them, occurred in a lineage or multiple lineages. Orthologous gene groups with the same evolutionary history can be retrieved by searching OrthNets with a network topology query.

As a proof-of-concept, we applied CLfinder-OrthNet to characterize gene transposition-duplication (*tr-d*) events among six Brassicaceae genomes, including those of *Arabidopsis thaliana* and two extremophytes, *Eutrema salsugineum* and *Schrenkiella parvula*. We identified subsets of lineage-specific *tr-d* events with signatures of selective retention and sub-functionalization in all six genomes. These included lineage-specific *tr-d* of genes that may be critical for the local adaptation of extremophytes, such as orthologs of *SALT TOLERANCE 32* and *ZINC TRANSPORTER 3*.

CLfinder-OrthNet offers a flexible toolset for systematic comparative studies of closely-related genomes. Beside the detection of all orthologs showing the same evolutionary history, the application includes but not limited to: (1) improving orthology inference assisted by co-linearity, (2) identification of gene duplication or transposition events co-occurring with certain phenotypic traits among closely-related genomes, and (3) detection of truncated, split, and chimeric gene models based on co-linearity.

CLfinder-OrthNet is available at https://github.com/ohdongha/CL_finder

C21: Digital Tools and Resources Session 4

Irisas: A Framework for Genome Wide Association with Structure Variants in *Arabidopsis thaliana* and *Drosophila melanogaster*

Xiangchao Gan, Max Planck Institute for Plant Breeding Research, Cologne, Germany

Short insertions, deletions (INDELs) and larger structural variants have been increasingly employed in genetic association studies, but few improvements over SNP-based association have been reported. In order to understand why this might be the case, we analysed two publicly available datasets and observed that 63% of INDELs called in *A. thaliana* and 64% in *D. melanogaster* populations are misrepresented as multiple alleles with different functional annotations, i.e. where the same underlying variant is represented by inconsistent alignments leading to different variant calls. To address this issue, we have developed the software Irisas to reclassify and re-annotate these variants, which we then used for single-locus tests of association. We also integrated them to predict the functional impact of SNPs, INDELs, and structural variants for burden testing. Using both approaches, we re-analysed the genetic architecture of complex traits in *A. thaliana* and *D. melanogaster*. Heritability analysis using SNPs alone explained on average 27% and 19% of phenotypic variance for *A. thaliana* and *D. melanogaster* respectively. Our method explained an additional 11% and 3%, respectively. We also identified novel trait loci that previous SNP-based association studies failed to map, and which contain established candidate genes. Our study shows the value of the association test with INDELs and integrating multiple types of variants in association studies in plants and animals.

C22: Digital Tools and Resources Session 4

Compute Resources Available to the Research Community for Microbiome Analysis

Bhavya Papudeshi[1], Sheri Sanders[2], Carrie Ganote[1] and Tom Doak[1], (1)Indiana University, Bloomington, IN, (2)National Center for Genome Analysis Support, Pervasive Technology Institute, Bloomington, IN

The National Center for Genome Analysis Support (NCGAS) is an NSF-funded center tasked with assisting biologists in getting access to computational resources they need in order to analyze genomic data. To support microbiome analysis, NCGAS provides preconfigured virtual

machines (VM) to identify taxa in 16S amplicon sequencing, and to identify both taxa and functions from whole genome metagenomes. Additionally, a pipeline to reconstruct genomes from metagenomes, to examine the role of specific microbes in a community, is available as a preconfigured VM hosting Anvi'o (https://ncgas.org/Blog_Posts/Running%20Anvio%20on%20Jetstream.php). Jetstream, a cloud computing resource, is both easy to use and flattens the learning curve for using the Linux operating system and for installing bioinformatics software. Jetstream provides an environment for both prototyping and publishing tailored workflows. Through an NCGAS allocation, a researcher can get access to Jetstream, and to other national compute clusters with more memory and for parallel processing. These compute resources have Globus connect subscriptions which assists in transferring terabytes of data quickly. In this workshop, we will demonstrate how to get an NCGAS allocation, set up a Jetstream account, spin up a preconfigured virtual machine for microbiome analysis (https://ncgas.org/Blog_Posts/Getting%20Started%20on%20Jetstream.php), and transfer data between compute clusters using Globus (https://ncgas.org/Blog_Posts/Getting%20Started%20with%20Globus.php).

C23: Digital Tools and Resources Session 4

iPat: A Genomics Analysis Tool for Everyone

Chunpeng James Chen and Zhiwu Zhang, Dept. of Crop and Soil Science, Washington State University, Pullman, WA

Genome-wide association study (GWAS) and genomic selection (GS) are two of the most critical analyses in practical breeding research. By far there're a couple of software tools available for the purposes, but few of them can provide a friendly environment for users to interact with. For example, most current tools only have interfaces in command-line environments, such as GAPIT, rrBLUP, and BGLR. This limitation imposes a serious barrier on breeders who were barely programming languages. Other than that, dealing with the compatibility of file formats is also a pain for users, none of the formats can globally work with all the tools. And it's also difficult to find a universal converter to resolve this problem.
In this study, we proposed iPat, intelligent prediction and association tool, as a solution to the mentioned issues. In iPat, a graphical user interface is friendly designed for everyone to use, users can easily perform profound analyses by multiple mouse clickings and draggings. Furthermore, more than three different formats are supported by iPat, including HapMap, VCF, PLINK, and numerical data. With files encoded in PLINK format, for instance, users can directly use them to perform GWAS in GAPIT or GS in BGLR without spending effort in converting the format. Besides, although it's a work published in 2018, there are several new features released in the recent update. One of the most important features is that iPat can now perform k-fold validation in GS. It can provide users with more convincing results and aid with their publication.
The software is available on http://zzlab.net/iPat/

C24: Digital Tools and Resources Session 4

Methods for Robust Bayesian Analyses of Complex Genetic Experiments

Anthony Greenberg, Bayesic Research, Ithaca, NY

Breeding programs rely on rich experimental designs to estimate or predict breeding values and advance their populations towards selection goals. Typically, multiple traits at several locations, across multiple years and conditions, are measured on replicated genotypes. Bayesian inference methods can be profitably employed to analyze such data sets, particularly if occasionally unfavorable trial conditions lead to partial data loss, presence of outlier observations, or imbalance in replication. Recent advances in numerical methods and computer infrastructure put Bayesian inference within reach of most public and private sector breeding programs. Although experimental designs may be complicated and vary from program to program, each organization usually employs a consistent approach. Thus, we need a set of methods and software tools that are customizable, yet can be deployed at a given institution and produce reliable results across years without significant adjustments.
Unfortunately, current model fitting methods are based on the Metropolis-Hastings algorithm which has severe limitations in this setting. This set of fitting methods relies on proposal distributions that require careful tuning. The simpler to implement variant of this approach, Gibbs sampling, does not require such adjustments, but can be numerically unstable and exhibits convergence problems even when fitting relatively simple models. Worse, these problems are influenced by parameter estimates themselves, in addition to the structure of the experimental design, and thus require constant supervision by experts.
A different approach, Hamiltonian Monte Carlo, has been in use for many years in some specialized settings. While it also requires a careful choice of tuning parameters, recent developments have produced methods to derive these adjustments automatically. I have implemented these approaches to analyze genetic experiments. I will show that they produce numerically robust estimates, with fast and reliable convergence. I will illustrate their application to simulated data, with comparisons to the widely-used Gibbs sampling. Software based on these methods will be publicly available. I hope it will empower even resource-limited organizations to deploy state-of-the-art statistical methods and thus accelerate progress in breeding programs world-wide.

C25: Digital Tools and Resources Session 4

OrthoQuery: A Tripal Database Module to Assess and Visualize Gene Family Evolution

Sumaira Zaman[1], Sean Buehler[2], Emily Grau[3], Stephen P. Ficklin[4] and Jill L. Wegrzyn[2], (1)Department of Computer Science and Engineering, University of Connecticut, Storrs, CT, (2)Department of Ecology and Evolutionary Biology, University of Connecticut, Storrs, CT, (3)University of Connecticut, Storrs, CT, (4)Dept of Horticulture, Washington State University, Pullman, WA

The abundance of transcriptomic resources for non-model organisms has enabled researchers to study comparative genomics on a larger scale. Generation of orthologous gene families facilitates comparative genomics by examining gene family evolution events related to selection pressures. Applications developed to study gene homology among species do not allow users to query data directly from external databases hosting resources that are not associated with a reference genome. Furthermore, real time computation of orthogroups for user selected subsets paired with interactive visualizations is lacking. OrthoQuery, a web-based Tripal module, provides a semi-automated analytical framework to enable comparisons among curated proteins and interactive visualizations in context of the resulting species tree. OrthoFinder, optimized with Diamond, is leveraged for protein level comparisons, and the Tripal database framework, coupled with Galaxy integration, supports various workflows and visualization options. OrthoQuery processes unigenes and stores a pre-computed set of orthogroups based on available species' resources in the local database. The module provides researchers with options to navigate the resulting species tree, identify ancestral/species-

specific groups of genes, and associate orthogroups with functional annotations. We demonstrate OrthoQuery's application when studying conifer genomes. Although the development of this module is intended for forest tree research, any of the other 30+ clade or organism specific databases supported Tripal will have access to OrthoQuery as part of Tripal's open source project. This robust and flexible Tripal module aims to enable researchers in conducting comparative genomics analysis for user selected species, with an emphasis on pre-processing transcriptomic resources to include non-model organisms.

C26: Digital Tools and Resources Session 4

Exploring a Landscape of Genetic Variation in Virtual Reality

Samuel Taler Westreich[1], Maria Nattestad[2] and Christopher Meyer[1], (1)DNAnexus, Mountain View, CA, (2)Cold Spring Harbor Laboratory, Cold Spring Harbor, NY

With the rise of large genomics and multi-omics datasets, it has become increasingly difficult for a human expert to leverage that data to its highest potential - it is simply too complex to intuitively understand. Great leaps in scientific thinking, including the double helix structure of DNA, have come from experts building three-dimensional models, as human brains are hard-wired for spatial reasoning.

Virtual reality now lets us more easily build and experiment with different representations of complex and large datasets. To use virtual reality to examine genomic data, we combined two iconic visualizations in genomics, the Manhattan plot and the circos plot, into a fully immersive data exploration experience called BigTop.

The results of a genome-wide association study (GWAS) on human height provides the starting point for our exploration. A scientist stands at the center of a circular room surrounded by points floating in the air. Each point represents a variant SNP, with a height according to its negative log10 transformed p-value, as in a regular Manhattan plot. The chromosomes of the human genome wrap around the circular room, similar to a circos plot. The third dimension indicates allele frequency, with rarer variants close to the center of the room, while the most common variants are close to the walls. This provides a useful perspective that allows a scientist to focus on rare or common alleles, and to discover patterns across all three dimensions. Using virtual reality controllers, the scientist can view the data from different perspectives and interacted with individual SNPs for additional information.

BigTop is open source at https://github.com/dnanexus/bigtop, and supports multiple datasets. It can be viewed using the most common commercial VR headsets or in a Chrome browser.