C01: Digital Tools and Resources Session 1

The Camelid Genome Variation Database: An Integrative Database for Studying Camelid Genetic Evolution, Biological Characteristics and Desert Adaption

Tuya Saren[1], Haiqing Liu[1] and Rimutu Ji[1,2], (1)Inner Mongolia Institute of Camel Research, Alxa, Inner Mongolia, China, (2)Inner Mongolia Agricultural University, Huhhot, China

Aims: The camelids including Bactrian camel (*Camelus bactrianus*), dromedary (*Camelus dromedaries*) and alpaca (*Vicugna pacos*), are important economically livestock for transportation and production (meat, milk and wool). During the long-term natural selection, they have acquired many unique abilities to adapt to the desert environment. In recent decades, their genetic diversity, evolutionary history and biological characteristics have been elucidated by independent studies. However, these studies of camelids were scattered and the amount of available data were still limited. Our database Camel GVD is designed to provide comprehensive resources for camelid genome studies. **Methods:** Camel GVD is a web-based large-scale whole genome variation database, including single nucleotide polymorphisms (SNPs), insertions and deletions (Indels). **Results:** It contains a collection of ~19 million variants identified by whole-genome sequencing of over 100 camels from our original research. It also contains all the public available camelid genome data. Camel GVD integrates a comprehensive collection of variation related information, including variation annotations, associated genes/proteins/transcripts, sample locations, population introductions, and allele frequencies. It offers several functions, for users to search, display and retrieve the variations and their annotations. Camel GVD also provides several data browsing functionalities under the "Browse" pull-down menu, including variant alignment, coverage and genome feature files. Users can select one sample or several samples to browse the information. It also provides a special function to display gene information of several camel features based on the existing studies to provide convenience for users. In addition, the vcf files can be downloaded for advanced data mining and analysis. Thus, Camel GVD is expected to become a bioinformatics platform for studying the evolution, genetic breeding, biological characteristics and functional genes of camelids. **Conclusion:** Our database Camel GVD is the first large-scale, and so far the most comprehensive collection of genomic variant data from over 100 individual camelids. Compared with one similar database, our database provides both SNPs and Indels information, and also provides functional gene information associated with camel genetic and biological characteristics. This database is helpful to facilitate the future studies of population genetics and molecular biology in camelids.

Key words: Bactrian camel; Dromedary; Alpaca; Camel genome variation databse;

The work was supported by International S & T Cooperation Program of China (2015DFR30680 and ky201401002); and Chinese National Natural Science Foundation project (31360397).

C02: Digital Tools and Resources Session 1

An Integrated Information System Dedicated to Oak Genomics and Genetics

Joelle Amselem[1], Nicolas Francillonne[1], Célia Michotey[1], Thomas Letellier[1], Jean Marc Aury[2], Corinne DaSilva[2], Sébastien Duplessis[3], François Ehrenmann[4], Sébastien Faye[2], Christine Gaspin[5], Christophe Klopp[5], Karine Labadie[2], Isabelle Lesur[4], Thibault Leroy[4], Florent Murat[6], Olivier Rué[7], Catherine Bodenes[4], Jean-Charles Leple[4], Grégoire Le Provost[4], Patricia Faivre Rampant[8], Antoine Kremer[9], Francis Martin[3], Hadi Quesneville[1], Jérôme Salse[6] and Christophe Plomion[9], (1)URGI, INRA, Université Paris-Saclay, Versailles, France, (2)CEA - Genoscope, Evry, France, (3)IAM, INRA, Université de Lorraine, Champenoux, France, (4)BIOGECO, INRA, Université Bordeaux, Cestas, France, (5)MIAT, INRA, Université de Toulouse, Castanet-Tolosan, France, (6)GDEC, INRA, Université Clermont II Blaise Pascal, Clermont-Ferrand, France, (7)MaIAGE, INRA, Université Paris-Saclay, Jouy-en-Josas, France, (8)URGV, INRA, Université Paris-Saclay, Evry, France, (9)INRA, Université de Bordeaux, Cestas, France

GnpIS is an information system designed to integrate and link genomic, genetic and environmental data into a single environment dedicated to plant (crops and forest trees) and fungi data. GnpIS is regularly improved with new functionalities answering specific needs raised by scientists and released several times a year. We propose to illustrate the integrated genome annotation system we set up with a focus on the interoperability between genomic and genetic data (e.g. Markers, QTL) present in GnpIS-core, through the use case *Quercus robur* (the pedunculate oak), a large, complex and highly heterozygous genome.

This genome annotation system relies on GMOD interfaces such as WebApollo/JBrowse and Intermine to make these data available under a user-friendly environment. All annotations and analysis results (Transposable Elements (TEs), genes, ncRNA ...) and functional annotation (protein-coding genes) were obtained using powerful and robust pipelines: (i) REPET used to detect, classify and annotate TEs representing 50% of the genome; (ii) Eugene which integrates *ab initio* and similarity gene finding softwares to predict gene models; (iii) ncRNA were annotated using different tools to annotate lncRNA, miRNA, rRNA, tRNA (iv) A functional annotation pipeline mainly based on Interproscan and comparative genomics was performed on the 25,808 highly confident predicted proteins. This system allows experts to analyze their protein families of interest and curate/validate gene structure.

All together these resources provide a framework to study the two key evolutionary processes that explain the remarkable diversity found within the Quercus genus: local adaptation and speciation.

C03: Digital Tools and Resources Session 1

POTAGE: A Visualisation Tool for Speeding up Gene Discovery in Wheat

Radosław Suchecki, Nathan S. Watson-Haigh and Ute Baumann, The University of Adelaide, Urrbrae, Australia

POPSEQ Ordered *Triticum aestivum* Gene Expression (POTAGE) is a web application which accelerates the process of identifying candidate genes for quantitative trait loci (QTL) in hexaploid wheat. This is achieved by leveraging several of the most commonly used data sets in wheat research. These include the Chromosome Survey Sequences, their order along the chromosomes determined by the population sequencing (POPSEQ) approach, the gene predictions and RNA-Seq expression data. POTAGE aggregates those data sets and provides an intuitive interface for biologists to explore the expression of the predicted genes and their functional annotation in a chromosomal context. The interface accelerates some of the laborious and repetitive tasks commonly undertaken in the process of identifying and prioritising genes which may underlie QTL. We illustrate the utility of POTAGE by showing how a short-list of candidate genes can quickly be identified for a QTL linked to pre-harvest sprouting - a major cause of quality and yield loss in wheat production. The candidate genes identified using POTAGE included

TaMKK3, which was recently reported as a causal gene for seed dormancy in wheat, and a mutation in its barley ortholog has been shown to reduce pre-harvest sprouting. In addition to the public version of POTAGE, we have also developed a Docker image for quickly deploying POTAGE locally and work-flows showing how to add your own expression data sets to a local installation. This is of particular relevance to those who work with unpublished data sets or would like to deploy POTAGE on their own hardware.

C04: Digital Tools and Resources Session 1

The AraGWAS Catalog: A Curated and Standardized *Arabidopsis thaliana* GWAS Catalog

Arthur Korte, Center for Computational and Theoretical Biology, University Wuerzburg, Wuerzburg, Germany

The abundance of high-quality genotype and phenotype data for the model organism *Arabidopsis thaliana* enables scientists to study the genetic architecture of many complex traits at an unprecedented level of detail using genome-wide association studies (GWAS). GWAS have been a great success in *A. Thaliana* and many SNP-trait associations have been published. With the AraGWAS Catalog (https://aragwas.1001genomes.org) we provide a publicly available, manually curated and standardized GWAS catalog for all publicly available phenotypes from the central *A. Thaliana* phenotype repository, AraPheno (https://arapheno.1001genomes.org). All GWAS have been recomputed on the latest imputed genotype release of the 1001 Genomes Consortium using a standardized GWAS pipeline to ensure comparability between results. The catalog includes currently 167 phenotypes and thousands of significant SNP-trait associations.

C05: Digital Tools and Resources Session 1

GrainGenes: New Content, New Tools, New Tutorials

Victoria Carollo Blake, USDA ARS WRRC, Albany, CA, Sarah Odell, University of California, Davis, Davis, CA, Gerard R. Lazo, USDA Agricultural Research Service, WRRC, Albany, CA, Margaret Woodhouse, ISU, Ames, IA, David Hane, USDA-ARS-WRRC, Albany, CA and Taner Z. Sen, USDA -ARS / GrainGenes, Albany, CA

GrainGenes (https://graingenes.org; https://wheat.pw.usda.gov) is the USDA-ARS database for wheat, barley, oat, and rye genetics and genomics. The GrainGenes project is moving toward a genome-centric resource to accommodate the 'big data' now available for the Triticeae and Avena. In this demo, we will 1) demonstrate the use of the new genome browsers on GrainGenes; 2) describe the variety-specific BLAST databases; 3) review the wealth of new content; and 4) share the collection of recently created topic-specific tutorials. Collaborations with The Triticeae Toolbox (T3), WheatIS, and Agriculture and Agri-Food Canada (AAFC) will assure that GrainGenes remains an important resource for the small grains research community. Mutual projects with our collaborators and future directions for the GrainGenes project will be discussed.

C06: Digital Tools and Resources Session 1

CottonGen: An Online Resource for the Cotton Community

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

CottonGen (www.cottongen.org) is a well curated, community-oriented informatics resource for cotton researchers that facilitates research discovery and cultivar improvement by curating, integrating, comparing, and maintaining a database that aims to serve as the central data repository for the cotton community. CottonGen not only contains genetic maps, QTLs, germplasm, markers, and genome sequences, but also has tools to view genomes (JBrowse and GBrowse), search DNA and protein sequences (BLAST+), view metabolic pathways (PathwayTools), and to view genetic maps (MapViewer and CMap). CottonGen now also uses a Cotton Trait Ontology for phenotype related data and has updated versions of the CottonGen Reference Transcriptomes (RefTrans), which are assemblies of peer-reviewed, published RNA-Seq and EST datasets for individual *Gossypium* species. CottonGen contains the whole genome sequences and annotations of three cultivated species (AD1, AD2, A2) and one wild species (D5), and the metabolic pathways of AD1 and D5. The CottonGen Breeders Information Management System (BIMS) is in development. BIMS is a convenient tool for management of cotton breeding programs and integrates both public and private breeding data for use with management tools. BIMS will also access data from several larger datasets such as the trial data from the Regional Breeders Testing Network (RBTN), the germplasm characterization data from the US National Cotton Germplasm Collection (NCGC), and the germplasm evaluations from the US The Germplasm Resources Information Network (GRIN), China, and Uzbekistan. CottonGen will continue to support the cotton research community by providing useful research tools and serve as a repository for data. CottonGen is directly supported by Cotton Incorporated, USDA-ARS, the cotton industry and USDA NRSP10.

C07: Digital Tools and Resources Session 1

DaTALbase: A Database for Genomic and Transcriptomic Data Related to TAL Effectors

Alvaro Perez-Quintero[1], Léo Lamy[2], Carlos Zarate[2], Boris Szurek[2] and **Alexis Dereeper**[3], (1)Institut de recherche pour le developpement France-Sud, UMR Interactions–Plantes–Microorganismes–Environnement (IPME), Cirad, Université Montpellier., Montpellier, France, (2)IRD, UMR IPME, Montpellier, France, (3)IRD, UMR IPME, F-34394 Montpellier, France

daTALbase is a curated relational database that integrates TALE-related data including:

- bacterial TALE sequences
- plant promoter sequences
- predicted TALE binding sites
- transcriptomic data of host plants in response to TALE-harboring bacteria
- and other associated data: SNPs, orthologs...

The database can be explored to uncover candidate new susceptibility genes, as well as to study variation in TALE repertories and their corresponding targets.

First instances of the database have been deployed for rice and cassava, and future versions will incorporate data related to *Xanthomonas* pathogens of beans, cabbage, citrus or barley. daTALbase can be accessed at http://bioinfo-web.mpl.ird.fr/cgi-bin2/datalbase/home.cgi

C08: Digital Tools and Resources Session 1

MaizeMine: A Data Mining Warehouse for MaizeGDB

Christine G. Elsik[1,2], Aditi Tayal[1], Deepak R. Unni[1], Hung N. Nguyen[1], Jack Gardiner[1], Justin Le Tourneau[1] and Carson M Andorf[3], (1)Division of Animal Sciences, University of Missouri, Columbia, MO, (2)Division of Plant Sciences, University of Missouri, Columbia, MO, (3)USDA-ARS Corn Insects and Crop Genetics Research Unit, Iowa State University, Ames, IA

MaizeMine (http://maizemine.maizegdb.org), the new data mining warehouse for MaizeGDB, accelerates genomic analysis by enabling researchers without scripting skills to create and export customized annotation datasets merged with their own research data for use in downstream analyses. MaizeMine uses the InterMine data warehousing system to integrate genomic sequences and gene annotations from the B73_RefGen_v3 and B73_RefGen_v4 genome assemblies, Gene Ontology (GO) annotations, protein annotations (UniProt), protein families and domains (InterPro), homologs (Ensembl Compara) and pathways (CornCyc, KEGG, Plant Reactome). MaizeMine also provides database-cross references between genes of the AGPv3.21, AGPv4 and RefSeq gene sets, as well as pre-computed expression levels for all three gene sets based on RNAseq data from the *Zea mays* Gene Expression Atlas (NCBI BioProject PRJNA171684).

MaizeMine provides simple and sophisticated search tools, including a keyword search, built-in template queries with intuitive search menus, and a QueryBuilder tool for creating custom queries. The Genomic Region search tool executes queries based on lists of genome coordinates, and supports both B73_RefGen_v3 and B73_RefGen_v4. The List tool allows users to upload identifiers to create custom lists, perform set operations such as unions and intersections, and execute template queries with lists. When used with gene identifiers, the List tool automatically provides gene set enrichment for GO and pathways, with a choice of statistical parameters and background gene sets. MaizeMine is particularly useful for tracking gene identifiers across gene sets to facilitate meta-analysis. Query results can be downloaded in several formats (tab delimited, GFF3, Fasta, BED, JSON, and XML).

C09: Digital Tools and Resources Session 2

A Statistical Framework for Detecting Mislabeled and Contaminated Samples using Shallow-Depth Sequence Data

Ariel W. Chan, Cornell University, Ithaca, NY

Researchers typically sequence a given individual multiple times, either re-sequencing the same DNA sample (technical replication) or sequencing different DNA samples collected on the same individual (biological replication) or both. Before merging the data from these replicate sequence runs, it is important to verify that no errors, such as DNA contamination or mix-ups, occurred during the data collection pipeline. Methods to detect such errors exist but are often ad hoc and require some combination of genotype calling, imputation, and haplotype phasing, making them unsuitable for error detection in low- to moderate-depth sequence data where such tasks are difficult to perform accurately. Additionally, because most existing methods employ a pairwise-comparison approach for error detection rather than joint analysis of the putative replicates, results may be difficult to interpret. We introduce a new method for error detection suitable for shallow-depth sequence data. Using Bayes Theorem, we calculate the posterior probability distribution over the set of relations describing the putative replicates and infer which of the samples originated from an identical genotypic source. Our method addresses key limitations of existing methods and produced highly accurate results in simulation experiments. Our method is implemented as an R package called BIGRED, which is freely available for download under the terms of the GNU General Public License: https://github.com/ac2278/BIGRED.

C10: Digital Tools and Resources Session 2

Ultrafast Analyses of NGS Datasets: Cleaning & Differential Expression

Aurelie Kapusta[1,2], Steven Flygare[3], Chase Miller[1,2], Yi Qiao[1,2], Gabor Marth[1,2], Edgar J. Hernandez[1,2], Guochun Liao[3], Martin Reese[3], Robert Schlaberg[4,5] and Mark Yandell[1,2], (1)Department of Human Genetics, University of Utah, Salt Lake City, UT, (2)USTAR Center for Genetic Discovery, Salt Lake City, UT, (3)IDbyDNA Inc., San Francisco, CA, (4)Department of Pathology, University of Utah, Salt Lake City, UT, (5)ARUP Institute for Clinical and Experimental Pathology, Salt Lake City, UT

NGS-based metagenomics has the potential to revolutionize science, but computation times have hindered adoption. In response, researchers at the University of Utah, the Centers for Disease Control and Prevention (CDC), ARUP Laboratories, and IDbyDNA Inc. have developed Taxonomer [1,2] – an ultrafast engine for comprehensive metagenomics data analysis and interactive results visualization (https://www.taxonomer.com). Taxonomer is unique in providing integrated nucleotide and protein-based classification, and it is extremely fast: Taxonomer can search every read in an Illumina RNA-seq dataset of ~7 million reads against a database of ~38 million sequences, identifying the most likely organism of origin for every read in less than 10 minutes using 16 threads [1]. Taxonomer thus opens new avenues for research applications. We will highlight two broad ones: RNA transcript profiling and detection of NGS sample contamination. For the latter, we will use examples from human and other organisms to demonstrate how researchers can use Taxonomer to profile metagenomics dataset; identify contamination in NGS datasets; and profile transcript abundances, all in real time. The demonstration will also include a step by step tutorial for building custom databases which enable researchers to customize Taxonomer for their own organisms of interest, specialized applications, and to embed Taxonomer in their bioinformatics workflows. NGS has already revolutionized genetics; now ultrafast metagenomics is about to revolutionize comparative genomics and our demo will provide a preview.

[1] Flygare, Simmon et al. (2016). *Genome Biology*

[2] https://github.com/Yandell-Lab/taxonomer_0.5 (Copyright (c) 2016 IDbyDNA Inc.)

C11: Digital Tools and Resources Session 2

Camoco: Identifying High Priority Candidate Genes from GWAS using Co-Expression Networks

Robert J Schaefer[1], Jean-Michel Michno[2], Elaine M Norton[1], Joseph R Jeffers[3], Owen Hoekenga[4], Brian P. Dilkes[5], Ivan Baxter[6], Molly E McCue[1] and Chad Myers[3], (1)University of Minnesota, St. Paul, MN, (2)Department of Agronomy and Plant Genetics,

University of Minnesota, St. Paul, MN, (3)University of Minnesota, Minneapolis, MN, (4)Cayuga Genetics Consulting Group, NA, NY, (5)Department of Biochemistry, Purdue University, West Lafayette, IN, (6)USDA-ARS/Donald Danforth Plant Science Center, St. Louis, MO

Camoco is a fully featured computational framework for building, analyzing and integrating gene co-expression networks with loci identified in genome wide association studies (GWAS). Hundreds of links between genetic markers (SNPs) and agro-economically important traits have been identified by GWAS. Yet, the causal gene or allele often remains unknown due to many genes being in linkage disequilibrium (LD) with each of potentially dozens of genetic markers. Co-expression networks identify genes that share similar response patterns of gene expression making them a powerful tool for inferring the biological function of under-characterized genes. In the right biological context, sets of causal genes related to a GWAS trait will exhibit strong co-expression while inconsequential genes in LD with the marker exhibit random patterns of co-expression.

Camoco features methods to build, analyze, and explore co-expression networks using either microarray or RNA-Seq data. Once built, Camoco establishes a biological context for networks by evaluating their ability to recapitulate previously described ontologies (e.g. GO, KEGG, or MapMan). Vetted networks are then used to determine subsets of genes in close proximity to GWAS loci that are strongly co-expressed. GWAS SNPs are mapped to genes using a SNP-to-gene mapping algorithm using user-defined or map-based haplotype windows. High priority candidate genes are identified by evaluating gene-specific co-expression among candidate genes. Demonstrations will be shown using GWAS datasets and co-expression networks generated in both plants and animals. Camoco is free and open source software and available at http://github.com/LinkageIO/Camoco.

C12: Digital Tools and Resources Session 2

coseq: An R/Bioconductor Package for Co-Expression Analyses of RNA-Seq Expression Data

Andrea Rau, UMR GABI, INRA, Université Paris Saclay, Jouy en Josas, France; University of Wisconsin-Milwaukee, Milwaukee, WI, Antoine Godichon-Baggioni, INSA de Rouen, Saint Etienne du Rouvray, France and Cathy Maugis-Rabusseau, INSA de Toulouse, Toulouse Cedex 4, France

Complex studies of transcriptome dynamics are now routinely carried out using RNA sequencing (RNA-seq). A common goal in such studies is to identify groups of co-expressed genes that share similar expression profiles across several treatment conditions, time points, or tissues. These co-expression analyses can in fact serve a double purpose: (1) as an exploratory tool to visualize cluster-specific profile trajectories; and (2) as a hypothesis-generating tool for poorly annotated genes, as co-expression clusters may correspond to genes involved in similar biological processes or that are candidates for co-regulation.

Although a large number of clustering algorithms have been proposed in the past to identify groups of co-expressed genes from microarray data, the question of if and how such methods may be applied to RNA-seq data has only recently been addressed. In this demo, using our R/Bioconductor package coseq (https://bioconductor.org/packages/release/bioc/html/coseq.html), I will illustrate how appropriately chosen data transformations in conjunction with Gaussian mixture models can be used to effectively identify RNA-seq co-expression clusters. Our package coseq provides a rigorous statistical framework for parameter estimation, an objective assessment of the number of clusters present in the data, diagnostic checks on the quality and homogeneity of identified clusters, and a suite of out-of-the-box publication-ready figures. In addition, coseq can integrate seamlessly into standard differential analysis pipelines such as DESeq2 and edgeR.

C13: Digital Tools and Resources Session 2

iDEP: Integrated Differential Expression and Pathway Analysis for RNA-Seq Data

Steven Xijin Ge, South Dakota State University, Brookings, SD

Analysis and interpretation of RNA-Seq data remain a challenge. We aim to develop a user-friendly web application for exploratory data analysis (EDA), differential expression, and pathway analysis. The key idea of iDEP (integrated Differential Expression and Pathway analysis) is to make many powerful R/Bioconductor packages easily accessible by wrapping them under a graphical interface, alongside annotation databases. For EDA, it performs hierarchical clustering, k-means clustering, and principal component analysis. iDEP detects differentially expressed genes using the limma and DESeq2 packages. For a group of co-expressed genes, it identifies enriched gene ontology (GO) terms as well as transcription factor binding motifs in promoters. Pathway analysis can be performed using packages like GAGE, GSEA, PGSEA, or ReactomePA. iDEP can also detect chromosomal gain or loss using the PREDA package. iDEP uses annotation of 69 metazoa and 44 plant genomes in Ensembl for ID mapping and GO functional categorization. Common gene IDs including microarray probe names can be automatically recognized. Pathway information was also compiled from databases like KEGG, Reactome, MSigDB, GSKB, and araPath. As an example, we analyzed an RNA-Seq dataset involving siRNA-mediated Hoxa1 knockdown in lung fibroblasts, and identified the down-regulation of cell-cycle genes, in agreement with previous studies. Our analyses also reveal the possible roles of E2F1 and its target genes, including microRNAs, in blocking G_1/S transition, and the upregulation of genes related to cytokines, lysosome, and neuronal parts. iDEP enables users to conduct in-depth bioinformatics analysis of transcriptomic data through a graphical interface. Freely available at http://ge-lab.org/idep/

C14: Digital Tools and Resources Session 2

Fast Ordered Sampling of DNA Sequence Variants

Anthony Greenberg, Bayesic Research, Ithaca, NY

While the amount of genomic data available is growing quickly, it is matched by increasing power and storage space of consumer-grade computers. Even applications and pipelines that require powerful servers can be quickly tested on desktop or laptop computers if we can generate representative samples from large data sets. Such subsets are also useful for statistical applications, such as repeated re-sampling. A fast and memory-efficient method that preserves the order of the original records would thus find many applications. A sampling method developed for tape drives 30 years ago appears to fit the requirements. I implement a version of this technique for files containing genetic variant genotype data. I test its performance on modern solid-state and spinning hard-drives, and show that it performs well compared to a

simple sampling scheme. I illustrate the utility of the technique by developing a sampling-based method to quickly estimate genome-wide patterns of linkage disequilibrium (LD) decay with between-variant distance. I provide open-source stand-alone software that samples loci from several variant format files, a separate program that performs LD decay estimates, and a C++ library that lets developers incorporate these methods into their own projects. A source code GitHub repository will be available by presentation time.

C15: Digital Tools and Resources Session 2

XMView: A Multiple Alignment XMap Viewer with Genetic Map Integration

Steve Wanamaker, Department of Botany & Plant Sciences, University of California Riverside, Riverside, CA

XMView is a program for drawing complete clusters of Refaligner optical molecules and contigs. With this program it is possible to display two optical maps at the same time, enabling comparisons between the optical maps. A track of genetic map coordinates allows the user to see the linearity of the assembly. Both of these features help the user to discover chimeras in contigs and optical molecules.

C16: Digital Tools and Resources Session 2

Introducing CAFE: Computational Analysis of (gene) Family Evolution.

Carrie Ganote, Fábio H. K. Mendes, Ben Fulton, Robert Henschel and Matthew W. Hahn, Indiana University, Bloomington, IN

Comparison of whole genomes has revealed large and frequent changes in the size of gene families, the result of gene duplication and loss. Comparative genomic analyses allow us to identify large-scale patterns of change and to make inferences regarding the role of natural selection in gene gain and loss. But genome assemblies constructed from these data are often fragmented and incomplete, resulting in annotation errors, especially in the number of genes present in a genome. To make these analyses possible, we have developed a stochastic birth-and-death model for gene family evolution—applied in the software package CAFE—which is robust in the face of less-than-ideal assemblies. Application of this method to data from multiple whole genomes of many groups has revealed remarkable patterns of gene gain and loss, including gene movement among chromosomes (especially sex chromosomes), polymorphic copy-number variants under local selection, and provides novel methods for carrying out genome assembly, to more accurately estimate gene number.

We will describe the application of CAFE to genome sets, and illustrate the conclusions possible from CAFE analysis. The demonstration will use a publicly available VM running CAFE, posted on the Jetstream cloud.

C17: Digital Tools and Resources Session 3

Genome Variation Map: A Repository of Genome Variations for Global Animals and Plants

Shuhui Song, Beijing Institute of Genomics, Beijing, China

The Genome Variation Map (GVM; http://bigd.big.ac.cn/gvm/) is a public data repository of genome variations. As a core resource in the BIG Data Center, Beijing Institute of Genomics, Chinese Academy of Sciences, GVM dedicates to collect, integrate and visualize genome variations for a wide range of animals and plants, accepts submissions of different types of genome variations from all over the world and provides free open access to all publicly available data in support of worldwide research activities. Unlike existing related databases, GVM features integration of a large number of genome variations for a broad diversity of species including human, cultivated plants and domesticated animals. Specifically, the current implementation of GVM not only houses a total of ~4.9 billion variants for 19 species including chicken, dog, goat, human, poplar, rice and tomato, but also incorporates 8,669 individual genotypes and 13,262 manually curated high-quality genotype-to-phenotype associations for non-human species. In addition, GVM provides friendly intuitive web interfaces for data submission, browse, search and visualization. Collectively, GVM serves as an important resource for archiving genomic variation data, helpful for better understanding population genetic diversity and deciphering complex mechanisms associated with different phenotypes.

C18: Digital Tools and Resources Session 3

Creating a Multi-Organism Disease Model Resource at RGD

Mary Shimoyama, Medical College of Wisconsin, Milwaukee, WI

The Rat Genome Database (RGD) was initiated in 1999 to standardize, integrate and present genomic and phenotype data for the laboratory rat. Because of its use as a model for multiple human diseases and the needs of a disease focused community of researchers, RGD has included human and mouse data from its beginning including genes, QTL, ClinVar variants, and functional data such as disease, phenotype, pathway and Gene Ontology annotations. With the rise of precision medicine initiatives, the need for researchers to access and compare genomic and phenotype data for a variety of models ideal for particular disease studies has increased. To accommodate these needs, RGD adapted its data formats, technical infrastructure and data mining and presentation tools to accommodate data from organisms that serve as important models for specific diseases. Each organism has a genome browser and users can access multiple organism data in the InterViewer for protein-protein interactions and the Variant Visualizer. The Object List Generator & Analyzer (OLGA) and the Gene Annotator allow users to create data sets based on genome region, functional information and combinations of function and then analyze data for genome location, variant pathogenicity and functional commonalities. RGD currently includes data for chinchilla, dog, bonobo and 13 lined ground squirrel. RGD's flexible infrastructure will easily accommodate its planned continuous expansion to multiple other disease model organisms.

C19: Digital Tools and Resources Session 3

Efficiency, Design, and Analytical Improvements in CartograTree, a Web-Framework for Association Mapping and Landscape Genomics in Forest Trees

Taylor Falk, Nic Herndon, Emily Grau, Sean Buehler, Peter Richter and Jill L. Wegrzyn, Department of Ecology and Evolutionary Biology, University of Connecticut, Storrs, CT

CartograTree is a web-based framework developed, in conjunction with the TreeGenes project, to create an efficient tool for researchers to display, select, analyze, and document model and non-model trees in conjunction with their associated genotypic and phenotypic metrics. The integration of environmental layers with genomic data associated with these georeferenced trees allows for analyses such as association

mapping and landscape genomics. Significant updates are implemented to increase the usability of CartograTree, including an advanced query interface, plans for updated plotting and selection tools, references to publications sourced from TreeGenes and Dryad repositories, as well as a wide range of environmental and climatic variables. CartograTree is also integrated into the new Tripal Plant PopGen Submit (TPPS) module, which allows detailed collection of data and metadata associated with publications. Since TreeGenes is running on the Tripal 3.0 platform, the recent integration of the web-based next general sequencing toolkit, Galaxy, is integrated to provide direct access to analytical pipelines. In CartograTree, this integrated enables cross-site querying with partner tree databases, and analysis focused on association mapping. CartograTree currently features a subset of the approximately 1,700 forest tree species found in TreeGenes, representing over 50 unique species and more than 48,000 individual tree accessions from TPPS and Dryad publications. The responsive and visually appealing CarograTree framework will help users perform efficient and comprehensive genomics analysis that can be applied to other plant systems.

C20: Digital Tools and Resources Session 3

The Cassava Genome Hub

Anestis Gkanogiannis[1], **Alexis Dereeper**[2], Boris Szurek[3], Carlos Zarate[3], Camilo Lopez[4], L. Augusto Becerra Lopez-Lavalle[5] and Manuel Ruiz[6], (1)CIAT, Cali, Colombia, (2)IRD, UMR IPME, F-34394 Montpellier, France, (3)IRD, UMR IPME, Montpellier, France, (4)National University of Colombia, bogota, Colombia, (5)International Center for Tropical Agriculture, CIAT, Cali, Valle del Cauca, Colombia, (6)CIRAD, UMR AGAP, Montpellier Cedex 5, France

This portal (www.cassavagenome.org) is an integrative genome information system that allows centralized access to genomics and genetics data, and analytical tools to facilitate translational and applied research in cassava. As we did previously for the Banana Genome Hub (Droc et al. 2013) and the Coffee Genome Hub (Dereeper et al. 2015), we opted for GMOD components that are open source, modular, portable and benefiting from a large community support in which we have been involved. We also plugged numerous in-house tools developed by the South Green bioinformatics platform () such as SNiPlay (Dereeper et al. 2015) and Gigwa (Sempéré et al. 2016) for the exploration of the huge amount of genomic variations available in Cassava, notably through diversity or GWAS studies, DiffExDB to access RNA-Seq and differential expression data, as well as new components dedicated to interactions with Cassava pathogens.

C21: Digital Tools and Resources Session 3

Comparative Analysis of Mammal and Angiosperm Phylogenomic Synteny Networks

Tao Zhao, Wageningen University, Wageningen, Netherlands

Comparative phylogenomic synteny (genomic context) analysis holds great promise for the inference of gene and genome evolutionary history. Utilizing the extensive available whole-genome resources, we have built complete microsynteny (local conserved gene order) networks for all genes of 87 mammalian and 107 angiosperms genomes, respectively. Thus, we can directly compare genome dynamics of these two major clades that have evolved and radiated during the last ~170 million years. To interpret the entire synteny network, we exploited network statistical parameters (i.e. average clustering coefficient, retention percentage, cluster sizes) to characterize and quantify various evolutionary features (i.e. conservation vs diversity) of gene families in a phylogenomic context. In addition, we dissected the composition and size distribution of all synteny clusters, which provide intriguing insights into the differing genomic architectures and dynamics of mammals and flowering plants. Sufficient representative genomes for synteny network construction in this study provide us clearer phylogenetic profiling patterns of synteny clusters. We will highlight several representative examples of lineage-specific clusters (i.e. unique genomic changes) that signal potential links between genomic context variation and the evolution of lineage-specific phenotypic traits.

C22: Digital Tools and Resources Session 3

GEAUniversal: A Web-Based Universal Gene Expression Atlas System for Managing, Analyzing and Sharing Large-Scale RNA-Seq-Based Transcriptome Data

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

We successfully developed *GEAUniversal*, which is a Web-based Universal Gene Expression Atlas System for Managing, Analyzing and Sharing Large-scale RNA-seq-based Transcriptome Data. The system is capable of hosting data from multi-species. *GEAUniversal* was implemented using Python, Flask and MySQL. The transcriptomic data are organized in hierarchical fashion, according to species, experiments, treatments and biological samples. *GEAUniversal* provides three core functions: 1) Expression Profile Query - to query the expression levels of genes in user-defined samples. We implemented an on-the-fly data normalization algorithm to enable querying gene expressions across experiments as the RNA-seq data normalization procedure is dependent upon user-selected dataset. 2) Differential Expression (DE) Analysis – to find differentially expressed genes between two samples or treatments using DESeq2 package. The returned differentially expressed genes can be filtered, sorted, and also displayed as bar or line charts. *GEAUniversal* can further identify enriched gene ontology (GO) terms in the list of user-defined genes from previous DE analysis. This function provides valuable insights to further identifying and analyzing biological pathways of the genes of interest. 3) Gene Co-Expression Analysis - to discover genes with similar expression pattern in user selected samples. The *GEAUniversal* also integrates a suite of scripts to simplify the system installation, including metadata and transcriptomic data population. To date, the *GEAUniversal* has been successfully deployed to empower several GEA projects, for example, the Gene Expression Atlas for Cultivated Alfalfa (*Medicago sativa*) at the Diploid Level (CADL) (www.alfalfatoolbox.org/atlasCADL/) and MtSSP-Atlas: A Gene Expression Atlas for Studying the Small Signaling Peptides in *Medicago truncatula* (mtsspdb.noble.org/atlas/).

C23: Digital Tools and Resources Session 3

RepeatExplorer Galaxy Server for in-Depth Characterization of Repetitive Sequences in Next-Generation Sequencing Data

Petr Novak, Pavel Neumann, Nina Hoštáková and Jiri Macas, Biology Centre CAS, Institute of Plant Molecular Biology, Ceske Budejovice, Czech Republic

Repetitive DNA makes up large portions of plant and animal nuclear genomes, yet it remains the least-characterized genome component in most species studied so far. Recent availability of high-throughput sequencing data together with novel bioinformatics tools provide necessary resources for in-depth investigation of genomic repeats and enable large-scale repeat analysis to be run by biologically oriented researchers. Here we present RepeatExplorer Galaxy server (https://repeatexplorer-elixir.cerit-sc.cz), a collection of software tools for in-depth characterization of repetitive elements, which is accessible via web interface. A key component of the server is the computational pipeline using a graph-based sequence clustering algorithm to facilitate *de novo* repeat identification without the need for reference databases of known elements. Since its first release, number of tools for repeat analysis available on public server has grown up. Today RepeatExplorer include several tools which can be used on both short unassembled NGS reads and genome assemblies. New tools include automatic classification of repetitive sequences based on comprehensive database of transposable element protein domains, genome annotation and classification of repetitive elements, and improved identification of tandem repeats using TAREAN pipeline.

References:

Novak, P., Avila Robledillo, L., Koblizkova, A., Vrbova, I., Neumann, P., Macas, J. (2017) – TAREAN: a computational tool for identification and characterization of satellite DNA from unassembled short reads. Nucleic Acids Res. 45(12): e111.
Novak, P., Neumann, P., Pech, J., Steinhaisl, J., Macas, J. (2013) – RepeatExplorer: a Galaxy-based web server for genome-wide characterization of eukaryotic repetitive elements from next generation sequence reads. Bioinformatics 29: 792-793.

C24: Digital Tools and Resources Session 3

MetaOmGraph for 'Omics Data: NoSQL-Enabled Big Data Visualization and Analysis

Eve Syrkin Wurtele, Manhoi Hur and Urminder Singh, Iowa State University, Ames, IA

MetaOmGraph (MOG) is a tool for plotting and analyzing large sets of data while using as little memory as possible. It was designed with transcriptomic data in mind, but is data-type agnostic. Curated compilations of RNA-seq datasets, including from Arabidopsis, maize, yeast and humans, most composed of thousands of samples, are available on our website. Alternately, a researcher can analyze her/his own dataset. Features include: visualizing gene expression patterns; sorting data by any metadata terms; finding groups of genes with common functions; determining which genes have expression patterns most/least correlated to a gene of interest; statistical determinations of significance. We will highlight MOG function using RNA-Seq and metabolomics datasets from maize and yeast.

C25: Digital Tools and Resources Session 4

PiRATE: A Pipeline to Retrieve and Annotate Transposable Elements of Non-Model Organisms

Jérémy Berthelier[1], Nathalie Casse[2], Nicolas Daccord[3], Véronique Jamilloux[4], Bruno Saint-Jean[1] and Grégory Carrier[1], (1)IFREMER/PBA, Nantes, France, (2)Le Mans University/MMS, Le Mans, France, (3)INRA/IRHS, Beaucouzé, France, (4)INRA/URGI, Versailles, France

With the emergence of long reads sequencing, a renewed interest in repetitive sequences is ongoing. Among them, transposable elements (TEs) are mobile DNA sequences which constitute powerful forces of genome evolution. An efficient strategy to conduct a *de novo* TE annotation is the following: TEs are detected from genomic data and automatically classified to construct a TE library, used to conduct the TE annotation. However, tools performing the classification step use databanks of known TEs and are less efficient to classify TEs belonging to poorly studied taxa. As example, we working on the Haptophyte *Tisochrysis lutea* and only 17 TE families are described for this phylum on RepBase (the most used TE database). In comparison 29,404 TE families are listed for the Metazoan. Facing this, we built a new bioinformatics pipeline named PiRATE. We optimized its detection step by using every existing TE detection approaches. The goal is to promote the full-length detection of every TE families, to facilitate their classification. PiRATE was controlled with genomic data of the model plant *A. thaliana* and is able to detect 80% of its TE families. With PiRATE we estimate that the genome of the Haptophyte *T. lutea* is constituted of 20.8% of TEs and that 3.8% represent putative mobile TEs. PiRATE is automated into a stand-alone Galaxy and is available through a virtual machine: http://doi.org/10.17882/51795

C26: Digital Tools and Resources Session 4

Automated Detection of Endogenous Viral Elements in Host Genomes

Jan P Buchman[1], Karina Zile[2], Cody Glickman[3], Kristyna Kupkova[4], Jacob Waldman[5], Mitchell A Ellison[5], Paul G Cantalupo[5] and **Ben Busby**[1], (1)NCBI, Bethesda, MD, (2)NCBI, NLM, NIH, Bethesda, MD, (3)University of Colorado Anschutz Medical Campus, Aurora, CO, (4)Brno University of Technology, Brno, Czech Republic, (5)University of Pittsburgh, Pittsburgh, PA

Endogenous Viral Elements (EVEs) are part of the host genome and allow horizontal transmission within a host, but the underlying evolutionary mechanisms are still unclear. Metagenomic sequencing data sets contain a wealth of information, including sequences from viruses. Such datasets present an opportunity to analyze known EVEs and discover new ones.

We introduce a new method for extending viral contiguous sequences or contigs through the **B**uilding **U**p **D**omains (BUD) algorithm (https://github.com/NCBI-Hackathons/ViruSpy) that identifies virus DNA from sequencing experiments. This methodology differs from current virus discovery tools by iteratively building upon sequences that are known to contain a viral protein domains, and searching for surrounding non-viral protein domains.

We designed EndoVir (https://github.com/NCBI-Hackathons/endovir) to implement BUD in Python3. To reduce the use of temporary files, data is streamed between processes where possible, e.g. we use MagicBLAST to screen the metagenomic datasets as it can read NCBI's SRA format directly. Our implementation analyzes the results from each individual pipeline step and allows real-time parameter adjustment. This approach allows us to adjust for the diversity present in metagenomic data sets.

C27: Digital Tools and Resources Session 4

CRISPRdirect & GGGenome: Web-Based Software for CRISPR-Cas9 Guide RNA Design with Fast and Sensitive Off-Target Searches

Yuki Naito and Hidemasa Bono, Database Center for Life Science, Mishima, Japan
CRISPRdirect (http://crispr.dbcls.jp/) is a simple and functional web-based online software for designing rational CRISPR-Cas9 guide RNAs. The software selects highly specific gRNAs by performing searches against entire genome using GGGenome (http://GGGenome.dbcls.jp/) nucleotide sequence search software and k-mer hash tables. GGGenome quickly searches short sequences from various kind of genomic sequences allowing mismatches and gaps. The query sequences may contain degenerate nucleotide characters (e.g. N, R, Y) which typically appear in PAM sequences. The recent updates of CRISPRdirect and GGGenome include support of more than 350 organisms including plants, animals and fungi. We also consider adding another species if their genomic sequences are publicly available. These tools also provide REST API which is useful for processing large number of sequences in an automated manner. All services of GGGenome and CRISPRdirect web servers are freely available to all users.

C28: Digital Tools and Resources Session 4

Mercator - a Fast Online Tool to Annotate Plant Genomes.

Marie E Bolger, IBG-2, Forschungszentrum Jülich, Jülich, Germany, Rainer Schwacke, forschungszentrum jülich, Juelich, Germany and Bjoern Usadel, Forschungszentrum Jülich GmbH, Jülich, Germany
Genome sequencing has become a relatively standard practice in recent years and has resulted in circa 200 plant genomes already been sequenced. This necessitates that the downstream tools are able quickly and efficiently process these data. Mercator, an online gene annotation tool has recently been upgraded in response to this need. The high quality annotations together with the cluster upgrades enables gene function prediction for whole plant genomes and transcriptomes (30-150k genes) within minutes. This has been further improved by the addition of visualizations which allow users to immediately compare their results with reference genomes. This tool can be found at www.plabipd.de/portal/web/guest/mercator-ii-alpha-version-

C29: Digital Tools and Resources Session 4

CausNet: A Causal Inference Algorithm for Gene Regulatory Network Reconstruction

Xiaohan Kang[1], Faqiang Wu[2], Bruce Hajek[1] and Yoshie Hanzawa[2], (1)University of Illinois at Urbana-Champaign, Urbana, IL, (2)California State University, Northridge, Northridge, CA
High-throughput sequencing has made large-scale transcriptional data under multiple experimental conditions available, yet gene regulatory network (GRN) reconstruction based on time-series data with statistical confidence is still a challenging task, partially due to limited data versus countless possible regulatory interactions, biological and technical variances, and multiple time scales of gene regulations. Reconstruction with confidence levels on the predicted regulatory interactions is essential for subsequent network analysis and experiment design.
We present CausNet, a causal inference algorithm that finds gene regulatory network reconstruction with confidence levels on the regulatory interactions. CausNet takes expression data and design matrix as input, and outputs a GRN with reliability scores that indicate the confidence levels. The algorithm is based on sparse linear regression model and Granger causality, where the former mitigates the limited data issue, and the latter removes indirect interactions. The reliability scores based on the biological replication data points are obtained by perturbation analysis, which is a Gaussian approximation of bootstrapping in statistics. The optimality of CausNet in the regression model makes it especially suitable for the study of a small number of core genes in a GRN.
We demonstrate the usage of CausNet to reconstruct clock gene networks of soybean and show its performance on simulated biologically plausible expression data. CausNet is implemented in Python 3 and is freely available at https://github.com/Veggente/soybean-network.

C30: Digital Tools and Resources Session 4

Dot: An Interactive Dot Plot Viewer for Comparative Genomics

Maria Nattestad[1], Michael C. Schatz[2] and Brett Hannigan[1], (1)DNAnexus, Mountain View, CA, (2)Johns Hopkins University, Baltimore, MD
Advances in long-read sequencing and scaffolding technologies are leading to unprecedented quality and quantity of genome assemblies. Comparing new assemblies to existing genomes of related species is crucial to understanding differences between organisms across the tree of life. The classic method for visualizing genome-genome alignments is the dot plot, which provides an excellent overview of alignments from the perspective of both genomes. However, dot plots have barely changed in the past decade and are still generated from the command-line as static images, limiting detailed investigation.
Here we present Dot, an interactive dot plot viewer that allows genome scientists to visualize genome-genome alignments in order to evaluate new assemblies and perform explorative comparative genomics. Dot enables scientists to explore regions of interest in detail by zooming in and inspecting unique and repetitive alignments. In addition to showing alignments, Dot allows scientists to load annotations for either or both genomes to show additional context, e.g. understanding how sequence differences map to gene differences. This might also allow scientists to explore how known repetitive elements in the reference genome affect assembly quality in specific regions. Dot supports the output of MUMmer, the most commonly used software method for aligning genome assemblies, with the potential to support outputs from future genome-genome alignment algorithms as they emerge. By leveraging D3 and canvas in JavaScript, Dot combines the benefits of interactivity with scalability, enabling scientists to explore large genomes and create publication-quality images. Dot is free, publicly available, and open source.
URL: https://dnanexus.github.io/dot/

C31: Digital Tools and Resources Session 4

Bioinformatic Analysis using Jetstream, a Cloud Computing Environment

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

National Center for Genome Analysis Support (NCGAS) assists researchers in addressing the scientific challenges of understanding and analyzing the wealth of gene sequence information now available. This includes on-boarding biology professionals who lack the necessary computational background to run their analyses on high-performance computing systems. Virtual machines help with the transition to command line use, software installation, and running analysis in the Linux environment (as most high-performance computing clusters). Jetstream (https://jetstream-cloud.org/) is a cloud computing resource that provides access to preconfigured virtual machines, making the transition relatively effortless, flattening the learning curve needed to get results from experiments that otherwise produce an untenable amount of data. Currently, over 14% of all allocations of usage on Jetstream are for biology other than protein folding – the majority of this being some sort of genomic analysis. NCGAS currently hosts over 123 genome analysis and bioinformatics software titles on Jetstream as preconfigured virtual machines. In this digital tool and resources workshop, we will demonstrate on how to set up Jetstream accounts, start a preconfigured virtual machine, and run genomic analysis on this virtual machine (https://ncgas.org/Blog_Posts/Getting%20Started%20on%20Jetstream.php). Jetstream also provides environments for prototyping and publishing tailored workflows that gives researchers access to interactive computing and data analysis resources on demand. optimality of CausNet in the regression model makes it especially suitable for the study of a small number of core genes in a GRN.
We demonstrate the usage of CausNet to reconstruct clock gene networks of soybean and show its performance on simulated biologically plausible expression data. CausNet is implemented in Python 3 and is freely available at https://github.com/Veggente/soybean-network.

C32: Digital Tools and Resources Session 4

NCGAS makes Robust Transcriptome Assembly Easier with a Readily Usable Workflow Following *de novo* Assembly Best Practices

Sheri A. Sanders[1], Carrie Ganote[2], Bhavya Papudeshi[2], Keithanne Mockaitis[1] 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. Best practices for such analyses include sample pooling, running multiple assembler algorithms with multiple parameters, combining the assemblies, and filtering the redundancy/erroneously assembled transcripts. These 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. NCGAS has created a workflow template to move client data through 19 parallelized assemblies using four software packages (Trinity, SOAP-denovo, transABySS, and VelvetOases) and multiple khmers. 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. While designed for IU machines and XSEDE's Bridges, allocations on these machines are available to any genomics researchers in US and the job scripts can be easily adjusted for other job handlers/clusters. 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. Scripts can be found at https://github.com/NCGAS/IndianaUniversity/tree/master/Transcriptome_Workflow_Mason.