workflows support multiple samples and multiple groups of samples and perform differential analysis between groups in a single workflow job submission. The calculated results are available for download and post-analysis.
The supported animal species include chicken, cow, duck, goat, pig, horse, rabbit, sheep, turkey, as well as several other model organisms including yeast, *C. elegans*, fruitfly, zebrafish, mouse and human, with genomic sequences and annotations obtained from ENSEMBL.
The portal is implemented with many state-of-the-art HPC, workflow and web development software tools including Galaxy, StarCluster, OGS/GE utilizing modern scalable cloud compute and storage sources from AWS.
The RNA-seq portal is freely available from http://weizhongli-lab.org/RNA-seq.

P0490: Bioinformatics: Software

AIR - Artificial Intelligence RNAseq

Riccardo Aiese Cigliano, Andreu Paytuvi Gallart, Ermanno Battista, Fabio Scippacercola and **Walter Sanseverino**, Sequentia Biotech, Barcelona, Spain

In the field of genomics, sequencing technologies have drastically changed in the last few years and the output of complex data generated has outpaced the solutions available for analysis, integration and interpretation. RNA Sequencing has emerged as the number one technique in transcriptomics and thus the solution we propose is based on this. A.I.R.: Artificial Intelligence RNASeq is the first free easy to use SaaS (Software as a Service) built with solid scientific methods. AIR is able to solve three important obstacles in the genomics field simultaneously: the informatics problem (specifically data storage, automatization of results and duration of analysis); the scientific problem (data interpretation and data integration, as well as providing new bioinformatic and statistical functions); the social problem (the lack of availability of skilled bioinformaticians). The overall objective of this project is to introduce a disruptive innovation that will allow researchers to perform transcriptomics data analysis easily, quickly and affordably.

P0491: Bioinformatics: Algorithms

BioBuilds: A Model for Long Term Sustainability of Open Source Bioinformatics

Cheng Lee, John Eargle and **Christopher Mueller**, Lab7 Systems, Austin, TX

Open Source software continues to play an important role in genomics, providing many of the scientific software tools used to analyze sequencing data. As sequencing continues to gain traction in industrial and clinical environments, users face productivity challenges from the effort required to install and maintain Open Source tools, as well as uncertainty about their regulatory acceptability, due in part to the lack of provenance, support, and verification services.BioBuilds, sponsored by Lab7 Systems, IBM, Intel and Continuum Analytics, is a self-contained, turnkey solution for deploying pre-built binaries of common Open Source bioinformatics tools that significantly reduces support and maintenance efforts.
Since its initial release in 2013, BioBuilds has grown into a large, community-driven solution with over 75 packages supported across three hardware architectures (x86, POWER8, and ARM7) on Linux and OS X. BioBuilds uses Continuum's Conda package manager to manage distributions. Users can create and share customized installations containing only needed applications and their dependencies, while still relying on a single source for pre-built binaries, simplifying reproducibility efforts.All BioBuilds packages are RSA-signed and checksummed, making verification and certification of computational environments easier. As with all previous releases, BioBuilds remains free and open. Build recipes for all packages are publicly available on GitHub, and we encourage contributions from the community.
The latest updates continue to support our mission of ensuring that Open Source bioinformatics tools remain readily available and accessible for a broad range of research, industrial, and clinical applications.

P0492: Bioinformatics: Algorithms

BLINK AD: A Multi-Locus Model for Detecting Additive and Dominance Effects in Genome-Wide Association Studies

Yao Zhou[1,2], Meng Huang[2] and Zhiwu Zhang[2], (1)Northeast Agricultural University, Harbin, China, (2)Washington State University, Pullman, WA

During the past decade, genome wide association study(GWAS) has been widely used for mapping genes in nature populations. Linear mixed model(LMM) is widely used to eliminate the false positive caused by population structure. However big data leads to extreme computation challenge, especially when non-additive effects were considered in the model. Currently, only a few methods can perform genome wide additive and dominance(AD) effects association analysis. Our group has newly developed an ultra-fast and powerful multi-locus additive model named BLINK. Here we implemented dominance effect into BLINK model. Simulation study showed that our model has higher power compared with PLINK AD model and Bolt-LMM under different scenarios. In a Finland human population, we analyzed 5 traits with significant dominance effects which were calculated by GREMLd model implemented in GCTA. BLINK AD model could find much more association loci, indicating a potential complex regulation mechanism for those traits. Our results also showed that most of the significant SNPs have complete dominance or over-dominance effects. Furthermore, we performed the prediction analysis using Bayesian LASSO and gBLUP. Results showed that prediction accuracies of Bayesian LASSO using GWAS signifcant loci are higher than gBLUP using whole genome SNPs. We also observed similar results from a rice hybrid population. We anticipate our model would help elucidate the genetic mechanism of complex traits and be useful for genomic selection in plant and animal breeding.

P0493: Bioinformatics: Algorithms

A Fast and Flexible Method for Improving Genomic Prediction with Biological Information

Jicai Jiang[1], Jeffrey R. O'Connell[2], Paul M. VanRaden[3] and Li Ma[1], (1)Department of Animal and Avian Sciences, University of Maryland, College Park, MD, (2)University of Maryland Medical School, Baltimore, MD, (3)Animal Genomics and Improvement Laboratory, ARS-USDA, Beltsville, MD

Genomic prediction has emerged as an effective approach in plant and animal breeding and in precision medicine. Much research has been devoted to achieve higher accuracy in genomic prediction, and one of the potential ways is to incorporate biological information. Due to the statistical and computational challenges, however, a fast and flexible method to incorporate such external information is still lacking. Here, we propose a linear mixed model that can incorporate biological information in a flexible way and develop a fast variational Bayes-based software package. In our proposed model, whole genome markers (m in total) are split into p ($1 \leq p \leq m$) groups in a user-defined manner, and the kth group of markers is given a common effect variance ($k = 1, \ldots, p$). Additionally, each marker has a pre-specified weight for which the rule can be flexibly defined. We provide examples on using the model to incorporate biological information. Previous functional genomics studies have accumulated much evidence on which genes or pathways are more/less important for milk yield in dairy cattle. Based on their levels of importance, we can divide genome-wide markers into a number of groups in our model. We also know rare variants are more likely to be QTLs than common variants for some traits, so we may define a rule, $\omega_j = Beta(MAF_j; \alpha = 1, \beta = 10)$ (beta distribution density function), to pre-specify weight for the jth marker. Our proposed model is implemented with the parameter expanded variational Bayesian method, making it fast and feasible to analyze very large data sets. The software is written in C++ with the Intel MKL library. As a test, we analyzed a large cattle data set consisting of ~20k old bulls (training population) and ~4k young bulls (validation population). All animals have genotyped or imputed ~760k whole-genome sequence genotypes. By grouping markers based on proximity, our software performed better than the industrial implementation of Bayes A in all five milk traits analyzed, with an increase of up to 10% in prediction accuracy and a similar running time. Collectively, the method and software show great potential to increase accuracy in genomic prediction, particularly in the future when more useful biological information is becoming available.

P0494: Bioinformatics: Algorithms

Genetic Study on Clonal F_1 and Double Cross Populations

Luyan Zhang and Jiankang Wang, Institiue of Crop Sciences, Chinese Academy of Agricultural Sciences, Beijing, China

This study focused on genetic analysis methods in clonal F_1 and double crosses. For linkage analysis, molecular markers were classified based on the number of distinguishable alleles at marker locus and the number of distinguishable genotypes in clonal F_1 progenies. Three recombination frequencies were estimated for different scenarios by maximum likelihood algorithm and Newton-Raphson method. Female, male and/or combined maps could be built using the estimates between markers. A combined algorithm of nearest neighbor and Two-opt algorithm of Traveling Salesman Problem was used for marker ordering. Extensive comparisons with software JoinMap4.1, OneMap and R/qtl shows that the new methodology can build more accurate linkage maps in shorter time. We also proposed a procedure to build the four haploids of the two parents. For QTL mapping, it was demonstrated that dominance effect between the female and male parents in clonal F_1 and double cross populations can cause the interactions between markers. We then developed an inclusive linear model that includes marker variables and marker interactions so as to completely control both additive effects and dominance effects. The linear model was finally used for background control in Inclusive Composite Interval Mapping (ICIM) of QTL. The efficiency of ICIM was demonstrated by extensive simulations and by comparisons with simple interval mapping, multiple-QTL models and composite interval mapping. The integrated software called GACD (Genetic Analysis of Clonal F_1 and Double cross populations) was developed for linkage analysis, map construction and QTL mapping in clonal F_1 and double cross populations, freely available from www.isbreeding.net.

P0495: Bioinformatics: Algorithms

EAGLE: Explicit Alternative Genome Likelihood Evaluator

Tony Kuo, Martin C. Frith, Jun Sese and Paul Horton, National Institute of Advanced Industrial Science and Technology, Tokyo, Japan

Reliable detection of genome variations, especially insertions and deletions (indels), from single sample DNA sequencing data remains challenging, partially due to the inherent uncertainty involved in aligning sequencing reads to the reference genome.

Here, we present EAGLE, a method for evaluating the degree to which sequencing data supports a given candidate genome variant. EAGLE incorporates candidate variants into explicit hypotheses about the individual's genome, and then computes the probability of the observed data (the sequencing reads) under each hypothesis. In comparison with methods which rely heavily on a particular alignment of the reads to the reference genome, EAGLE readily accounts for uncertainties that may arise from multi-mapping or local misalignment and uses the entire length of each read.

We compared the scores assigned by several well-known variant callers to EAGLE for the task of ranking true putative variants over false putative variants on both simulated data and real genome sequencing based benchmarks. Overall, EAGLE tended to rank true variants higher than the scores reported by the callers. For indels (<50 bp) EAGLE obtained marked improvement on simulated data and a whole genome sequencing benchmark, and modest but statistically significant improvement on an exome sequencing benchmark.

P0496: Bioinformatics: Algorithms

Impact of the Genetic Diversity of the Reference Population on the Accuracy of Imputation of Rare Variants

Adrien M. Butty[1], Filippo Miglior[1,2], Paul Stothard[3], Flavio Schenkel[1], Birgit Gredler[4], Mehdi Sargolzaei[1,5] and Christine F. Baes[1], (1)Centre for the Genetic Improvement of Livestock, Department of Animal Biosciences, University of Guelph, Guelph, ON, Canada, (2)Canadian Dairy Network & University of Guelph, Guelph, ON, Canada, (3)Livestock Gentec, Department of Agricultural, Food & Nutritional Sciences, University of Alberta, Edmonton, AB, Canada, (4)Qualitas AG, Zug, Switzerland, (5)Unknown, Unknown, ON, Canada

The cost for whole-genome sequencing is still too high for use on large number of animals. SNP genotyping with chips is routinely used and common cattle breeds are widely genotyped. Imputation algorithms have been developed to allow for *in silico* inference of whole-genome sequence (WGS) for all genotyped individuals. High average imputation accuracies are commonly observed. Accurate imputation of rare variants, however, is still difficult and strongly dependent on the composition of the reference population. Up to now most of the sequenced cattle individuals have been selected using the key ancestor approach. This method relies on pedigree or genomic relationships and aims to optimize the genotypic variance of a population explained by the selected animals. In order to reach better imputation accuracies of rare variants in a genotyped population, we have developed a new algorithm for choosing the candidates for sequencing. Following a simulated annealing iteration method, we optimize the Genetic Diversity Index (GDI) of a group of animals that will be the future reference population for imputation. This GDI is computed by summing the count of unique haplotypes at each 20-SNP-window over all animals present in a selection.