Cyberinfrastructure for Life Sciences - *iAnimal* Resources for Genomics and Other Data Driven Biology

J.M. Reecy[1], J. P. Carson[2], F. McCarthy[3], J. E. Koltes[1], E. Fritz-Waters[1], J. Williams[4,5], E. Lyons[4,6], C. F. Baes[7], M. W. Vaughn[2, 4]

[1] Department of Animal Science, Iowa State University, Ames, IA, [2] Texas Advanced Computing Center, University of Texas, Austin, TX, [3] School of Animal and Comparative Biomedical Sciences, University of Arizona, Tucson, Arizona, [4] The iPlant Collaborative, Thomas W. Keating Bioresearch Building, University of Arizona, Tucson, AZ, [5] Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, [6] School of Plant Sciences, University of Arizona, Tucson, AZ, [7] Bern University of Applied Sciences, Switzerland

ABSTRACT: Whole genome sequence, SNPs, copy number variation, phenotypes and other "-omics" data underlie evidence-based estimations of breeding value. Unfortunately, the computational resources (data storage, high-performance computing, analysis pipelines, etc.) that exploit this knowledge are limited in availability – many investigations are therefore restricted to the commercial sector or well-funded academic programs. Cyberinfrastructure developed by the *iPlant Collaborative* (NSF-#DBI0735191) and its extension *iAnimal* (USDA-#2013-67015-21231) provides the animal breeding community a comprehensive and freely available platform for the storage, sharing, and analyses of large datasets – from genomes to phenotype data. *iPlant*/*iAnimal* tools support a variety of genotype-phenotype related analyses in a platform that accommodates every level of user – from breeder to bioinformatician. These tools have been used to develop scalable, accessible versions of common workflows required for applying sequencing to livestock genomics.

Keywords: bioinformatics; breeding; analysis pipeline; high-performance computing; next generation sequencing; variant calling

Introduction

Genomic technologies are now solving previously impossible problems in animal agriculture. High-throughput sequencing and related advances in all areas of information acquisition (phenotypes, climate data, etc.) signal biology's transition to "Big Data" science. Mapping, sequencing and now the re-sequencing of large numbers of individuals—human, chicken, cattle, swine, sheep, etc.—are feasible projects for individual laboratories, rather than the exclusive domain of international collaborations. 1000-fold reductions in sequencing costs (NHGRI (2014)) make it practical for any lab to become their own genome-sequencing center.

In contrast to the increasing accessibility of data, access to computational resources and, perhaps more importantly, broadly-usable interfaces to analysis tools is still lacking; bioinformatics remains a bottleneck (Pérez-Enciso and Ferretti, (2010)). In 2008 The National Science Foundation funded The *iPlant Collaborative* to develop a national cyberinfrastructure for plant sciences. *iPlant* services life science researchers and educators working in all domains of life, enabling them to understand and make increasingly powerful predictions about biological systems. In its first 5 years, *iPlant* successfully developed a platform of integrated technologies and computational resources that provide access to large replicated data storage, high-performance computing, grid computing, and cloud computing. These resources are made available to scientists by providing access at multiple levels including application programming interfaces (APIs), RESTful services, and web-based systems for data access, tool integration, and analysis. *iAnimal* is a natural extension of *iPlant* resources made broadly available to an animal community already concerned with similar sets of biological questions.

This paper presents the *iPlant/iAnimal* vision for solving problems in animal breeding/livestock production, and demonstrates a bovine genotyping pipeline as an example of how *iPlant*/*iAnimal* resources can be leveraged in investigations relevant to animal science. Beyond specific hardware and software, a key part of the cyberinfrastructure that we are developing includes the people concerned both with producing tools suitable for users with various levels of computational background, and with providing training and learning materials that are key to accelerating discovery.

Materials and Methods

iAnimal. Recent advances in biotechnology have permitted animal scientists to sequence all the DNA, or genome, of any organism at relatively little cost. In addition, these technologies are used to understand the activity of genes in a genome and the natural variability in genes between individuals. While such data hold the promise of improving US agriculture by enabling animal breeding and genetics, there exists substantial challenges in transforming all those data into usable knowledge by the widest community of animal scientists and breeders. *iAnimal* is developing an ecosystem of integrated computational resources that leverage prior national investments in cyberinfrastructure (iPlant Collaborative, CoGe, AgBase, and VCMap) to enable agricultural researchers to accelerate their research towards improving US agriculture. We will develop cyberinfrastructure for researchers to manage, analyze, and visualize their quantitative and functional genomics data through an integrated computational platform of existing resources.

Three important aspects of this work will enable scientists to more easily make sense of genomic DNA

source: https://doi.org/10.24451/arbor.7399 | downloaded: 26.3.2020

sequence data. The first is the development of web-based tools for visualizing and interacting with genomic data. These visualization systems are being developed in collaboration with more than a dozen animal research groups working on cattle, pigs, horses, fish, and honeybee. The second is the development of tools to make it easier for researchers to share and reuse data. As more data is generated, it is important for scientists to both quickly share data with collaborators and make it open for other scientists to reuse. The third is the development of cyberinfrastructure training resources for 21st century researchers. Online training videos and written tutorials are being generated, and a series of workshops will be held at scientific meetings frequented by a diversity of livestock and aquaculture scientists. These training materials and workshops are specifically targeting researchers funded by the USDA, as well as young scientists at the postdoc and graduate student levels. An example of how new genomic data is being managed, visualized, kept private, and shared is available at: http://youtu.be/JPZ8IrPnh_8

Infrastructure. The *iPlant/iAnimal* cyberinfrastructure physical resources and computational capacities are both comprehensive and substantial. Data housed in replicate on iRODS (https://www.irods.org/) servers at the University of Arizona and the University of Texas has a current capacity of 500+ TB, and is expansible to 7 PB. Data is backed up by up to 100 PB of SunTek tape storage. On-demand virtualization and cloud computing is provided by iPlant Atmosphere, with the capacity to host 100 to 1000 concurrent 2 GB RAM virtual machines. iPlant currently has access to Stampede and Lonestar at the Texas Advanced Computing Center (TACC), Blacklight at the Pittsburgh Supercomputing Center (PSC), and Trestles at San Diego Supercomputing Center (SDSC). Blacklight is an SGI UV 1000 system powered by 4096 Intel Xeon X7560 cores that can share up to 32 TB RAM, for extremely memory-intensive computing tasks. Trestles, a 100 TFLOP Appro Linux system powered by 10,368 2.4 GHz AMD Magny-Cours processor cores, has a 4 PB parallel file system, and uses Mellanox QDR Infiniband for communications. Lonestar, a 302 TFLOP Dell Linux system, is powered by 22,656 3.33GHz Intel Xeon 5680 CPU cores, a 1+ PB Lustre file system, and a Mellanox QDR Infiniband interconnect. Lonestar features 16 large-memory nodes with 24 cores/node and 1TB of memory, as well as eight dual-NVIDIA M2090 GPU nodes. Stampede (TACC), a Dell Power Edge C8220 system with 9+ PetaFLOPS (PF) peak performance, has over 96,000 Intel Xeon E5-2680 CPU cores and 6400+ Intel Xeon Phi Coprocessors, a 14 PB global parallel file system, and is interconnected via Mellanox FDR Infiniband. In addition to the core HPC cluster, Stampede has 16 large-memory nodes with 32 cores/node and 1TB of memory and 128 nodes containing NVIDIA K20 GPUs. The list of brokered computing systems continues to expand.

Livestock genotyping pipeline. The dramatic growth in capacity and availability of genome sequencing is driving substantial interest in the use of this technology in livestock genomics. However, the algorithms used for sequence alignment, variant calling, variant effect prediction, and imputation analyses are difficult for many scientists to apply efficiently and easily. This is the case for several reasons. First of all, scientific workflows are often composed of multiple software components, each with their own invocation and usage semantics, software dependencies, and data formats—successful adoption of a workflow entails gaining mastery of all components, satisfying all dependencies, and providing data in suitable formats. Next, workflows applicable to the high-throughput genomics space often require access to dozens or even hundreds of computer processors, numerous TB of physical storage, several GB of RAM per processor, and the expertise to apply these resources to the workflow at hand. This is an expensive and technically challenging proposition for most scientists that work on genetics and genomics. Finally, the components of many workflows are only minimally scalable; often, they are written in interpreted languages and operate in serial mode. Occasionally, workflow components are able to take advantage of multiple processors on a single computing node, but are unaware of the rest of the computing cluster. In addition, some workflow components have disproportionately high disk input/output bandwidth requirements for the amount of work being performed. To solve these issues, we have implemented scalable, accessible versions of commonly used workflows required for applying next-generation sequencing to livestock genomics in a user-friendly format to democratize access to genomic selection technology and allow researchers to address important issues facing livestock production.

Results

Our initial demonstration cattle genotyping pipeline included BWA 0.5.9 (Li and Durbin (2009)), SAMtools (Li et al. (2009)), and GATK (McKenna et al. (2010)). We refactored the components of this pipeline for extreme scalability. Specific improvements include: running individual alignments on up to 288 processing cores on TACC Lonestar supercomputer; distributing variant detection across 432 cores; adding basic support for functional annotation of SNPs using SnpEff 3.1h (Cingolani et al. (2012)). Combined, these improvements utilizing *iPlant/iAnimal* resources and Discovery Environment resulted in a 48X improvement in pipeline execution speed, thus allowing the full genotyping of an animal with functional annotation of SNPs in less than 4 hours elapsed time. Using Lonestar, up to 50 concurrent animals can be genotyped at once. Using TACC's newest supercomputer, Stampede, this number increases by a factor of 30. These tools have been used to process over 40 Terabytes of sequence data from the 1000 Bull Genomes and Water Buffalo Genome Consortium projects, using over 300,000 hours of computing. These pipelines are being expanded to allow variant calling on individual as well as pooled animal data. Furthermore, variant calling is being expanded to include Platypus (Rimmer et al. (2012)) and the GATK HaplotypeCaller (Figure 1). The inclusion of additional callers allows for identification of concordance of variants between the callers, increases the confidence of identification of true variants and allows for more complete detection of variants. Currently the pipeline is based around bovine data, but plans include expanding this to other livestock species including sheep, chicken, pig, horse, water buffalo, camel, and bison.

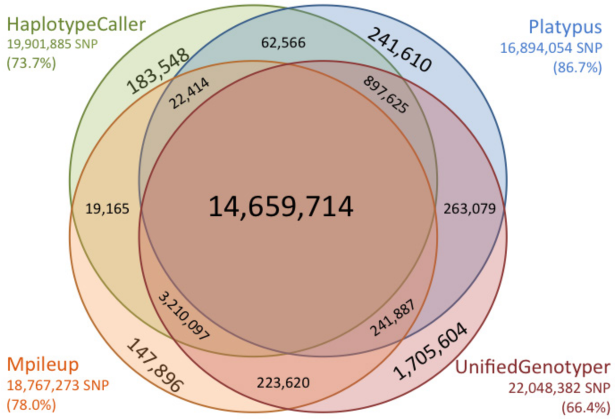

Figure 1. **Concordance of Variants Across Multiple Variant Calling Software.** 14,659,714 variants detected across all software with 89.6% of all variants being detected by 2 or more software.

Conclusion

The *iPlant/iAnimal* cyberinfrastructure is a comprehensive collection of computational capacities. However, it is not solely an issue of computational power to process data that is problematic. Researchers must also learn how to string together multiple programs to process the data. Using *iPlant*/*iAnimal* resources, we provide facilitated computational pipelines that can process large-scale sequence data, enabling agricultural researchers to process data efficiently and effectively. This is especially important because, with the ease to generate genotype/phenotype associations and the associated increase in rate of publications, there is a real need to effectively compare new results with previously published studies.

Literature Cited

Cingolani, P., Platts, A., Wang le, L., et al. (2012) Fly, 6(2):80-92.

Li, H., Durbin, R. (2009). Bioinformatics, 25:1754-60.

Li, H., Handsaker, B., Wysoker, A. et al. (2009). Bioinformatics, 25:2078-9.

McKenna, A., Hanna, M., Banks, E., et al. (2010). Genome Res., 20:1297-303.

NHGRI (Accessed 2/2014). *Website*: https://www.genome.gov/sequencingcosts/

Pérez-Enciso, M., and Ferretti, L. (2010). *Anim. Genet.* 41(6):561-569.

Rimmer A., Mathieson I., Lunter G., McVean G. (2012). Platypus: An Integrated Variant Caller (www.well.ox.ac.uk/platypus).