C01: Computer Demo 1

The Agronomic Linked Data (AgroLD) Project

Pierre Larmande, IRD, UMR DIADE, Institut de Biologie Computationnelle, Montpellier, France, Manuel Ruiz, CIRAD, UMR AGAP / CIAT, Montpellier Cedex 5, France, Nordine El Hassouni, INRA, Montpellier, France and Aravind Venkatesan, Institut de Biologie Computationnelle, Montpellier, France

The drastic growth in data in the recent years, within the Agronomic sciences has brought the concept of knowledge management to the forefront. Some of the factors that contribute to this change include a) conducting high-throughput experiments have become affordable, the time spent in generating data through these experiments are minuscule when compared to its integration and analysis; b) publishing data over the web is fairly trivial and c) multiple databases exist for each type of data (i.e. 'omics' data) with a possible overlap or slight variation in its coverage. In most cases these sources remain autonomous and disconnected. Hence, efficiently managed data and the underlying knowledge in principle will make data analysis straightforward aiding in more efficient decision making. At the Institute of Computational Biology (IBC), we are involved in developing methods to aid data integration and knowledge management within the domain of Agronomic sciences to improve information accessibility and interoperability. To this end, we address the challenge by pursuing several complementary research directions towards: distributed, heterogeneous data integration.

This talk will focus mainly on,Agronomic Linked Data (AgroLD) wich is a Semantic Web knowledge base designed to integrate data from various publically available plant centric data sources. These include Gramene, Oryzabase, TAIR and resources from the South Green platform among many others. The aim of AgroLD project is to provide a portal for bioinformaticians and domain experts to exploit the homogenized data towards enabling to bridge the knowledge.

C02: Computer Demo 1

Forest Tree GnpIS: an Information System Dedicated to Forest Tree Genetics, Genomics and Phenomics

Célia Michotey[1], Christel Anger[2], François Ehrenmann[3], Véronique Jorge[4], Odile Rogier[4], Catherine Bastien[4], Christophe Plomion[5], Christian Pichot[6], Hadi Quesneville[1] and Anne-Francoise Adam-Blondon[1], (1)INRA - URGI, Versailles, France, (2)INRA - GBFOR, Orléans, France, (3)INRA, Cestas, France, (4)INRA - AGPF, Orléans, France, (5)INRA - BIOGECO, Cestas, France, (6)INRA - URFM, Avignon, France

Due to the major technological advances both in genomics and phenomics, it is now possible to quickly obtain large amounts of data at low cost. One of the consequences is the critical need for data management, with the opportunity to make these datasets interoperable, thus enhancing their reuse and enrichment.

GnpIS is an original information system (IS) able to manage these data. It is designed to integrate and link genetic, genomic and phenomic data into a single environment, allowing researchers to explore information from different angles.

"Forest tree GnpIS" is a GnpIS focused on forest tree data. These resources are accessible through a web portal (https://urgi.versailles.inra.fr/gnpis/) whose main entry point is a tool using keywords for data discovery and dedicated tools facilitating more specific queries and data retrieval.

In order to make data submission easier for research teams, workflows are implemented to automate data flow between local sources, GnpIS and other international IS if necessary. This way, data from several forest tree species have been integrated into the Forest tree GnpIS (e.g. 16800 accessions, 18 phenotyping trials, 13700 markers and 1000 QTLs, cf. https://urgi.versailles.inra.fr/Species/Forest-trees/Database-overview). I will show you how to navigate through these data and more precisely: 1) from a genome browser and its annotation access genetic maps (via markers and QTLs), 2) from a QTL of interest get the phenotyping results of its trait and select the accessions with the most interesting results, 3) refine your selection by studying an accession pedigree and genotyping results.

C03: Computer Demo 1

Exploring Grapevine Gene Expression Patterns with Vespucci.

Marco Moretto, Paolo Sonego, Stefania Pilati, Giulia Malacarne, Laura Costantini, Lukasz Grzeskowiak, Giorgia Bagagli, Claudio Moser, M. Stella Grando and Kristof Engelen, Fondazione Edmund Mach, San Michele all'Adige, Italy

In the era of high-throughput omics biology, the most used technologies for large-scale transcriptional studies are microarrays and RNA-Seq. These data are required to be deposited in public repositories upon publication. Such repositories have the enormous potential to provide a broad view of how different experimental conditions lead to expression changes, by comparing gene expression changes across all possible measured conditions. Unfortunately, this is not a task easily achievable due to differences among experimental platforms that make direct comparisons difficult. Using an adapted version of COLOMBOS technology (http://colombos.fmach.it/), geared towards eukaryotes and crop species in particular, we developed the Vitis Expression Studies Platform Using COLOMBOS Compendia Instances (VESPUCCI), a gene expression compendium for grapevine that integrates thousands of gene expression samples from several technological platforms, both microarray and RNA-Seq. Each sample has been manually annotated using a controlled vocabulary developed ad hoc to ensure both human readability and computational tractability. Expression data in the compendium can be visually explored using several tools provided by the web interface or can be programmatically accessed using the REST interface. VESPUCCI is freely accessible at http://vespucci.colombos.fmach.it.

References:

http://journal.frontiersin.org/article/10.3389/fpls.2016.00633/full

C04: Computer Demo 1

Bio -TDS : BioQuery Tool Discovery System

Carol Lushbough, University of South Dakota, Vermillion, SD

Bioinformatics and computational biology play a critical role in bioscience and biomedical research. As researchers design their experimental projects, one major challenge is to find the most relevant bioinformatics toolkits that will lead them to new knowledge discovery from their data. The Bio-TDS (**B**io**Q**uery **T**ools **D**iscovery **S**ystems, http://biotds.org/) has been developed to assist researchers in retrieving the most applicable computational toolkits by posing questions as free text. The Bio-TDS is a flexible, three-layer architecture retrieval system that provides users from multiple bioscience domains (e.g. genomic, proteomic, bio-imaging) the ability to query over 12,000 analytic tool descriptions integrated from well-established, community repositories. The query expressions posed by potential researchers can vary from precise questions such as "RNASeq reads mapper for mammalian" or "reads mapping", to very domain specific free text expressions like "best tool for microbiology genome mapping using RNASeq".One of the primary components of the Bio-TDS system is the ontology and natural language processing workflow for annotation, curation, query processing, and evaluation. The Bio-TDS's scientific impact was evaluated using sample researcher questions retrieved from Biostars, a site focusing on **biological data analysis**. The Bio-TDS was compared to five similar bioscience analytic tool retrieval systems with the Bio-TDS outperforming the others in terms of relevance and completeness. The Bio-TDS provides researcher with the ability to associate their bioscience question with the most relevant computational tool-set required for the data analysis in their knowledge discovery process.

C05: Computer Demo 1

Spell-QTL, a New Tool for QTL Analysis on Modern Datasets

Damien Leroux and Sylvain Jasson, MIAT, Institut National de la Recherche Agronomique, Castanet-Tolosan CEDEX, France

Spell-QTL (https://mulcyber.toulouse.inra.fr/projects/spel/) performs classical QTL analysis in modern datasets. Complex pedigrees have recently emerged, such as MAGIC8, that make classic population models obsolete. Inferring Parental Origins has become a tedious task and genotype data is often raw SNP data. We implemented innovative methods to handle virtually any pedigree and precisely compute the Parental Origin Probabilities (POP) at any locus given any number of observed markers.

We construct a Bayesian Network matching the structure of the pedigree. The exact POP at each observed marker are inferred using this network. Typically, marker observations involve the ancestors and the phenotyped generation, but our method can take observations on any individuals and produce POP for any subset of individuals. Along a linkage group, the probabilities of recombination in each generation are modelled with Continuous-Time Markov Chain (time here being genetic distance) which enable precise computation of the POP at any locus, taking into account all the information provided by the neighboring markers. The QTL analysis is performed with a classical model selection approach.

The software suite is divided in three distinct binaries. *Spell-pedigree* analyzes the pedigree and computes the CTMCs. *Spell-marker* computes the POP at the markers. *Spell-QTL* performs the QTL detection and analysis. We emphasize speed and allow concurrency or even distributed computations wherever appropriate. We provide a rich and fully-documented command-line interface and the results are output in text files for convenient import in R or Excel.

C06: Computer Demo 1

Analysis of Genomic Variability within Inbred Populations with NGSEP

Jorge Duitama[1], Juan David Lobaton[2] and Juan Fernando De la Hoz[2], (1)Universidad de los Andes, Bogotá, Colombia, (2)International Center for Tropical Agriculture (CIAT), Cali, Colombia

The development and availability of high throughput sequencing (HTS) technologies and different bioinformatic tools revolutionized the research on genomics allowing to obtain genome-wide data on entire populations of nearly every form of life. Research communities of inbred species, such as rice or beans, are developing large datasets of variability from Whole Genome Sequencing (WGS) data, and also performing Genotype by Sequencing (GBS) experiments on structured mapping and breeding populations. Unlike outbred species, haplotypes in these populations can be obtained and analyzed directly from genotype calls. However, most bioinformatics tools can not take advantage of this feature because they were mostly developed for human genetics. Our bioinformatics group recently developed the Next Generation Sequencing Experience Platform (NGSEP) for accurate, efficient, and user-friendly analysis of HTS data. We showed using both WGS and GBS data that, for inbred species, NGSEP provides better accuracy compared to other commonly used software for variants detection. Here we present the latest functionalities implemented in NGSEP for analysis of genomic variability datasets in VCF format from populations of inbred samples: 1) A Hidden Markov Model (HMM) for genotype imputation, from either known or unknown parental haplotypes, 2) Assessment of diversity across the genome through allele sharing statistics, and 3) Analysis of introgression of foreign haplotypes within subpopulations. These analysis options have been useful to analyze large genomic datasets such as that provided by the 3000 rice genomes project and we expect they will be useful for further sequencing efforts in different inbred crops.

C07: Computer Demo 1

NCBI Genomic Data Viewer

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

The National Center for Biotechnology Information (NCBI) has developed a new browser resource, the Genomic Data Viewer (GDV), to display and analyze genome assemblies for a wide variety of organisms. Built on NCBI's existing browser technology used for the Variation Viewer and 1000 Genomes Browser, GDV makes use of the wealth of sequence associated data stored at NCBI to provide browser tracks and also supports displays for other datatypes, such as genotypes and allele frequencies. In addition to the display of NCBI-provided tracks, GDV supports the upload and display of user-provided data, accepting a variety of common file formats. GDV browser instances are available for any RefSeq assembly in the NCBI Assembly database that has track data. NCBI resources that refer out to GDV can customize the default track display to include the tracks most relevant to their content area. For instance, GDV referrals from NCBI Variation resources can be configured to show SNVs and structural variations from dbSNP and dbVar in genomic context with gene annotations, and a dynamic table of sample genotypes and population allele frequency. In addition, the browser display is easily user-configured, either to custom track combinations or with preset combinations of tracks relevant to various analysis types. Track configurations can be saved and shared with other users. This presentation will demonstrate some of the available functionality in GDV, including searching, navigating, accessing and configuring tracks, uploading user data, and provide examples of the flexible use of the GDV at NCBI.

C08: Computer Demo 1

Easymirror and Easyimport: Set up Your Own Ensembl Site in 2 Hours for Your Favourite Taxa

Sujai Kumar, The University of Edinburgh, Edinburgh, United Kingdom and **Richard J Challis**, Edinburgh University, Edinburgh, United Kingdom

A powerful reason to choose Ensembl over alternative genome browsers is the integration of comparative genomics and other datatypes (such as structural variants and functional annotation) in a genome browser with a mature API. However Ensembl is generally considered to be difficult to install and adding new genomes requires considerable effort. We have developed tools to make this process simpler. EasyMirror allows you to set up a new site with remote and local data within a few minutes while EasyImport allows you to import genomic sequence, gene models and annotations for any taxon in around an hour. We will demonstrate how to set up an Ensembl database and website, and populate it with your own genome assemblies and annotations, along with data from other existing Ensembl species databases. We will also show how to perform additional analyses, such as orthology analyses, to view your data in the context of other species hosted on other Ensembl servers.

C09: Computer Demo 2

GOBII: Genomic Open-source Breeding Informatics Initiative

Philip Glaser[1], Yanxin Gao[2], Elizabeth Jones[3], Yaw A. Nti-Addae[2], Kevin Palis[4], Syed Raza[2], Joshua Lamos-Sweeney[2], Angel Villahoz-Baleta[2], **Kelly Robbins**[2], Jean-Luc Jannink[5], Lukas Mueller[4], Mark E Sorrells[2], Qi Sun[6], Edward S. Buckler[7] and Susan McCouch[2], (1)Cornell Institute of Biotechnology, Ithaca, NY, (2)Cornell University, Ithaca, NY, (3)GOBII Project, Biotechnology Dept, Cornell University, Ithaca, NY, (4)Boyce Thompson Institute, Ithaca, NY, (5)USDA-ARS, Ithaca, NY, (6)Institute for Genomic Diversity, Cornell University, Ithaca, NY, (7)USDA-ARS-Cornell University, Ithaca, NY

In the last ten years, genotyping costs have dropped significantly, making feasible powerful new breeding approaches that can take advantage of the vast amounts of genomic data that have been generated. The Genomic Open-source Breeding Informatics Initiative (GOBII) is the first large-scale public-sector effort to provide an open-source solution for storing, querying, and applying analysis pipelines to high-density genotypic information. The GOBII system is comprised of a data warehouse (PostgreSQL, MonetDB, and HDF5), a web service layer with a BrAPI compliant RESTful API, and GUIs for loading and extracting data. Technologies and key features comprising the system will be presented.

C10: Computer Demo 2

Gigwa - Genotype Investigator for Genome-Wide Analyses.

Manuel Ruiz, CIRAD, UMR AGAP / CIAT, Montpellier Cedex 5, France, Guilhem Sempéré, CIRAD, UMR Intertryp, Montpellier, France, Adrien Petel, CIRAD, St Pierre, French Southern Territories, Alexis Dereeper, IRD, UMR RPB, F-34394 Montpellier, France, Gautier Sarah, INRA - UMR AGAP, Montpellier, France and Pierre Larmande, IRD, UMR DIADE, Institut de Biologie Computationnelle, Montpellier, France

Exploring the structure of genomes and analyzing their evolution is essential to understanding the ecological adaptation of organisms. However, with the large amounts of data being produced by next-generation sequencing, computational challenges arise in terms of storage, search, sharing, analysis and visualization. This is particularly true with regards to studies of genomic variation, which are currently lacking scalable and user-friendly data exploration solutions. Here we present Gigwa, a web-based tool that provides an easy and intuitive way to explore large amounts of genotyping data by filtering it not only on the basis of variant features, including functional annotations, but also on genotype patterns. The data storage relies on MongoDB, which offers good scalability properties. Gigwa can handle multiple databases and may be deployed in either single- or multi-user mode. In addition, it provides a wide range of popular export formats. The Gigwa application is suitable for managing large amounts of genomic variation data. Its user-friendly web interface makes such processing widely accessible. It can either be simply deployed on a workstation or be used to provide a shared data portal for a given community of researchers.

C11: Computer Demo 2

The Brassica Information Portal: Towards Integrating Phenotype and Genotype Data.

Annemarie Eckes[1], Tomasz Gubala[1,2], Piotr Nowakowski[1,2], Tomasz Szymczyszyn[1] and Wiktor Jurkowski[1], (1)Earlham Institute, Norwich, United Kingdom, (2)Cyfronet AGH, Krakow, Poland

The Brassica Information Portal (BIP) is a web repository for population and trait scoring information related to Brassica breeding. High throughput phenotyping has increased the need to make the rapidly accumulating volumes of phenotype data available and accessible in a standardised manner. Serving as a community resource, the BIP helps define common data and metadata standards relevant for describing Brassica diversity. This is especially important for genotype-phenotype association studies in crop improvement.

BIP provides access to trial data including metadata and trait data as well as QTL and marker data linked to traits. It further includes population data with accession and line information, cross-linked to their associated sequences in SRA.

BIP hosts lines assessed for development and agricultural traits which can be used for GWAS analysis, for marker development and ultimately crop improvement. Submission and extraction of data is possible using the web interface and the API. Future aims are to expand our database to hosting SNP next to marker data. We also plan to further integrate our resource with ensembl plants and CyVerse to provide single-entry point access to analysis tools and relevant visualisations.

The participant will learn how to navigate the resource and submit data based on a use-case scenario. We will report on our efforts to identify minimum metadata as well as trait definitions. Our aim is to demonstrate BIP's value and utility, both for single research groups and the global Brassica research and breeding community.

C12: Computer Demo 2

GnpIS-Ephesis, Plant Phenotype Field Experimentations Resources – Data Discovery and Dataset Building Use Cases.

Cyril Pommier[1], **Guillaume Cornut**[1], Michael Alaux[1], Thomas Letellier[1], Célia Michotey[1], Erik Kimmel[1], Sophie Durand[1], Raphael Flores[1], Florian Philippe[1], Mathide Lainé[1], Eric Duchene[2], Thierry Lacombe[3], Francois-Xavier Oury[4], Gilles Charmet[5], Arnaud Gauffreteau[6], Hadi Quesneville[1] and Anne-Francoise Adam-Blondon[1], (1)INRA - URGI, Versailles, France, (2)INRA-

SVQV, COLMAR, France, (3)INRA - Domaine de Vassal, Marseillan-plage, France, (4)INRA Site de Crouël, CLERMONT-FERRAND, France, (5)INRA UMR GDEC Clermont-Fd, France, Clermont-Ferrand, France, (6)French National Institute of Agronomy (INRA) Agronomy, Thiverval Grignon, France

Integrating phenotyping data from divers sources allow to conduct meta-analysis for climate change studies, genetic studies or genotype by environment studies. This integration relies on pivot object and data formatting. Due to the heterogeneity of Phenotyping data, this process is often complex, but we will show what are its benefices through GnpIS. This Information System has been built to allow the integration of genomic, genetic and phenomic data for plant and their pathogens and relies on identifying common pivot resources like germplasm, observation variables following the Crop Ontology framework, experimental locations and years. This allows to link together trials conducted by experimental facilities including field networks and high throughput phenotyping facilities in controlled environments or fields.

We will first demonstrate two approaches, datadiscovery and dataset building. The first one, implemented in GnpIS portal (https://urgi.versailles.inra.fr/gnpis) and the WheatIS (http://wheatis.org/Search.php) shows an overview of available data. The second approach allows to build and download datasets for yield evolution genotype by environment studies. This demonstration relies on public data (doi:10.15454/1.4489666216568333E12) provided by INRA . It includes fifteen years of observations on eleven experimental sites including agronomic and disease observations. We will then show how to build a climate change study dataset by getting phenology data for perennial species like Grape or Poplar, from which we can extract dataset for statistical analysis tools or model to evaluate adaptability of several hundreds of variety.

Aknowledgments
Duchène E, Lacombe T, Oury FX, Charmet G, Gauffretau A, Labernardière M, Lebreton A., Roumet P.

C13: Computer Demo 2

Knowledge.Bio: A Web Application for Collaboratively Building and Exploring Networks of Biological Relationships.

Richard Michael Bruskiewich[1], Kenneth Casimir Huellas-Bruskiewicz[1,2], Chandan Kumar Mishra[1,2], Farzin Ahmed[1,2], Yinglun Colin Qiao[1,2], Jarielle D Lim[1,2], Lance Hannestad[1,2], Rudy Kong Tin Lun[1,2] and Benjamin M Good[3], (1)STAR Informatics / Delphinai Corporation, Port Moody, BC, Canada, (2)Simon Fraser University, Burnaby, BC, Canada, (3)The Scripps Research Institute, La Jolla, CA

Knowledge.Bio (Bruskiewich et al., 2016) is an open source web resource to enhance access and interpretation of networks of concepts and relationships derived from text mining or curated public resources. Each edge in these networks is linked to underlying evidence (e.g. an abstract in PubMed), which may be accessed within the context of the application. Along the way, users construct their own graphical map of related concepts which can be exported and shared either locally as a file on a user's computer, or remotely within an indexed map library. The initial research literature knowledge base of Knowledge.Bio is drawn from the Semantic Medline Database (https://skr3.nlm.nih.gov/SemMed/; Kilicoglu et al., 2012) and the Implicitome (Hettne et al., 2016), two complementary resources derived from text mining of NCBI Pubmed abstracts. Additional cross-linkages are provided to external biological databases via integration with WikiData (https://www.wikidata.org; Burgstaller-Muehlbacher 2016). In addition, Knowledge.Bio data loading protocols allow additional concept and relationship data to be loaded from other data sources.

Availability and Implementation: hosted at http://knowledge.bio/; open source code available at http://bitbucket.org/sulab/kb2/.

Contact: richard@plantinformatics.com

References:

Bruskiewich, RM, et al. (2016) Knowledge.Bio: A Web application for exploring, building and sharing webs of biomedical relationships mined from PubMed.doi: http://dx.doi.org/10.1101/055525

Burgstaller-Muehlbacher, Sebastian, et al. (2016) "Wikidata as a semantic framework for the Gene Wiki initiative." Database 2016 : baw015.

Hettne, KM, et al. (2016) The Implicitome: A Resource for Rationalizing Gene-Disease Associations.PLoS ONE 11(2): e0149621. doi:10.1371/journal.pone.0149621

Kilicoglu, H, et al. (2012) SemMedDB: a PubMed-scale repository of biomedical semantic predications.Bioinformatics, 28(23), 3158-60.

C14: Computer Demo 2

The Triticeae Toolbox (T3): Connecting Phenotypes, Genotypes, and Biological Knowledge

Clay Birkett, USDA-ARS, Ithaca, NY, Clare Saied, Cornell University, Dept. of Plant Breeding and Genetics, Ithaca, NY, David Matthews, Cornell University, Ithaca, NY, David Hane, University of California / USDA-ARS-WRRC, Albany, CA and Jean-Luc Jannink, USDA-ARS / Cornell University, Ithaca, NY

The Triticeaetoolbox (T3) is a database for wheat, barley, and oat that contains phenotype and genotype data used by plant breeders. The database facilitates collaboration by providing reports, statistical analysis, and genome wide association studies. T3 provides links to GRIN-Global, Ensembl Plant, URGI, Planteome and Crop Ontologies, and GrainGenes. Data is integrated with reference sequence resources using JBrowse genome browser. We provide a plant breeding application programing interface (BrAPI), which allows 3rd party analysis tools to directly access the data. T3 software is available under the GNU General Public License and is freely downloadable from GitHub. An overview of the advanced T3 tools and installation requirements for a T3-like database will be discussed.

C15: Computer Demo 2

e!Dal - a Open Source Software to Store, Share and Publish Research Data

Matthias Lange, Daniel Arend and Uwe Scholz, Leibniz Institute of Plant Genetics and Crop Plant Research (IPK) Gatersleben, Stadt Seeland, Germany

e!DAL (https://edal.ipk-gatersleben.de/) is an open source software to **publish** and **share** research data in compliance to journal, funding agencies and open data policies (Arend et al. 2014 BMC Bioinformatics). Designed as broker between in-house file storage and data registries, it is provided as an all-in-one but small JAVA package that can be operated as micro service at any network accessible server. Prerequisites are a mounted, sufficient sized file system, access to an e-mail server, and open http(s) ports. Recommendable is a registration as DataCite data centre to enable e!DALs support for an automatic registration of published data sets as **journal and founding agency proven citable digital object identifiers** (**DOI**). Already embedded features are:

- http(s) server to support web and DOI access to data sets,
- database to store and associate DataCite and Dublin core (ISO 15836) compliant meta data,
- data submission tool,
- e-mail based reviewing and approval system,
- access logging and statistics for data publication metrics,
- API to embed into custom JAVA clients,
- WebDAV support to mount e!DAL as NAS file system,
- OAI-PMH meta data harvesting protocol
- version control system
- text search engine

e!DAL is already the basis for the PGP research data repository (Arend et al. 2016 Database: The Journal of Biological Databases and Curation) where 113 DOIs and more than 150.000 files are published. e!DAL driven PGP repository is accepted as data repository for **Nature Publishig Group**, is registered in **re3data.org**, **OpenAIRE** and **DataCite.** This shows e!DAL capabilities to build individually configured research data from the scratch or use e!DAL as API in custom bioinformatics tools or databases.

In the software demo session we will show step-by-step how to set-up an e!DAL instance, demonstrate the data submission process and give some best practice how to operate an in-house e!DAL based research data repository.

C16: Computer Demo 2

User-Friendly Whole Genome DNA Methylation Analysis With FlowGe

Jeffrey Grover[1], Matthew Bomhoff[1], Sean Davey[1], Brian D. Gregory[2], Rebecca A. Mosher[1] and Eric Lyons[1], (1)University of Arizona, Tucson, AZ, (2)University of Pennsylvania, Philadelphia, PA

Acquiring the specialized knowledge and resources required to manage, process, analyze, and visualize data from next generation sequencing (NGS) experiments is a bottleneck for many researchers. To address this issue we have integrated a suite of bioinformatics workflows into the web-based comparative genomics platform, CoGe. We term CoGe's integrated NGS workflows FlowGe. FlowGe's workflows include RNAseq, whole-genome bisulfite sequencing, ChIP-seq, variant calling (SNPs), and population genetics calculations. Unlike other standalone next-generation sequencing applications, FlowGe leverages the over 27,000 genomes currently available within CoGe and enables users to take advantage of CoGe's intuitive user interface. This facilitates rapid data processing, and integration of visualization and analysis in a single platform. Importantly, FlowGe allows researchers to control access to their data and share it with others, as well as combine public and private NGS data to answer genome-scale questions. CoGe's genome visualization system, NGS analysis workflows, and its connection to the CyVerse Data Store provide an integrated ecosystem to manage NGS data throughout its life cycle and facilitate greater accessibility of genomics analyses to researchers of all levels. To demonstrate FlowGe's ease of use and utility to all researchers we will demonstrate analysis of whole-genome bisulfite sequencing data to determine global DNA methylation levels.

FlowGe is freely available on the web at https://genomevolution.org under the MIT open source license (source code on GitHub https://github.com/LyonsLab/coge).

C17: Computer Demo 3

GrainGenes: Supporting the Small Grains Community

Taner Z. Sen, USDA -ARS / GrainGenes, Albany, CA, Gerard R. Lazo, USDA Agricultural Research Service, WRRC, Albany, CA, Yong Q. Gu, USDA ARS, Western Regional Research Center, Albany, CA, David Hane, University of California / USDA-ARS-WRRC, Albany, CA and Sarah Odell, USDA-ARS, Albany, CA

Funded with hard funds by USDA, GrainGenes provides long-term data sustainability for small grains researchers and hosts a range of community newsletters, databases, and digital workspaces for wheat, barley, rye, and oats. GrainGenes is a gateway for integrated access to several types of peer-reviewed and curated genomic, genetic, and phenotypic data, along with QTLs and other experimental outcomes. The availability of reference genome assemblies of wheat and barley, along with their diversity data, is making a significant impact at GrainGenes and we are creating genome-centric views on our interface with rich links to data that is already housed at GrainGenes, curated over decades. We recently updated the GrainGenes Genome Browser with JBrowse, and created training videos for our users to smooth the learning curve for the new interface. GrainGenes will continue creating/implementing new tools and views for the small grains community, supporting them in their research, and providing them a long-term repository for their peer-reviewed experimental and computational data.

C18: Computer Demo 3

RefEx, a Reference Gene Expression Dataset As a Web Tool for the Functional Analysis of Genes.

Hiromasa Ono and Hidemasa Bono, Database Center for Life Science, Mishima, Japan

RefEx (Reference Expression dataset; http://refex.dbcls.jp) provides suitable datasets as a reference for gene expression data of normal human, mouse, and rat tissues and cells. Four different measurement strategies are used in our collected gene expression data. These are: expressed sequence tag (EST), Affymetrix GeneChip, cap analysis of gene expression (CAGE), and transcriptome sequencing (RNA-seq). By showing in parallel with gene expression dataset in normal 40 organs (10 major groups) was obtained by four different experimental methods, you can make an intuitive comparison among gene expression values but also the methods. Users can examine the expression profiles of unfamiliar genes in normal tissues of the body, cells, and cell lines, from actual measurement data, rather than only from a description in a journal article. Recently, we incorporated CAGE data from the FANTOM5 project into RefEx. RefEx contains unique lists of genes whose expression pattern is prominent in a specific tissue compared with other tissues. Clicking the tissue icons on the RefEx top page easily retrieves genes with tissue specific expression patterns. In addition, up to three genes that have been added to a user's list can be compared at the same time. Users can compare all the detailed information about genes in that list, including expression data, in parallel. This enables users to easily find differences among the genes. In this way, RefEx is also useful as a tool for investigating the relationships of unknown genes found in gene expression analyses.

C19: Computer Demo 3

Newly Designed Genome Database for Rosaceae (GDR)

Sook Jung[1], Taein Lee[1], Chun-Huai Cheng[1], Jodi L. Humann[1], Ping Zheng[1], Sushan Ru[1], Kristin Scott[1], Morgan Frank[1], Deah McGaughey[1], Ksenija Gasic[2], Jim McFerson[3], Cameron Peace[1], Katherine M. Evans[4], Lisa DeVetter[1] and Dorrie Main[1], (1)Washington State University, Pullman, WA, (2)Clemson University, Clemson, SC, (3)Washington State Univesity, Wenatchee, WA, (4)Washington State University, Wenatchee, WA

A newly designed Genome Database for Rosaceae (GDR: www.rosaceae.org), with features such as a major genera and tools quick start in the homepage, has been launched. New features to quickly familiarize users with GDR data and functionality also include a dynamically generated data overview page browseable by data type and number and short video tutorials for the site overview, species pages and all the search pages. New data includes GDR generated datasets such as reference transcriptomes for Rosaceae genera that combines published RNA-Seq and EST data sets, and conserved syntenic regions among Prunus persica v2.0, Fragaria vesca v2.0, Malus x domestica v3.0p, and Pyrus communis v1.0. New genome assemblies and annotations along with GDR functional annotations are available, together with significant additions to map, marker and QTL data. Our effort toward data integration across databases, organisms and data types continues with development of Rosaceae Trait Ontology, curation of more data types with ontologies and developing interfaces that allow searching of integrated data more efficiently.

C20: Computer Demo 3

Quality Assessment Using BUSCO v2

Felipe A. Simão, Robert M. Waterhouse, Mathieu Seppey, Panagiotis Ioannidis, Evgenia V. Kriventseva and Evgeny M. Zdobnov, University of Geneva Medical School & Swiss Institute of Bioinformatics, Geneva, Switzerland

An important driving force behind genome projects is the acquisition of a complete catalog of genes, thus, it is vital to assess the integrity and completeness of the assembled genome. Although indications of assembly quality may be gleaned from statistical measures, a key measure of assembly quality is completeness in terms of the expected gene content. The identification of genes from many diverse species that are evolving under single-copy control defines an evolutionarily-informed expectation regarding gene content. Selected from the major species clades of the OrthoDB catalog of orthologs, the Benchmarking Universal Single-Copy Orthologs (BUSCOs) defines sets of genes expected to be present in any newly-sequenced genome from the appropriate species clade. The BUSCO assessment tool implements a computational pipeline to identify and classify matches from genome assemblies, annotated gene sets, or transcriptomes. The new version offers improved features and greatly expands the number of clades covered by BUSCO, including plants and protists. During this demonstration we will show how BUSCO can be used to facilitate genomic research, including examples on: de novo assemblies, re-annotation, assembly updates and phylogenetic analyses. Project website: http://busco.ezlab.org/v2/

C21: Computer Demo 3

PMN, A Unified Resource For Plant Metabolism

Peifen Zhang, Carnegie Institution for Science, Stanford, CA

Plant metabolism is responsible for the majority of food, feed, and medicine production. However, despite the significant role that plant metabolism plays we still know very little about it. To better understand plant metabolism and facilitate enzyme/pathway discovery and metabolic engineering, we developed a comprehensive infrastructure through which known information about plant metabolism can be easily accessed and a sequenced genome can be annotated into enzymes, reactions and pathways with high quality. The infrastructure includes a machine-learning-based enzyme function prediction pipeline (E2P2), a pathway inference software (Pathway Tools, developed by SRI International), and a semi-automated pipeline for validating inferred pathways (SAVI). Using this infrastructure, we predicted the metabolic networks of 22 sequenced genomes in the green lineage (10 monocots, 9 eudicots, 2 lower plants and 1 green algal species). We have also developed an extensively curated pan-plant biochemical pathway database called PlantCyc that catalogs a comprehensive set of known plant biochemical pathways involved in primary and specialized metabolism. Major data types in all databases include pathway diagrams and summaries, reaction equations, compound chemical structures, and enzyme properties. Evidence codes along with references are attached to pathways and enzymes for quality assurance. The databases we generated are available online at the Plant Metabolic Network (PMN) website (www.plantcyc.org). Tools provided at our web site allow users to display and examine large scale “omics” data in a metabolic context, to compare pathways between species, and to BLAST against specific enzyme datasets.

C22: Computer Demo 3

Planteome: Reference Ontologies and a Platform for Integrative Plant Genomics

Austin Meier[1], Laurel Cooper[2], Pankaj Jaiswal[2], Barry Smith[3], Christopher Mungall[4], Marie-Angélique Laporte[5], Elizabeth Arnaud[5], Justin L. Elser[2], Justin Preece[2], Sinisa Todorovic[1] and Eugene Zhang[1], (1)Oregon State University, Corvallis, OR, (2)Department of Botany & Plant Pathology, Oregon State University, Corvallis, OR, (3)University at Buffalo, Buffalo, NY, (4)Lawrence Berkeley National Laboratory, Berkeley, CA, (5)Bioversity International, Montpellier Cedex 5, France

The Planteome project is a centralized online plant informatics portal which provides semantic integration of widely diverse datasets with the goal of plant improvement. Traditional plant breeding methods for crop improvement may be combined with next-generation analysis methods and automated scoring of traits and phenotypes to develop improved varieties. The Planteome project (www.planteome.org) develops and hosts a suite of reference ontologies for plants associated with a growing corpus of genomics data. Data annotations linking phenotypes and germplasm to genomics resources are achieved by data transformation and mapping species-specific controlled vocabularies to the reference ontologies. Analysis and annotation tools are being developed to facilitate studies of plant traits, phenotypes, diseases, gene function and expression and genetic diversity data across a wide range of plant species. The project database and the online resources provide researchers

tools to search and browse and access remotely via APIs for semantic integration in annotation tools and data repositories providing resources for plant biology, breeding, genomics and genetics.

C23: Computer Demo 4

COPO: A Data Stewardship Platform for Plant Scientists

Felix Shaw[1], Anthony Etuk[1], Alejandra Gonzalez-Beltran[2], David Johnson[2], Philippe Rocca-Serra[2], Paul J. Kersey[3], Ruth Bastow[4], Katherine Denby[5], Susanna A. Sansone[2] and Robert P. Davey[1], (1)Earlham Institute, Norwich, United Kingdom, (2)Oxford e-Research Centre, University of Oxford, Oxford, United Kingdom, (3)EMBL - The European Bioinformatics Institute, Cambridge, United Kingdom, (4)Global Plant Council; University of York, York, United Kingdom, (5)University of York, York, United Kingdom

We present Collaborative Open Plant Omics (COPO), a brokering service between plant scientists and public repositories, which enables management, aggregation and publication of research outputs described and integrated using linked data. COPO provides consolidated access tc services and disparate information sources via a web interface and Application Programming Interfaces.

Users create profiles which represent a logical span of research, such as a grant funding round or PhD project. 'Research objects' comprising a broad spectrum of potential outputs (such as sequence data, images, manuscripts, source code, posters) can be uploaded into the profile. Annotation of these objects with community-supported standards is facilitated using simple user-interface wizards which aim to reduce the complexity of this task, supported by ISA components (http://isa-tools.org/) for metadata interoperability and automated metadata format conversion. COPO uses the Ontology Lookup Service (http://www.ebi.ac.uk/ols) to provide the crucial contextual metadata required for standardised data description. Currently, deposition of both data and metadata to the European Nucleotide Archive for sequence data, and Figshare for data types such as images, posters, and presentations is supported.

In the future, we will support more public repositories for multi-omic data submission, and users will be able to search for and pull such data into analysis environments such as CyVerse and Galaxy. We will subsequently track the outputs and associated metadata in COPO, thus creating a provenance trail from data to publication.

http://copo-project.org/

https://github.com/collaborative-open-plant-omics/COPO

C24: Computer Demo 4

Systems Biology Resources for the Citrusgreening Disease Complex.

Prashant S Hosmani[1], Surya Saha[1], Mirella Flores-Gonzalez[1], Susan Brown[2] and Lukas Mueller[1], (1)Boyce Thompson Institute, Ithaca, NY, (2)Kansas State University, KSU Bioinformatics Center, Manhattan, KS

Huanglongbing (HLB) is a tritrophic disease complex involving citrus host trees, the Asian citrus psyllid (ACP) insect and a phloem restricted, bacterial pathogen *Candidatus* Liberibacter asiaticus (CLas). HLB is considered to be the most devastating of all citrus diseases, and there is currently no adequate control strategy. In Florida, an estimated 40-70% of all citrus trees are infected, and HLB effects include production declines (10-20% per year), diminished fruit quality and increased production costs. We have designed a web portal with information for consumers and growers as well as genomics and bioinformatics resources for citrus, ACP and CLas. The JBrowse browser provides the context for expression data and annotated features on the genome. Biocyc Pathway Tools databases model biochemical pathways within each organism and will be used to explore the entire disease complex. Micro-CT analysis of the ACP will be combined with transcriptomics data from different tissues, life stages and sexes to create a 3D atlas that will reveal the internal anatomy of ACP overlaid with the expression profile of different tissues across major life stages. All tools like JBrowse, Blast and the Atlas will connect to a central database containing gene models for citrus, ACP and multiple *Candidatus* Liberibacter pathogens. The database will allow manual curation so that the community can continuously improve the knowledgebase as more experimental research is published. The database architecture combines custom schemas and the community-developed Chado schema (http://gmod.org/wiki/Chado) for compatibility with other genome databases. The portal can be accessed at https://citrusgreening.org/.

C25: Computer Demo 4

SNP-Seek Resource for Rice Research

Locedie Mansueto, Roven Rommel Fuentes, Nikki Borja, Jeffrey Detras, Juan Miguel Abriol-Santos, Dmytro Chebotarov, Millicent D. Sanciangco, N. Ruaraidh Sackville Hamilton, Ramil P. Mauleon, Kenneth L. McNally and Nickolai Alexandrov, International Rice Research Institute, Los Baños, Philippines

The 3,000 rice genome project generated a large dataset of genomic variation to the world's most important crop. We identified ~40M SNPs using BWA aligner and GATK variant calling on five reference genomes of rice, representing the major variety groups: Nipponbare (*jap*), IR 64 (*ind*), 93-11(*ind*), DJ 123 (*aus*) and Kasalath (*aus*). The results are accessible through the Rice SNP-Seek website (http://snp-seek.irri.org) or by web-services with defined APIs. SNP-Seek provides legacy phenotypic and passport data for all the sequenced varieties from the International Rice Genebank Information System at IRRI and incorporates gene models from several rice annotation projects.

Inside SNP-Seek, the massive genotypic data are stored as HDF5 files for fast retrieval, while the germplasm, phenotypic and genomics data are in RDMS using the CHADO schema, allowing the use of controlled terms from biological ontologies as query constraints. Several visualization tools are embedded including: JBrowse for SNPs, indels and genes; VISTA for the genome alignments; phylogenetic tree and MDS plots to explore the evolutionary relationships between the varieties; allele frequency chart and the genotype matrix viewer. A list manager for SNPs, varieties and genes allows the user to pass results as constraints in their queries. The SNP Annotator can add traits, ontology terms, functional effects and interactions to the markers in a list. In this demo we will present the data and features of the Rice SNP-Seek Database, and show typical use cases focusing on allele-mining and breeding applications.

C26: Computer Demo 4

Elasticsearch Indexing/Search and Expression Data Visualization with Two New Tripal Extension Modules

Ming Chen[1], Nate Henry[1], Thomas Lane[1], Xiao Zhou[1], Jill L. Wegrzyn[2], Stephen P. Ficklin[3] and Margaret Staton[1], (1)University of Tennessee, Knoxville, TN, (2)Department of Ecology and Evolutionary Biology - University of Connecticut, Storrs, CT, (3)Washington State University, Pullman, WA

The Hardwood Genomics Databases (HGD) is a Tripal[1]-based web database offering access to genetic and genomic data from ecologically and phylogenetically important hardwood tree species. HGD currently hosts transcriptomes from 15 species, low coverage whole genome sequence data from 10 species, SSR markers from 8 species, and the Chinese chestnut reference genome. With the rapid increase in the number of open access tree transcriptomes sequences and the accompanying gene expression data, the HWD has built new tools for researchers to search, download, and visualize this data. First, we have a new search engine, built from the open access elasticsearch[2] software, that provides powerful, full text search and includes features such as fuzzy searching that handle misspelled or alternately spelled words. The general site-wide search for all content types is now complemented by an advanced transcript-specific search allowing users to specify functional annotation details and filter by organism. Second, a new Tripal Expression Analysis module supports three new content types relevant to transcriptome projects: biological sample records, expression protocol records and gene expression levels for sequence features. This allows the user to explore a public RNASeq experiment fully, from collection of tissues, to lab methods and analysis of data, to gene expression levels. The feature expression levels across multiple biomaterials are visualized with interactive bar plots or heat maps that are generated with javascript graphic libraries D3JS[3] and plotly[4]. Additionally, we offer a tool for users to build customized, two dimensional heat maps for gene expression comparison by submitting a list of feature IDs. The customized heat map provides detailed library information and links to the feature content pages. The implementation of these two modules in the HGD will facilitate hardwood trees research in aspects of data acquisition and discovery, exploration of candidate genes, and comparative gene expression analysis among tree species.

Keywords: Tripal, database, elasticsearch, searching, gene expression, heatmap

1. Ficklin, Stephen P., et al. "Tripal: a construction toolkit for online genome databases." Database 2011 (2011): bar044.
2. Elasticsearch. https://www.elastic.co/. Accessed Oct 24, 2016.
3. D3JS. https://d3js.org/. Accessed Oct 24, 2016
4. Plotly. https://plot.ly/. Accessed Oct 24, 2016

C27: Computer Demo 4

Toggle-3 : A Framework to Build Quickly Pipelines and to Perform Large-Scale NGS Analysis

Christine Tranchant-Dubreuil[1], Sebastien Ravel[2], Cécile Monat[1], Laura Helou[1], Abdoulaye Diallo[1], Gautier Sarah[3], Julie Orjuela-Bouniol[4] and François Sabot[1], (1)IRD - UMR DIADE, Montpellier, France, (2)CIRAD, Montpellier, France, (3)INRA - UMR AGAP, Montpellier, France, (4)ADNid company, Montpellier, France

Dear biologist, have you ever dreamed of using the whole power of those numerous NGS tools that your bioinformatician colleagues uses through this awful list of command lines ?

Dear bioinformatician, have you ever wished for a really quick way of designing a new NGS pipeline without having to retype again dozens of code lines to readapt your scripts or starting from scratch ?

So, be happy ! TOGGLE is for you !

With TOGGLE (TOolbox for Generic nGs anaLysEs), you can create your own pipeline through an easy and user-friendly approach. Indeed, TOGGLE integrate a large set of NGS softwares and utilities to easily design pipelines able to handle hundreds of samples. The pipelines can start from Fastq (plain or compressed), SAM, BAM or VCF (plain or compressed) files, with parallel (by sample) and global analyses (by multi samples). Moreover, TOGGLE offers an easy way to configure your pipeline with a single configuration file:

– Organizing the different steps of workflow,
– Setting the parameters for the different softwares,
– Managing storage space through compressing/deleting intermediate data,
– Determining the way the jobs are managed (serial or parallel jobs through scheduler SGE, SLURM and LSF).

TOGGLE can work on your laptop, on a single machine server as well as on a HPC system, as a local instance or in a Docker machine. The only limit will be your available space on the storage system, not the amount of samples to be treated or the number of steps. TOGGLE was used on different organisms, from a single sample to more than one hundred at a time, in RNAseq, DNAreseq/SNP discovery and GBS analyses.

List of bioinformatics tools included:

– Cleaning and Quality checking : FastQC, CutAdapt, fastxTrimmer
– Assembly : TransAbyss, Trinity, TGI-CL
– Mapping : BWA (aln/sampe and mem), TopHat
– SAM/BAM management : picardTools, SAMtools, GATK
– SNP calling/cleaning/annotation : SAMtools, GATK, VarScan, snpEff
– ReadCount : HTseq-count
– Structural Variations : BreakDancer, Pindel

Web site: https://github.com/SouthGreenPlatform/TOGGLE

C28: Computer Demo 4

Bgee: Database and R Package for Retrieving the Preferred Anatomical Expression Localization of a List of Genes, or of a Single Gene, in Animals.

Frederic B. Bastian, SIB Swiss Institute of Bioinformatics - University of Lausanne, Lausanne, Switzerland

A wealth of gene expression data is deposited in public repositories, but the problem is how to extract, from this wealth of data, the biologically relevant information about expression patterns of genes. We present here Bgee, a database and R package allowing to identify the preferred anatomical localizations with expression from a list of genes, or for a single gene. The R package also allows to download curated and re-analyzed expression data from Bgee directly into R.

Bgee is a database to retrieve and compare gene expression patterns in multiple animal species. It is based exclusively on curated "normal" healthy expression data, and produces calls of presence/absence of gene expression, in anatomy, life cycles and developmental stages, to allow comparisons between species. Bgee release 14 includes 29 species.
In this demo I will introduce the use of an innovative tool, allowing to discover in which anatomical structures genes from a given set are preferentially expressed: TopAnat. TopAnat analyses are similar to Gene Ontology enrichment tests, but applied to terms from an anatomical ontology, with genes mapped to anatomical terms using the expression calls from Bgee. We hope that TopAnat will prove to be as useful as standard GO enrichment analyses. I will also present how to retrieve from Bgee gene expression ranks, allowing to identify the most relevant anatomical information for each gene. Finally, the Bgee R package also allows to retrieve into R data from thousands of wild-type healthy samples of multiple animal species, curated and re-analyzed in Bgee.

C29: Computer Demo 4

Using the Wheat Tilling Resource to Find Mutations of Interest

Hans Vasquez-Gross[1], Ksenia V Krasileva[2], Tyson R. Howell[1], Paul C. Bailey[3], Francine Paraiso[1], James Simmonds[4], Ricardo H. Ramirez-Gonzalez[3], Xiaodong Wang[1], Christine Fosker[3], Sarah Ayling[3], Andy L Phillips[5], Cristobal Uauy[6] and Jorge Dubcovsky[7], (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 Center, Norwich, United Kingdom, (5)Rothamsted Research, Harpenden, United Kingdom, (6)John Innes Centre, Norwich, United Kingdom, (7)University of California, Davis, Davis, CA

The exome sequencing of a population comprised of 1500 EMS mutagenized tetraploid wheat lines (cv. Kronos) and 1200 mutagenized hexaploid wheat lines (cv. Cadenza) for in silico targeting of induced local legions in genomes (TILLING) provides an invaluable reverse genetics tool for wheat researchers. With a high probability to knockout a given gene, there are many lines with knockouts in targeted genes of interest. During the session, we will cover finding a gene of interest through our BLAST tool, finding potential high quality hits, then visualizing the results on a reference genome browser (JBrowse), and submitting an order form to request the seeds. For each annotated gene, we have calculated mutation effects for synonymous and intergenic changes, but also non-synonoymous, stop-gained, stop-loss, and splice site mutations. This resource will be publicly available from the UCDavis - Dubcovsky Lab at: http://dubcovskylab.ucdavis.edu/wheat-tilling

C30: Computer Demo 4

Using the Public Plant Breeding API (BrAPI) to Access Data from the Statistical Platform R

Reinhard Simon[1], Nicolas Morales[2], Maikel Verouden[3] and Lukas Mueller[2], (1)International Potato Center (CIP), Lima, Peru, (2)Boyce Thompson Institute, Ithaca, NY, (3)Biometris, Wageningen Plant Research, Wageningen, Netherlands

Several databases exist that store data from plant breeding experiments. To facilitate data access and data re-use a common Breeding Application Programming Interface (BrAPI) has been agreed between database developers from an international community representing the public and private sector. It is based on the principles of open linked data and a web-based architecture. This allows others to build additional tools including tools on platforms like R: A language and environment for Statistical Computing. Here we present a new R package called ‘brapi’ that retrieves data via the BrAPI interface. The main advantages are ease of use for data analysts, statisticians or breeders since the package converts the data into R objects and simplifies access for the R user. The R functions are largely named to be consistent with the BrAPI base calls including parameter names. The R functions return objects that are designed for easy further processing while retaining information about their origin to assist in tracking different data versions. The package has additional features to integrate with the popular RStudio: Integrated Development Environment for R to optimize workflows. We show how to connect to different databases and use the obtained data in subsequent R steps and visualization tools. The use of standards and protocols like the BrAPI increases the reusability of breeding data, contributing opportunities for more efficient breeding programs.