KBase Performance Metric for FY17: Develop improved open access platforms for computational analysis of large genomic datasets.

Q3 Metric: Report on the latest developments for working with and analyzing plant genomes.

1. Background

The DOE Systems Biology Knowledgebase (KBase) is a collaborative, open environment for systems biology of plants, microbes and their communities. KBase enables researchers to collaboratively generate, test, compare, and share hypotheses about biological functions; perform large-scale analyses on scalable computing infrastructure; and combine experimental evidence and conclusions that lead to accurate models of plant and microbial physiology and community dynamics. In KBase, you can create shareable, reproducible workflows called *Narratives* that include data, analysis steps, results, visualizations, scripts, commentary, and conclusions of an experiment.

KBase was conceived from the beginning as a knowledgebase that would bring together relevant computational systems biology tools and data for microbes and plants and interactions between the two. Traditionally, some scientists have worked on microbes and others have investigated plants, but in fact many of the available analysis tools and datasets are useful for both. KBase's data model integrates reference data and shared user data for both microbes and plants so that results for one organism can be applied to others, helping to accelerate the pace of systems biology research.

2. Overview of plant analysis capabilities in KBase

KBase has a range of analysis tools and data resources that facilitate plant science research by allowing researchers to gain insight into the evolution of genes and genomes, to profile transcriptomes, to perform genome functional modeling with metabolic networks, and to identify differential expression between tissues, developmental stages, environmental conditions and genetic backgrounds. These capabilities are directly relevant to important DOE research targets such as optimizing biomass production in biofuel feedstocks.

The current KBase capabilities provide the foundation for both the project and external developers to build a powerful predictive plant systems biology resource. The ultimate goal is to enable scientists to infer *causal* models that predict plant phenotypes in environmental context from measurements of plant and plant-associated microbial sequence, gene function and expression/activity dynamics and as a function of time, space and condition. A related capability that would simultaneously be enabled is the ability to design interventions--changes of environmental input, microbes and microbial variants, and plant genetic modifications--to optimize desired phenotypes.

To meet these goals, KBase must be able to manage plant and microbial datasets and provide capabilities for functional genomics. In previous reports we have detailed how we have supported the microbial side of this equation. Here we describe tool chains that are required to support the large size

and structural complexity of plant genomes and to map from sequence through gene function to gene expression and expression pattern analysis. As outlined in Figure 1, this phase of the project created tools for upload/download, annotation reconciliation with KBase modeling tools that enable plant and plant/microbe metabolic modeling, basic functional analysis for gene function prediction which links plant gene sequences by homology to other sequences in the system, and gene expression analysis to link dynamics of gene activity to phenotypes.

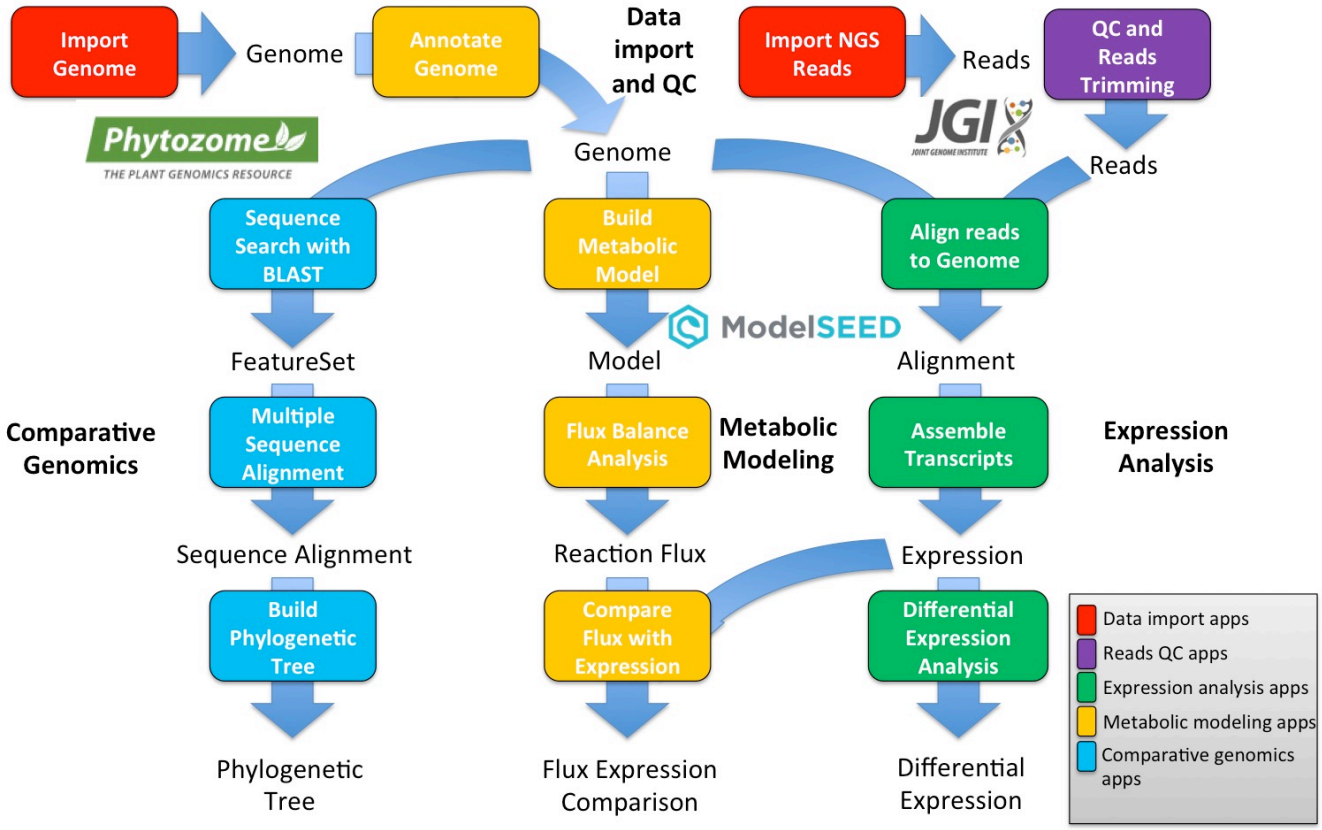

***Figure 1**: The analysis and data import apps that can be used to build plant analysis workflows in KBase.*

Figure 1 shows the current KBase plant analysis toolkit, which includes applications (apps) that enable users to analyze large genomes and next-generation sequencing (NGS) datasets. NGS reads can be trimmed and quality-checked and then assembled and run through RNA-seq pipelines and downstream analysis tools to study differential expression. Genomes can be annotated and then used as input to KBase's metabolic modeling tools, which can reconstruct models for eukaryotes as well as prokaryotes. Another analysis workflow starts with sequence alignment and proceeds to phylogenetic tree reconstruction to elucidate the evolutionary relationships between multiple plant species.

To support this work, we have integrated several key sources of eukaryotic reference data, such as annotated plant genomes from JGI's gold-standard resource, Phytozome [27], to give KBase users immediate access to this valuable community resource and enable them to integrate it with their own data. Integration of these resources into KBase allows users to not only apply KBase analyses to their organism of specific interest, but also to compare the results between different organisms and combine them. KBase also supports user upload of large next-generation sequencing (NGS) reads and genome sequences that can be analyzed together with the public data. Together, these tools--

some of which, such as the plant-microbiome metabolic modeling capability, are not available elsewhere--and data resources bring immediately useful sophisticated analysis capabilities into the hands of KBase users. Combined with the platform infrastructure, these integrated resources significantly lower the bar for future community development and integration of plant systems biology tools and data.

3. KBase resources for analyzing plant data

This section details KBase's plant-relevant data resources and tools as well as listing some example Narratives that demonstrate their use. These annotated, reproducible workflows (which can be found in KBase's Narrative Library, http://kbase.us/narrative-library) can be viewed, copied, and re-run, possibly with different parameters or new input data.

Comparative genomics

Comparative genomics is an area in which the power of KBase is growing. It is through comparison of genomes and genes that the evolution and identify of genes and their functional consequences are often best understood and it is the mechanism by which results from one gene/organism can be transferred to an evolutionarily related one. We are providing a unified set of tools for eukaryotes and prokaryotes that respect the sometimes different approaches needed for the larger and structurally more complex genes and genomes of plants and other eukaryotes.

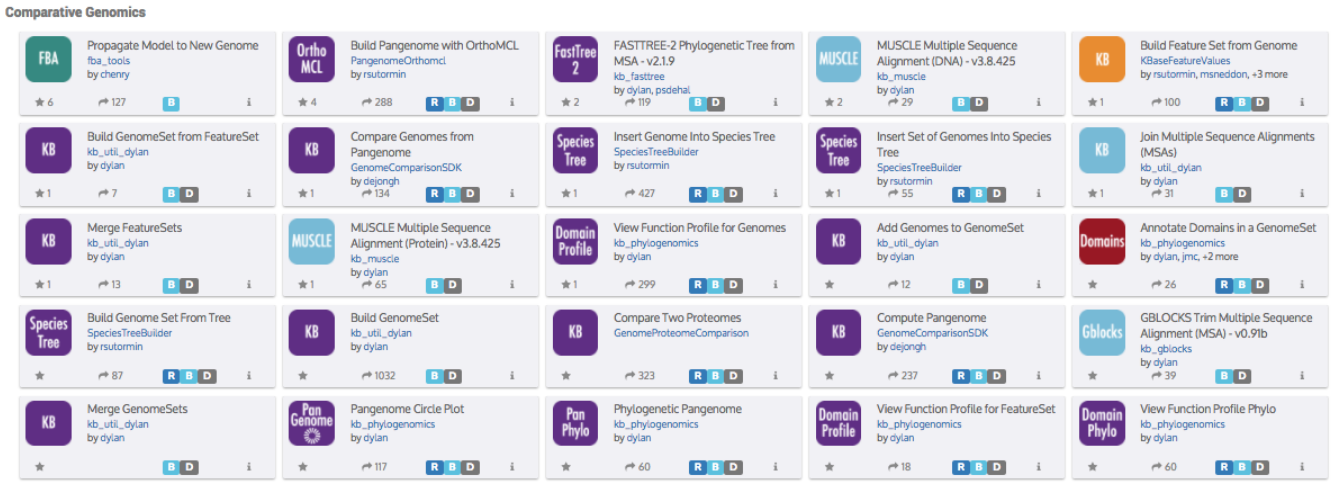

Figure 2: *Screenshot of the comparative genomics tools in the KBase App Catalog.*

KBase has apps in active development (see Figure 2) to enable gene family analysis within and across species, including BLAST-based homology search, multiple sequence alignment and construction of phylogenetic trees. Phylogenetic reconstruction In KBase works in three steps: 1) Generate a feature set (a list of genome references, used by the next steps in the pipeline) 2) Multiple sequence alignment, and 3) Build phylogenetic tree. There are apps for generating feature sets via a BLAST search against genomes (*BLASTp prot-prot Search*) or by searching the gene annotation/identifiers (*Build Feature Set from Genome*). A feature set can be used as input for multiple sequence alignment (*MUSCLE Multiple Sequence Alignment (MSA) - protein seqs*) [22], and in turn, a multiple sequence alignment can be used to build a phylogenetic tree (*FASTTREE-2*

Phylogenetic Tree from MSA) [26]. Most of these apps are currently in beta, as they are still being tested and improved.

The Narrative entitled “Discovery and characterization of ERF protein family members in Arabidopsis thaliana” (https://narrative.kbase.us/narrative/ws.22292.obj.1) demonstrates some of the comparative genomics functionality in KBase. The ERF family of genes encode transcription factors that are involved in various developmental and physiological processes in plants. In this Narrative, the ERF gene family members in the model plant *A. thaliana* are analyzed and characterized. This approach is based on mining homologs of the tobacco AP2 domain containing the ERF2 protein in *A. thaliana* using BLASTP and HMMER searches. Altogether, we were able to find 138 ERF protein family members in *A. thaliana* and build a phylogenetic tree (see Figure 3) based on this gene family.

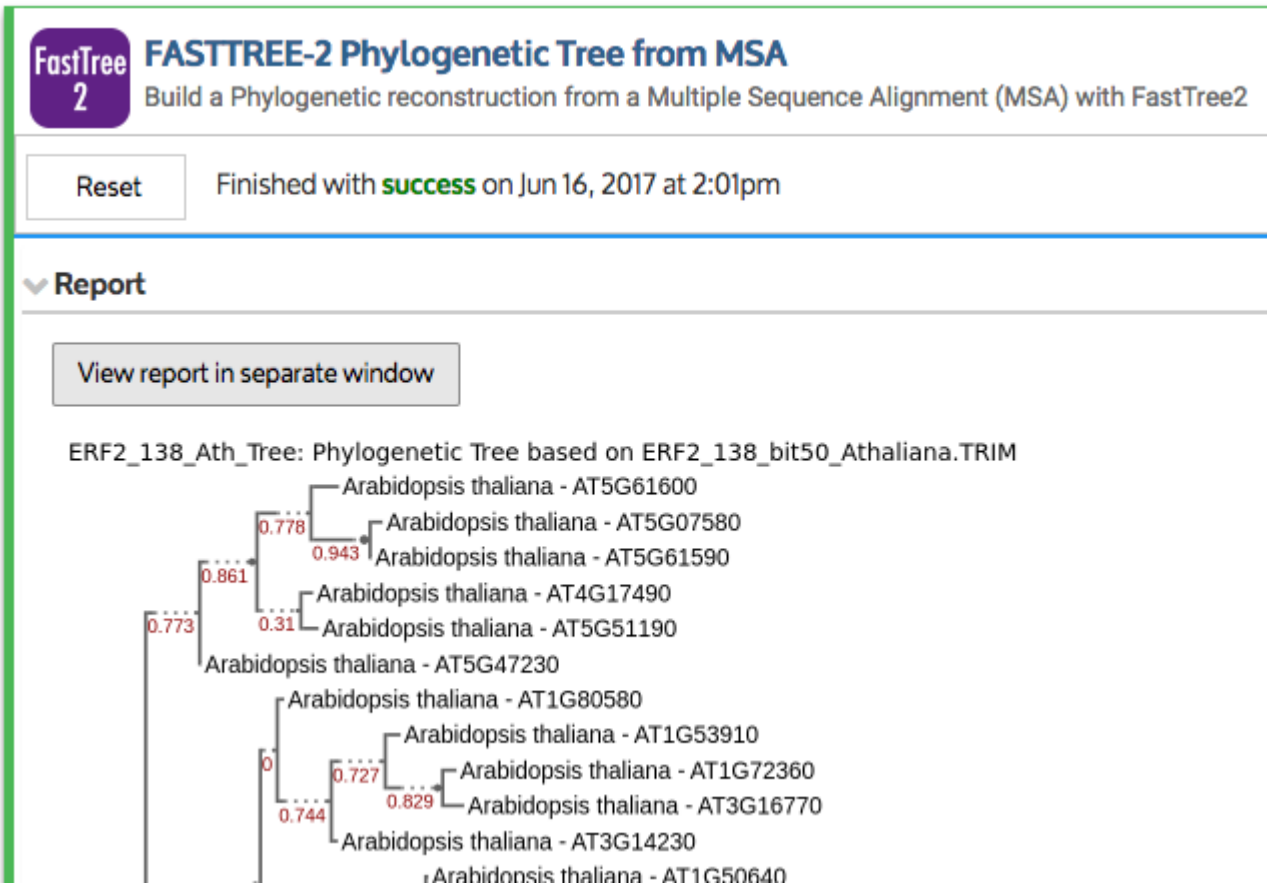

***Figure 3**: Part of the phylogenetic tree for the ERF protein family members in A. thaliana as constructed by the FASTTREE-2 Phylogenetic Tree from MSA app.*

Plant genome annotation and metabolic modeling

Metabolic modeling is a powerful methodology for predicting organism and community growth and production phenotypes from well-annotated genotypes. KBase’s integration with PlantSEED and plant annotation and enzyme-reaction mapping tools enable the creation of metabolic models for plants and plant microbial communities. This toolkit can be used to, for example, predict condition-specific plant

growth and production phenotypes from a plant genotype and that of its microbiome, or design soil amendments or gene modifications to improve these outcomes.

The metabolic modeling tools in KBase can be used to model eukaryotes such as plants, as well as microbes and other prokaryotes. The first step is to use the *Annotate Plant Transcripts with Metabolic Functions* app to assign predicted functions to plant transcript sequences of interest. To generate a reconstruction of plant primary metabolism, run the *Build Metabolic Model* app. To see how the behavior of the metabolic network changes, users can try simulated gene knockouts as part of the FBA runs. They can apply gene expression datasets generated from any environmental conditions to constrain enzyme activity in the FBA simulations. In doing this, scientists are able to combine experimental and model information to observe how the metabolic network changes its behavior from condition to condition, leveraging the expression analysis tools that are discussed in the Expression Analysis section.

Building community models for plants and microbes

One interesting type of analysis that can be performed in KBase is constructing a merged metabolic model for multiple organisms by using the *Build Community Model* app. The ability to merge metabolic reconstructions and model their interactions enables researchers to study interactions between tissues and organs within a single species, as well as cross-species interactions between multiple types of plants, or between plants and microbes. This community modeling capability is particularly useful to plant scientists who may wish to examine how multiple tissues, such as the leaf and the root, interact via the phloem/xylem, or to examine how a microbial pathogen or mutualist may interact with a plant's leaves or roots. By examining the impact of such metabolic interactions, researchers can generate new hypotheses as to what may be happening in the metabolic networks *in vivo*.

In a collaboration between ANL and ORNL [28], scientists explored a mutually beneficial metabolic partnership between a moss and a bacterium, using KBase data and tools to build a merged community metabolic model in which a nitrogen-fixing diazotrophic microbe (*Anabena*) fixes nitrogen to allow a plant (*Sphagnum*) to grow. *Sphagnum* (peat moss) is a genus of plants (Bryophyta) that associate with nitrogen-fixing diazotrophs and are quintessential ecosystem engineers. The Narrative that captures the first part of the analytical workflow used in the study, "Cyanobacteria-Sphagnum Metabolic Interactions: Modeling a Plant/Microbe Interaction" (https://narrative.kbase.us/narrative/ws.9667.obj.2), demonstrates how to find or import microbial and plant genomes in KBase, annotate them with metabolic functions, reconstruct and gapfill metabolic models using appropriately composed media, and finally (using the *Build Community Model* app) merge the two models into a community model which exhibits nitrogen fixation and exchange, showing that the plant portion of the model consumes the nitrogen fixed by the microbial portion of the model. The two-species merged metabolic reconstruction generated in KBase predicts that the *Sphagnum* requires less light when utilizing nitrogen fixed by the microbe than when fixing nitrates on its own.

Expression analysis

RNA-seq analysis is emerging as one of the most powerful approaches for assessing differential gene expression. RNA-seq uses next-generation sequencing to account for all the transcripts in the

biological sample at a particular time, and can be used for a variety of applications such as transcriptome assembly, gene discovery/annotation, and detection of differential transcript abundances between tissues, developmental stages, genetic backgrounds and environmental conditions. It can shed light on questions such as which genes are over- or underexpressed at different stages of development, which genes are expressed differently in a disease state compared with normal cells, or how changing environmental conditions such as temperature changes, drought, or differences in soil chemistry can affect the production of plant biomass. Similarly, samples can be obtained from a group of wild and mutant genotypes to identify the candidate genes responsible for the genetic differences that might explain the observed phenotype differences. The overarching goal of the RNA-seq pipeline in KBase is to create differential expression estimates and use these to inform metabolic models and to perform functional analysis of genes with similar expression patterns.

RNA-seq analysis typically consists of (i) mapping short sequence reads to the reference genome; (ii) assembling the transcripts into full-length transcripts and expression quantification; and (iii) differential analysis of the gene expression. KBase provides a set of apps that allow users to run the tools from the popular Tuxedo RNA-seq suites [5-9] to generate the normalized full and differential expression matrix of the reads obtained from Illumina sequencing platforms using the reference prokaryotic and eukaryotic genome. The RNA-seq apps in KBase can be combined into multiple workflows, allowing users to select their choice of reads aligner and assembler for the differential gene expression analysis (see Figure 4). For alignment of the reads to the reference genome, transcriptome profiling, and identification of differentially expressed genes, the original Tuxedo suite uses the tools TopHat2, Cufflinks, and Cuffdiff [5,6] respectively; the new Tuxedo suite uses HISAT2, StringTie and Ballgown [7,8,9]. All of these tools are available as KBase apps, which are described below; detailed usage instructions can be found on our website at http://kbase.us/transcriptomics-and-expression-analysis/.

There are two Narrative tutorials that demonstrate how to use the KBase RNA-seq pipeline end-to-end on plant reads. You can copy and re-run them to become acquainted with building RNA-seq analysis workflows in KBase.

- Arabidopsis RNA-seq Analysis using Original Tuxedo Suite: https://narrative.kbase.us/narrative/ws.19393.obj.1
- Arabidopsis RNA-seq Analysis using New Tuxedo Suite: https://narrative.kbase.us/narrative/ws.19391.obj.1

The RNA-seq pipeline in KBase has three basic steps, which are described in more detail below.

(i) Aligning reads to the reference genome

KBase has incorporated three different alignment algorithms from the Tuxedo suite for mapping short reads to the reference genome. TopHat2 and HISAT2 are splice aligners and can identify known and novel exon-exon splicing junctions in eukaryotes, whereas Bowtie2 only does unspliced alignment and is preferred for prokaryotic genomes. These are available in KBase as the *Align Reads using Bowtie2* app, the *Align Reads using TopHat2* app, and the *Align Reads using HISAT2* app.

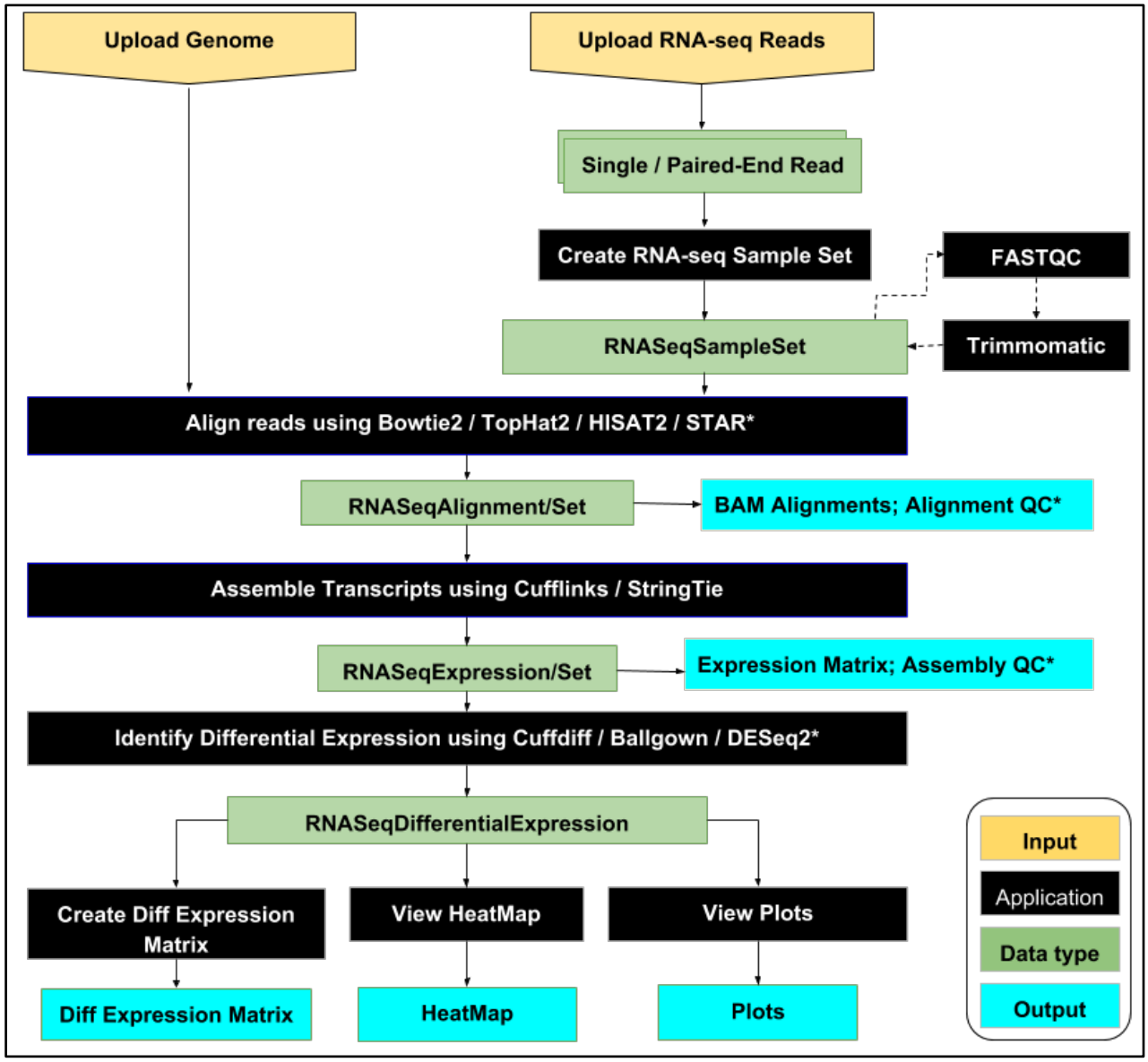

Figure 4: *KBase RNA-seq analysis workflow using original Tuxedo suite (Bowtie2/TopHat2, Cufflinks, Cuffdiff) and new Tuxedo suite (HISAT2, StringTie, Ballgown). The apps marked with * are still in development.*

(ii) Assembling transcripts and quantifying expression

KBase currently has two apps to assemble transcripts for and estimate the gene level abundance: *Assemble Transcripts using Cufflinks* and *Assemble Transcripts using StringTie*. Both Cufflinks and StringTie provide normalized full expression matrices in FPKM (fragments per kilobase of exon model per million mapped reads) and TPM (transcripts per million) format that can be downloaded. The RNA-seq expression object generated by StringTie also provides additional read-count data and a gene-count matrix that are used by Ballgown for detecting differential gene expression.

(iii) Differential analysis of gene expression

KBase provides several differential gene expression analysis tools. The apps *Create Differential Expression Matrix using Cuffdiff* and *Create Differential Expression Matrix using Ballgown* take the

genes and expression levels from Cufflinks and StringTie and apply rigorous statistical methods (q-value and fold change) to determine which genes are differentially expressed between two or more experimental conditions. These apps generate a differential gene expression matrix based on the user-specified threshold cutoff parameters and static plots that help visualize the results. In addition, the *Interactive Volcano Plot* visualization app (see Figure 5) helps users to select appropriate q-value and fold change cutoffs to help fine-tune the threshold cutoff as an input parameter for the differential expression analysis apps.

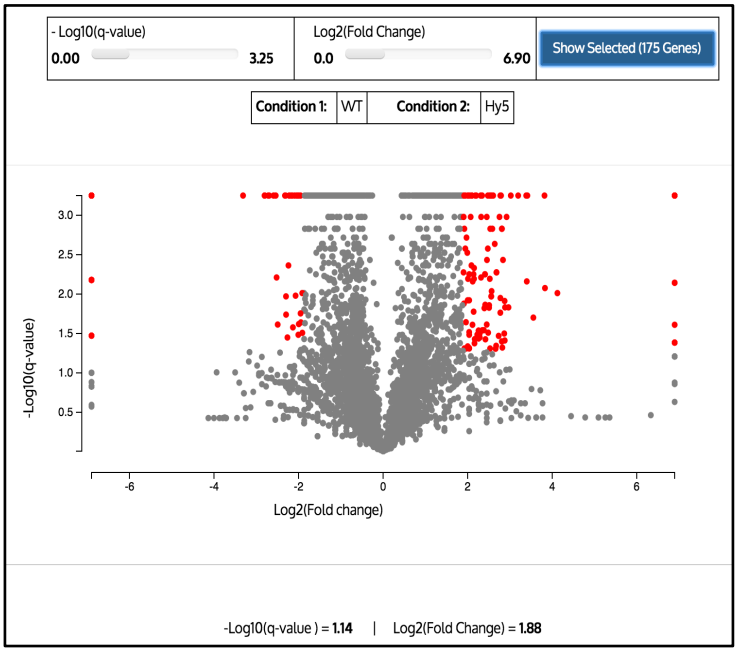

Figure 5*: The Interactive Volcano Plot visualization app allows users to visually study the effect of changing the q-value and fold change cutoffs for differential expression analysis.*

Next-generation sequencing (NGS) reads management and QC

RNA-seq analysis requires the upload, management and quality-checking (QC) of large amounts of next-generation sequencing (NGS) data KBase provides multiple ways to upload NGS reads and perform QC. Users can upload single or paired-end read files to their KBase account from their computer or from an online site with a publicly available URL (FTP, HTTP, Dropbox or Box). KBase provides FastQC [1] for quality check, Trimmomatic [2] for adapter cleaning and removal of poor quality reads, and Cutadapt [3] for efficient management and QC of NGS datasets. Running these apps on reads data helps to improve the accuracy of subsequent analyses.

Data resources in KBase

One of KBase's strengths is that relevant public data is integrated into the platform, so it can be used in analyses along with user-uploaded data. KBase's public reference data currently includes the publicly released non-embargoed annotated plant genomes from the JGI Phytozome database, which are updated on a regular basis. As of June 2017, KBase has 63 plant genomes comprising 41 different plant species--multiple versions of some genomes allow users to compare different genome assemblies and annotations and choose the version they want to analyze. KBase currently has 225

fungal genomes imported from NCBI RefSeq. A collaboration with JGI/Mycocosm will soon make Mycocosm's fungal genomes with high-quality annotations and models available in KBase.

KBase has several data resources that arose out of the PlantSEED project [19]. PlantSEED was independently developed to combine plant comparative genomics, functional annotation of enzymes, and reconstruction of plant primary metabolism for individual species. Plant-specific compounds and reactions, collected from public sources such as KEGG, MetaCyc, and AraCyc, have been integrated into PlantSEED and made available in KBase, where they can be used for plant metabolic modeling. PlantSEED's derived collection of plant k-mers (peptides of 8 amino acids in length) that are unique to each catalytic function underlie the functionality of the *Annotate Plant Transcripts with Metabolic Functions* app, allowing it to assign predicted metabolic functions to the plant sequences.

A Narrative tutorial (https://narrative.kbase.us/narrative/ws.15250.obj.1) demonstrates one way to utilize PlantSEED in KBase. We start with plant coding or protein sequences, annotate the sequences using the *Annotate Plant Coding Sequences with Metabolic Functions* app, and then build a metabolic model of plant primary metabolism using the annotated sequences as the input.

4. Coming soon

KBase's resources for performing plant science are being actively expanded and improved. Tools are in development to make it easier to upload, manage and download large compendia of reads. KBase's access to high-performance computing (HPC) offers opportunities for compute-intensive analyses that would take days or months for most users to run on their own hardware. We are adding new HPC-backed algorithms for large scale assembly of eukaryotic genomes and metagenomes, including an app (currently in beta) that wraps HipMer [2], a fast "extreme-scale" assembler suited for large genomes such as those of plants.

Another active area of development is the expression analysis toolkit, which is being expanded and improved. Several additional expression analysis tools will be available soon in KBase. STAR (Spliced Transcripts Aligner to a Reference) [10] is an ultrafast RNA-seq aligner tool to align reads to genomes; it can map reads of any length and can detect canonical and non-canonical splice junctions and fusion transcripts. DEseq2 [11] is a counts-based differential expression tool that generates an expression matrix based on the total counts of a gene including all of its isoforms. QualiMap [12] provides a comprehensive multi-BAM QC and RNA-seq QC report.

To offer additional biological insight into differentially expressed genes, KBase is working on adding additional tools that will support downstream analyses such as functional enrichment. These tools characterize the molecular functions of differentially expressed genes based on Gene Ontology [14] and SEED [15] terms by comparing a list of differentially expressed genes against the rest of the genome to identify overrepresented functions and applying statistical methods to test for enrichment of each annotated gene set. Several apps for downstream analysis of expression data are already available in a preliminary form, with improvements underway. The expression data generated by *Create Differential Expression Matrix using Cuffdiff* and *Create Differential Expression Matrix using Ballgown* can be used by clustering apps including *Cluster Expression Data – Hierarchical, Cluster Expression Data - K-Means* and *Cluster Expression Data – WGCNA* to enable users to analyze

patterns of gene expression by grouping expression data [16,17,18]. The clusters generated by these apps can be viewed as a heat map using the *View Multi-cluster Heatmap app.* The expression matrices generated by the RNA-seq apps can be fed to the metabolic modeling apps *Compare Flux with Expression* and *View FBA Expression Comparison* to compare reaction flux with gene expression and to identify the pathways where expression and predicted flux agree or conflict.

5. Conclusions

KBase provides a range of tools and datasets that can be used to perform plant science analysis workflows. Many of KBase's apps can be applied to plants and other eukaryotes as well as to prokaryotes; a few are specifically optimized for plant data. The plant-specific analysis apps include uploaders for large eukaryotic genomes and the *Annotate Plant Transcripts with Metabolic Functions* app. Domain-agnostic tools in KBase that can be applied to plant data include NGS reads management and QC apps, two RNA-seq expression analysis suites, apps for comparative genomics and phylogenetic analysis, and a set of metabolic modeling tools that can be applied to plants and plant/microbe interactions. These capabilities can be explored interactively via the exemplar Narratives mentioned in this report.

KBase is actively engaging the external community to help us improve our tools and workflows for plant science, including support for large-scale reads upload and analysis, plant genome annotation, functional genomic clustering and enrichment, physiological modeling and variation and trait-based modeling analysis. We welcome your feedback on our current tools and your suggestions on what new functionality we should add, and we invite you to share your plant science Narratives with the community.

Q1 Report Authors: Priya Ranjan, Sunita Kumari, Sam Seaver, Vivek Kumar, Doreen Ware, Nomi Harris
KBase PIs: Adam Arkin, Robert Cottingham, Chris Henry

REFERENCES

1. Andrews S, FastQC A Quality Control tool for High Throughput Sequence Data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/
2. Bolger, AM, Lohse M, Usadel B (2014) Trimmomatic: A flexible trimmer for Illumina Sequence Data. Bioinformatics, btu170.
3. Martin, M (2012) Cutadapt removes adapter sequences from high-throughput sequencing reads. Bioinformatics in Action, 17 (1),10-12
4. Georganas C et al. (2015) HipMer: an extreme-scale de novo genome assembler. Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis. ACM, 2015.
5. Trapnell C, Pachter L, Salzberg SL. (2009) TopHat: discovering splice junctions with RNA-Seq. Bioinformatics. Vol 25, 9:1105-1111. http://bioinformatics.oxfordjournals.org/content/25/9/1105.abstract
6. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL, Pachter, L (2012) Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nature Protocols, 7(3), 562 578. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3334321/

7. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL (2013) TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biology. 14:R36 http://www.genomebiology.com/2013/14/4/R36/abstract
8. Pertea M, Pertea GM, Antonescu CM, Chang TC, Mendell JT & Salzberg SL (2015) StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nature Biotechnology http://www.nature.com/nbt/journal/v33/n3/full/nbt.3122.html
9. Pertea M, Kim D, Pertea G, Leek JT and Salzberg SL (2016) Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie, and Ballgown. Nature Protocols 11, 1650–1667. http://www.nature.com/nprot/journal/v11/n9/full/nprot.2016.095.html
10. Dobin A, Davis,CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 2013; 29 (1): 15-21. doi: 10.1093/bioinformatics/bts635
11. Love MI, Huber W and Anders S (2014). "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2." *Genome Biology*, **15**, pp. 550. doi: 10.1186/s13059-014-0550-8.
12. Okonechnikov K, Conesa A, García-Alcalde F. Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. *Bioinformatics*. 2016;32(2):292-294. doi:10.1093/bioinformatics/btv566.
13. Robinson JT, Thorvaldsdóttir H, Winckler W, et al. Integrative Genomics Viewer. *Nature biotechnology*. 2011;29(1):24-26. doi:10.1038/nbt.1754.
14. Ashburner et al. Gene ontology: tool for the unification of biology (2000) *Nat Genet* **25(1)**:25-9
15. Overbeek R, Olson R, Pusch GD, Olsen GJ, Davis JJ, et al. (2014) The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). Nucl Acids Res 42: D206 D214. doi:10.1093/nar/gkt1226 http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3965101/
16. Langfelder P, Horvath S (2012) Fast R Functions for Robust Correlations and Hierarchical Clustering. Journal of Statistical Software, 46(11), 1-17. http://www.jstatsoft.org/v46/i11/
17. Lloyd, Stuart P. "Least squares quantization in PCM." Information Theory, IEEE Transactions on 28.2 (1982): 129-137.
18. Langfelder P and Horvath S (2008) WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics, 9:559
19. Seaver SMD et al. (2014) High-throughput comparison, functional annotation, and metabolic modeling of plant genomes using the PlantSEED resource. Proc Natl Acad Sci U S A. 2014 Jul 1;111(26):9645-50. doi: 10.1073/pnas.1401329111.
20. Seaver SMD, Bradbury LMT, Frelin O, Zarecki R, Ruppin E, Hanson AD, et al. Improved evidence-based genome-scale metabolic models for maize leaf, embryo, and endosperm. Frontiers in Plant Science. 2015;6. doi:10.3389/fpls.2015.00142
21. Altschul SF, Madden TL, Sch ffer AA, Zhang J, Zhang Z, Miller W, & Lipman DJ. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25:3389-3402. doi: 10.1093/nar/25.17.3389
22. Edgar, RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32(5):1792-1797. doi:10.1093/nar/gkh340
23. Eddy SR. (2011) Accelerated Profile HMM Searches. PLoS Comp. Biol., 7:e1002195 doi: 10.1371/journal.pcbi.1002195
24. Talavera G1, Castresana J. (2007) Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. Syst Biol. 2007 Aug;56(4):564-77. doi:10.1080/10635150701472164
25. Castresana J. (2000) Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol. 2000 Apr;17(4):540-52. PMID: 10742046
26. Price, MN, Dehal, PS and Arkin, AP (2010) FastTree 2–approximately maximum-likelihood trees for large alignments. *PloS one*, *5*(3), p.e9490.
27. Goodstein DM, Shu S, Howson R, et al. Phytozome: a comparative platform for green plant genomics. Nucleic Acids Res. 2012; 40: D1178-1186.

28. Weston DJ, Timm CM, Walker AP, Gu L, Muchero W, Schmutz J, Shaw AJ, Tuskan GA, Warren JM, Wullschleger SD (2015). Sphagnum physiology in the context of changing climate: emergent influences of genomics, modelling and host–microbiome interactions on understanding ecosystem function. Plant, cell & environment 38(9): 1737-51