Writing a Paper and/while using KBase

B3 program publishes their work – the Capstone

June 2017

Writing a Paper – what are the parts?

- Background and Introduction – why is this problem important, in what context is it important? What was the hypothesis you were testing?
- Materials and Methods – what is the experimental design - justify that it fulfills the requirement to test the hypothesis.
  - Give enough details that someone could repeat the experiment and get the same results (workflow or provenance, references for methods, sources for materials)
- Results – show the results of your quality control steps and your final measurements but do not interpret them. Usually there will be figures and graphs to summarize the data. There might be a
- Discussion – how do you interpret the results? Do they indicate that your hypothesis is correct, incorrect, or partly both with more experiments needed?
- References – there should be no unsubstantiated statements, a reference to an authority you trust should back up every assertion.

- Robert A Day “How to Write and Publish a Scientific Paper”

Journal/Title/Authors

Steiner, K. C. and Carlson, J. E, eds. 2006. Restoration of American Chestnut To Forest Lands - Proceedings of a Conference and Workshop. May 4-6. 2004. The North Carolina Arboretum. Natural Resources Report NPS/NCR/CUE/NRR - 2006/001, National Park Service. Washington, DC.

GENETIC STRUCTURE OF AMERICAN CHESTNUT POPULATIONS BASED ON NEUTRAL DNA MARKERS

Thomas L. Kubisiak and James H. Roberds
USDA Forest Service, Southern Research Station, Southern Institute of Forest Genetics,
23332 Hwy 67, Saucier, MS 39574 USA (tkubisiak@fs.fed.us)

Introduction/Background

INTRODUCTION

The American chestnut (*Castanea dentata* Borkh.) was once one of the most important timber and nut-producing tree species in eastern North America (U.S. Census Bureau 1908). Its native range extended from southern Maine and Ontario in the north to Georgia, Alabama and Mississippi in the south (Sargent 1905). The species now exists primarily as stump sprouts across this entire range, the victim of a devastating canker disease. The disease, chestnut blight, is caused by an exotic fungal pathogen now known systematically as *Cryphonectria parasitica* (Barr 1979). After more than half a century of blight, numerous living stems of American chestnut still exist in the understory of upland forests in the mid-Appalachians (Stephenson et al. 1991). Prolific stump sprouting has enabled American chestnut to persist, but as sexual reproduction is infrequent, its gene pool will likely face serious erosion when old root systems fail to produce sprouts and perish.

Materials and Methods

MATERIALS AND METHODS

Population sampling and DNA extraction

A rangewide sampling of expanded leaves or dormant buds of American chestnut were collected at 22 sites across its natural range (refer to Figure 1). Most of the samples were collected from sites in State or National Forests, but a few sites were located on private land holdings. Each sample was assigned a unique ID and sent to the Southern Institute of Forest Genetics in Saucier, Mississippi for DNA extraction and analysis. Total nucleic acids were isolated from tree tissues as described in Kubisiak et al. (1997).

Results

RESULTS

Putative species identification

Primers that amplified the intergenic spacer region between *trn*T (UGU) and the *trn*L (UAA) 5' exon of the chloroplast genome (primers a and b: 5'-CATTACAAATGCGATGCTCT-3' and 5'-TCTACCGATTTCGCCATATC-3', respectively; Taberlet et al. 1991) were found to uniquely differentiate American chestnut chloroplast DNA from all other *Castanea* (chestnut and chinkapin) species. Based on DNA sequence data (data courtesy F. Dane and P. Lang of Auburn University) this primer pair was found to amplify a band 857 base pairs (bp) in length in American chestnut, and bands ranging from 942 to 945 bp in all other *Castanea* species including the native chinkapin (both *C. pumila* var. *alleghaniensis* and *C. pumila* var. *ozarkensis*). Much of the size difference observed between American chestnut and the other *Castanea* species was due to two unique deletions (one 12 bp and the other 75 bp in length) contained within this region of the American chestnut chloroplast genome. A larger sampling of native chinkapin (specifically *C. pumila*; var. *alleghaniensis* - 48 trees) has yet to show the presence of these large deletions.

Results – Summary Table

Table 1. Microsatellite and RAPD primer sequence, repeat type, allele size, and number of unique alleles identified in samples collected from 18 populations of *Castanea dentata* Borkh. located throughout the species natural range in eastern North America.

Locus	Primer Sequence 5'-3'	Repeat type	Allele size (bp)	Number of unique alleles
Microsatellites				
*Cs*CAT01[a]	F[b]:AGAATGCCCACTTTTGCA R:CTCCCTTATGGTCTCG	$(AC)_nAT(AC)_n$	167-211	31
*Cs*CAT14	F:GAGGTTGTTGTTCATCATTAC R:ATCTCAAGTCAAAAGGTGTC	$(AC)_n$	121-151	15
*Cs*CAT15	F:TCTGCGACCTCGAAACCGA R:CTAGGGTTTTCATTTCTAG	$(AG)_n$	115-141	15
*Qa*CA022	F:AACAATAGGAGTTGGTTTGAG R:GTTAGGGTTTGGAAAATAGGA	$(AC)_n$	160-188	13
*Qa*GA068	F:GCTTTTCTTTCCAGGGCTAC R:GTGGGACAGTGAGGCAGAG	$(AG)_n$	156-192	17
*Qa*GA209	F:CAAGCAGTATTGTTTTATCTC R:GTTGCCCCTGTGAACTAC	$(AG)_n$	227-265	15
RAPDs				
106	CGTCTGCCCG	NA	500	2

Discussion

DISCUSSION

One of our main concerns in this investigation was inclusion of trees that are not pure American chestnut. Inappropriate trees include interspecific hybrids or pure species other than American chestnut, especially the native congener species chinkapin (*Castanea pumila*). Inclusion of such contaminants could have inflated our estimates of genetic diversity, especially in populations containing the non-American chestnut samples, as well as clouded true patterns of genetic variability. Chloroplast DNA sequence variations have been widely used to investigate interspecific relationships among plant species (Palmer et al. 1988, Clegg et al. 1991) because they evolve slowly. We identified a chloroplast-specific marker (primers a and b; Taberlet et al. 1991) that quickly differentiates American chestnut chloroplast DNA from all other *Castanea* species, including the native *C. pumila*. Unfortunately, maternal inheritance of chloroplasts precludes our ability to distinguish interspecific hybrids of maternal American chestnut origin. As a result, our sample set might still contain some interspecific hybrids, however, the number should be small as most collections were made in either State Forests or National Forests where non-native *Castanea* species do not extensively occur.

References

LITERATURE CITED

Barr, M.E. 1979. The diaporthales in North America with emphasis on gnomonia and its segregants. Mycologia Memoir 7, J. Cramer, Lehre, Germany.

Burnham, C.R. 1981. Blight-resistant American chestnut: There's hope. Plant Dis. 65:459-460.

Clapper, R.B. 1954. Chestnut breeding techniques and results. II. Inheritance of characters, breeding for vigor, and mutations. J. Hered. 45:201-208.

Clegg, M.T., G.H. Learn, and E.M. Goldberg. 1991. Molecular evolution of chloroplast DNA. P. 135-149. in Evolution at the molecular level, Selander, R.K., A.G. Clark, and T.S. Whittam (eds.) Sinauer Associates Inc., Sunderland, MA.

Second example (more recent, with bioinformatics)

Complete Plastid Genome Sequences of Three Rosids (*Castanea, Prunus, Theobroma*): Evidence for At Least Two Independent Transfers of *rpl22* to the Nucleus

Robert K. Jansen[1] Christopher Saski,[2] Seung-Bum Lee,[3] Anne K. Hansen,[1] and Henry Daniell[*,3]

[1]Section of Integrative Biology and Institute of Cellular and Molecular Biology, The University of Texas at Austin

[2]Clemson University, Genomics Institute

[3]Department of Molecular Biology and Microbiology, University of Central Florida

***Corresponding author:** E-mail: daniell@mail.ucf.edu.

Associate editor: Charles Delwiche

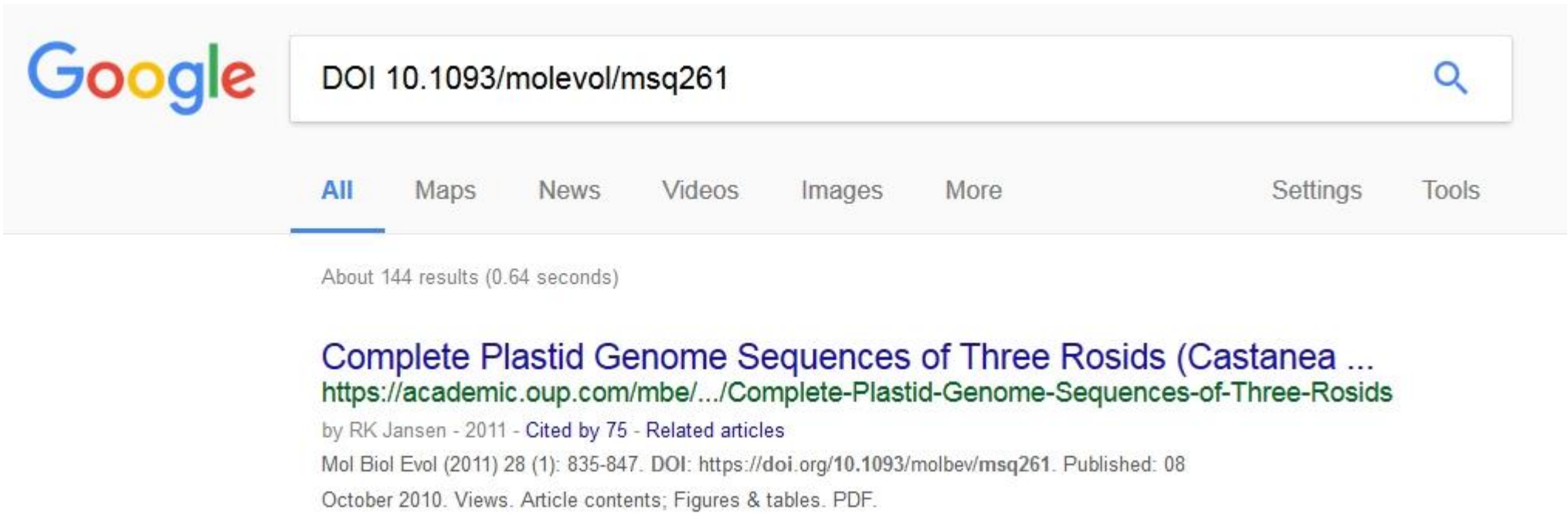
Abstract

Functional gene transfer from the plastid to the nucleus is rare among land plants despite evidence that DNA transfer to the nucleus is relatively frequent. During the course of sequencing plastid genomes from representative species from three rosid genera (*Castanea*, *Prunus*, *Theobroma*) and ongoing projects focusing on the Fagaceae and Passifloraceae, we identified putative losses of *rpl22* in these two angiosperm families. We further characterized *rpl22* from three species of *Passiflora* and one species of *Quercus* and identified sequences that likely represent pseudogenes. In *Castanea* and *Quercus*, both members of the Fagaceae, we identified a nuclear copy of *rpl22*, which consisted of two exons separated by an intron. Exon 1 encodes a transit peptide that likely targets the protein product back to the plastid and exon 2 encodes *rpl22*. We performed phylogenetic analyses of 97 taxa, including 93 angiosperms and four gymnosperm outgroups using alignments of 81 plastid genes to examine the phylogenetic distribution of *rpl22* loss and transfer to the nucleus. Our results indicate that within rosids there have been independent transfers of *rpl22* to the nucleus in Fabaceae and Fagaceae and a putative third transfer in *Passiflora*. The high level of sequence divergence between the transit peptides in Fabaceae and Fagaceae strongly suggest that these represent independent transfers. Furthermore, Blast searches did not identify the "donor" genes of the transit peptides, suggesting a de novo origin. We also performed phylogenetic analyses of *rpl22* for 87 angiosperms and four gymnosperms, including nuclear-encoded copies for five species of Fabaceae and Fagaceae. The resulting trees indicated that the transfer of *rpl22* to the nucleus does not predate the origin of angiosperms as suggested in an earlier study. Using previously published angiosperm divergence time estimates, we suggest that these transfers occurred approximately 56–58, 34–37, and 26–27 Ma for the Fabaceae, Fagaceae, and Passifloraceae, respectively.

Key words: plastid genome, *rpl22*, gene transfer, rosids.

The DOI

Mol. Biol. Evol. 28(1):835–847. 2011 doi:10.1093/molbev/msq261 Advance Access publication October 8, 2010



MOLECULAR BIOLOGY AND EVOLUTION

Volume 28, Issue 1
January 2011

Article Contents

Complete Plastid Genome Sequences of Three Rosids (*Castanea*, *Prunus*, *Theobroma*): Evidence for At Least Two Independent Transfers of *rpl22* to the Nucleus FREE

Robert K. Jansen, Christopher Saski, Seung-Bum Lee, Anne K. Hansen, Henry Daniell ✉

Mol Biol Evol (2011) 28 (1): 835-847. **DOI:** https://doi.org/10.1093/molbev/msq261
Published: 08 October 2010

Abstract

Functional gene transfer from the plastid to the nucleus is rare among land plants despite evidence that DNA transfer to the nucleus is relatively frequent. During the course of sequencing plastid genomes from representative species

Related Records: 54,443
(from Web of Science Core Collection)

For: Complete Plastid Genome Sequences of Three Rosids (Castanea, Prunus, Theobroma): Evidence for At Lea **...More**

Refine Results

Search within results for...

Web of Science Categories

- PLANT SCIENCES (9,880)
- GENETICS HEREDITY (9,827)
- BIOCHEMISTRY MOLECULAR BIOLOGY (9,635)
- EVOLUTIONARY BIOLOGY (9,488)
- MICROBIOLOGY (8,229)

more options / values...

Refine

Document Types

- ARTICLE (50,952)
- REVIEW (2,523)
- PROCEEDINGS PAPER (1,160)
- BOOK CHAPTER (606)
- EDITORIAL MATERIAL (332)

Sort by: Relevance

Select Page | Save to EndNote online | Add to Marked List

1. **The Plastid Genomes of Flowering Plants**

By: Ruhlman, Tracey A.; Jansen, Robert K.
Edited by: Maliga, P
CHLOROPLAST BIOTECHNOLOGY: METHODS AND PROTOCOLS Book Series: Methods in Molecular Biology Volume: 1132 Pages: 3-38 Published: 2014

GET IT! View Abstract

2. **The evolution of the plastid chromosome in land plants: gene content, gene order, gene function**

By: Wicke, Susann; Schneeweiss, Gerald M.; dePamphilis, Claude W.; et al.
Conference: 1st International Symposium on Chloroplast Genomics Engineering Location: Changchun, PEOPLES R CHINA Date: DEC 07-09, 2007
PLANT MOLECULAR BIOLOGY Volume: 76 Issue: 3-5 Special Issue: SI Pages: 273-297 Published: JUL 2011

GET IT! View Abstract

3. **Chloroplast genomes: diversity, evolution, and applications in genetic engineering**

By: Daniell, Henry; Lin, Choun-Sea; Yu, Ming; et al.
GENOME BIOLOGY Volume: 17 Article Number: 134 Published: JUN 23 2016

GET IT! View Abstract

Results

Genome Organization of Three New Rosid Sequences

The three newly sequenced rosid plastid genomes are quite similar to each other in terms of overall organization, gene/intron content, gene order, and GC content (fig. 1, table 1, with accession numbers), and they fall within the typical size range for photosynthetic angiosperm plastid genomes that have not been rearranged (Raubeson and Jansen 2005; Bock 2007). The only exceptional feature is the putative loss of one ribosomal protein gene, *rpl22*, in *Castanea*. There is a pseudogene with 16 internal stop codons remaining in the plastid genome at the correct location within the highly conserved S10 operon (fig. 2). Another

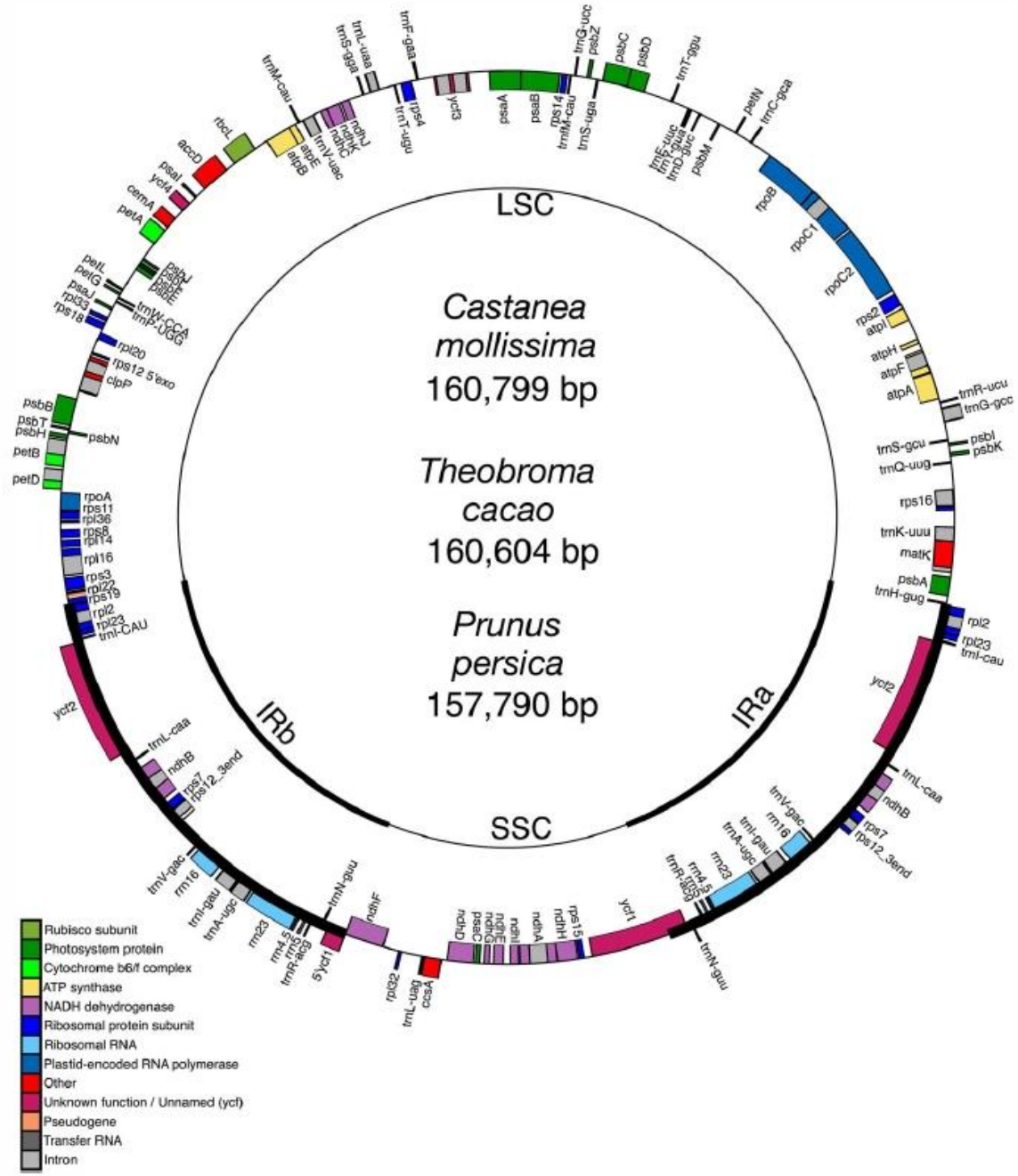


FIG. 1 Circularized gene map of the plastid genomes of three rosids. The thick lines indicate the extent of the inverted repeats (IRa and IRb), which separate the genomes into small (SSC) and large (LSC) single-copy regions. Genes on the outside of map are transcribed in the counterclockwise direction, and genes on the inside of the map are transcribed in the clockwise direction.

Supplementary Material

Supplementary table 1 is available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

Support for this work was provided by grants from National Sciecne Foundation (DEB-0717372 to R.K.J.) and National Institutes of Health GM 63879 and USDA 58-3611-7-610 (to H.D.). We thank Chris Blazier and two anonymous reviewers for critical comments on an earlier version of this manuscript.

B3 Authors

- Tamdan Le
- Kayla Jones
- Caleb Judd
- Melanie Loor
- Kim Kien
- Jennifer Weller
- Paul Sisco
- Steve Barilovits
- Taylor Perkins
- Jeanne Smith
- Erica Putnam

The B3 Chloroplast paper – Background and Introduction

- American chestnut background - Caleb
- Chloroplast genome background - Tamdan
- Context - Choice of chloroplast genome as a focus for the class – why? Melanie
- Other? (discriminating interspecific hybrids? Maternal donor?) Paper – Cytoplasmic male sterility in chestnuts; Cp as marker for the MT. = Jennifer

- Assignments: (a few references and printouts of papers provided as a guideline – be sure not to plagiarize, use quotes as needed, but sparingly)
  - Helen Thompson "The Chestnut Resurrection" Nature (2012) 490,22-23 doi:10.1038/49022a
  - Thomas L. Kubisiak and James H. Roberds "Genetic Structure of American Chestnut Populatons Based on Neutral DNA Markers" (2006) Proceedings of Conference on Restoration of American Chestnut to Forest Lands May 4-6, 2004. Ed KC Steiner and JE Carlson
  - TL Kubisiak et al. "Molecular Mapping of Resistance to Blight in an Interspecific Cross in the Genus Castenea" (1997) Host Genetics and Resistance 87(7) 751- 759
  - Sequencing cucumber (cucumis sativus L.) chloroplast genomes identifies differences between chilling-tolerant and –susceptible cucmber lines By Sang-Min Chug, Vanessa S Gordon and Jack E. Staub, Genome Research (2006), doi:10.1139/G07-003
  - P Lang, F Dane, TL Kubisiak, H Huang "Molecular evidence for an Asian origin and a unique wetward migration of species in the genus Castanea via Europe to North America" (2006). Molecular Phylogenetics and Evolution. 43:49-59 doi:10.1016/j.ympev.2006.07.022
  - J Shaw, JH Craddock, MA Binkley "Phylogeny and Phylogeography of North American Castenea Mill. (Fagaceae) Using cpDNA Suggests Gene Sharing in the Southern Appalachians (Castenea Mill., Fagaceae) CASTENEA 77(2): 186-211 (2012) doi:10.2179/11-033.
  - P Sisco et al., "An Improved Genetic Map for Castenea mollissima/Castenea dentate and its Relationship to the Genetic Map of Castanea sativa" (2005) Proc. IIIrd Intl. Chestnut Congress Acta Hort. 693, ISHS 491 – 496
  - S Wicke et al. "The evolution of the plastid chromosome in land plants: gene content, gene order, gene function (2011) Plant Mol Biol 76:273-297. doi:10.1007/s11103-011-9762-4.
  - AM Ellison et al "Loss of foundation species: consequences for the structure and dynamics of forested ecosystems – a review" Front. Ecol Environ (2005) 3(9): 479-486
  - X Yang et al "Using Next-Generation Sequencing to Explore Genetics and Race in the High School Classroom" (2016) CBE – Life Sciences Education 16:ar22,2

The B3 Chloroplast paper – Materials and Methods

- Sample collection Kayla
- DNA purification and QC - Kim
- PCR primer design, amplification and QC - Jennifer
- Illumina library preparation and QC - Jennifer
- Raw Data = sequence output – Cathy Moore
- Derived Data = analysis – the Kbase environment with workflow and intermediate data sets comprises another set of methods – Steve and Jennifer
- Assignments: (protocols are Web-accessible, you can summarize and provide the link – a couple of the papers also have DNA extraction, PCR and sequencing references).

The B3 Chloroplast paper - Results

- Sequence results
  - Raw data statistics for each chloroplast genome
  - What sorts of images and graphs do we want to provide?
- Bioinformatics results
  - Derived data output for each chloroplast genome – coverage, gaps, frequency per base?
  - What sorts of images do we want?
  - Annotated data – what genes are recognized in each cp genome?
    - 01 – Erica
    - 02 – Caleb
    - 03 – Tamdan
    - 04 – Melanie
    - 05 – Kim
    - 06- Kayla
  - What sorts of images do we want? – graphical workflow?
- Assignments (Papers with examples are provided, discuss more later)
  - J-Y S Yap et al “Complete Chloroplast Genome of the Wollemi Pine (Wollemi nobilis): Structure and Evolution (2015) PLOS One e0128126. doi:10.1371.journal.pone.0128126.
  - L Chaney, R Mangelson, T Ramaraj, EN Jellen, PJ Maughan “The Complete Chloroplast Genome Sequences for Four Amaranthus Species (Amaranthaceae)” (2016) Applications in Plant Sciences 4(9):1600063. doi:10.3732/apps.1600063.

The B3 Chloroplast paper - Discussion

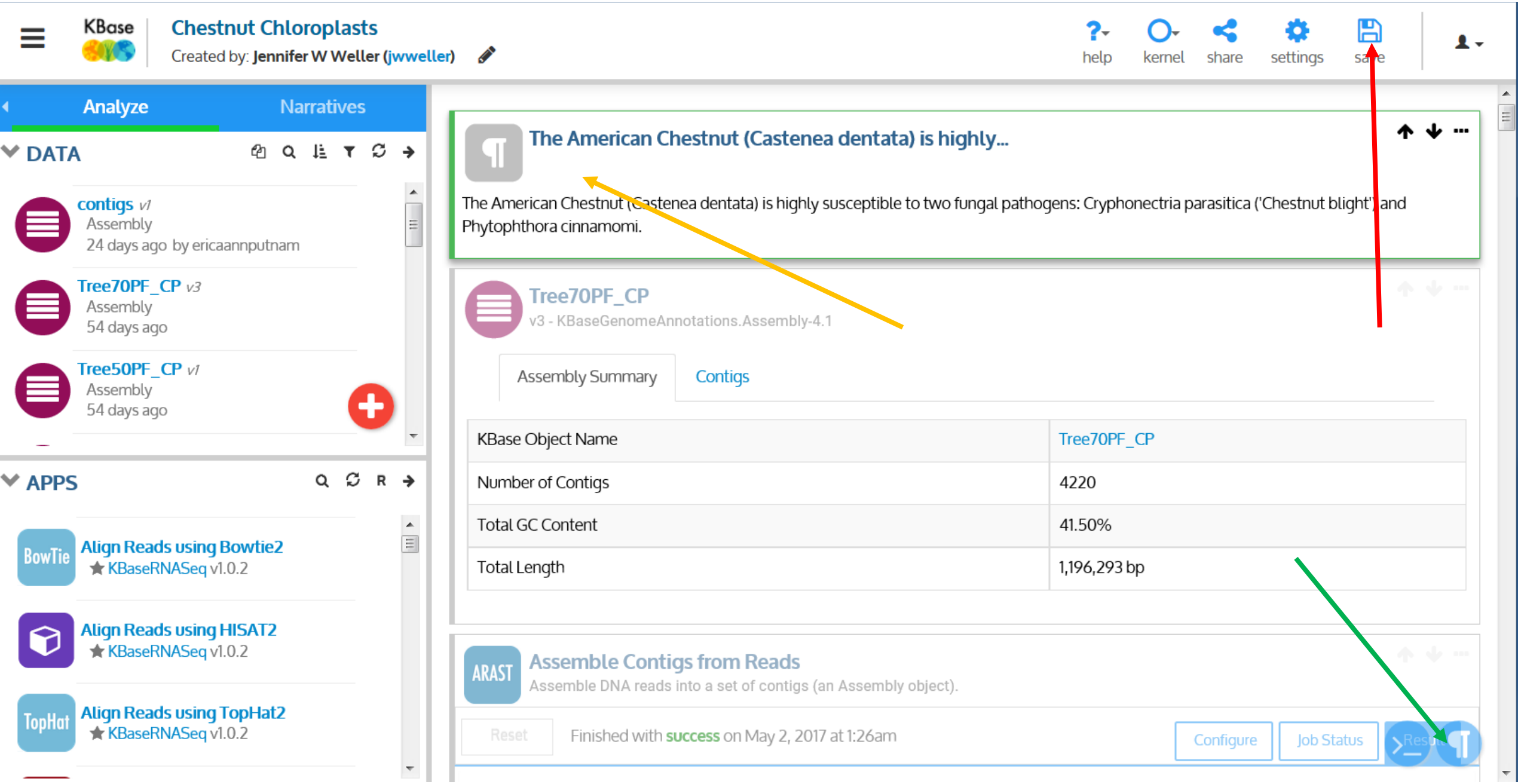
- Did we get complete coverage of any of the genomes?
  - Which were best and which were worst? Why?
- Are there any surprises for genome length or gene content?
- Do the chloroplasts correspond to what we think we know about their lineage?
- Other? (compare them to each other? To other reference tree cp?) Dane, Tree Genetics and Genomes – Chinkapin cp (C. pumila) Maybe ease of use or challenges to using Kbase?
- Assignments: (we have to analyze our results first)

The B3 Chloroplast paper - References

- Keep a running list - different journals use different styles but as long as you include the basic information your can intercovert as needed:
- Author names, Paper title, Date, Journal title and volume and page information, digital object identifier (doi where possible)
- Web references – example:

Using Kbase for both analysis and for collaboration

- Adding content
- Adding structured content (html)


KBase
Chestnut Chloroplasts
Created by: Jennifer W Weller (jwweller)
help
kernel
share
settings
save
Analyze
Narratives
DATA
contigs v1
Assembly
24 days ago by ericaannputnam
Tree70PF_CP v3
Assembly
54 days ago
Tree50PF_CP v1
Assembly
54 days ago
APPS
Align Reads using Bowtie2
KBaseRNASeq v1.0.2
Align Reads using HISAT2
KBaseRNASeq v1.0.2
Align Reads using TopHat2
KBaseRNASeq v1.0.2
The American Chestnut (Castenea dentata) is highly...
The American Chestnut (Castenea dentata) is highly susceptible to two fungal pathogens: Cryphonectria parasitica ('Chestnut blight') and Phytophthora cinnamomi.
Tree70PF_CP
v3 - KBaseGenomeAnnotations.Assembly-4.1
Assembly Summary
Contigs
KBase Object Name
Tree70PF_CP
Number of Contigs
4220
Total GC Content
41.50%
Total Length
1,196,293 bp
Assemble Contigs from Reads
Assemble DNA reads into a set of contigs (an Assembly object).
Reset
Finished with success on May 2, 2017 at 1:26am
Configure
Job Status

KBase

Chestnut Chloroplasts

Created by: **Jennifer W Weller** (jwweller)

help | kernel | share | settings | save

Analyze | Narratives

DATA

- **contigs** *v1*
  Assembly
  24 days ago by ericaannputnam
- **Tree70PF_CP** *v3*
  Assembly
  54 days ago
- **Tree50PF_CP** *v1*
  Assembly
  54 days ago

APPS

- **Align Reads using Bowtie2**
  ★ KBaseRNASeq v1.0.2
- **Align Reads using HISAT2**
  ★ KBaseRNASeq v1.0.2
- **Align Reads using TopHat2**
  ★ KBaseRNASeq v1.0.2

Introduction

Introduction

The researchers responsible for the data production, analysis and creation of this paper include: Dr. Jennifer Weller (UNCC), Dr. Paul Sisco (NCSU emeritus), Jeanne Smith, Erica Putnam (Olympic High School), Steve Barilovits (TACF), and the students of the Olympic High School B3 Summer Science camp 2016 (Caleb Judd, Melanie Loor, Tamdan Le, (note to add remaining names) and we gratefully note technical help from Dr. Cathy Moore (UNCC).

Background The American Chestnut (Castenea dentata) is highly susceptible to two fungal pathogens: Cryphonectria parasitica ('Chestnut blight') and Phytophthora cinnamomi.

Materials and Methods DNA was extracted from leaves collected from 6 individuals, described below. The published chloroplast sequence of a Chinese Chestnut (Castenea mollissima) was used to design overlapping long amplicons. for each individual the long amplicons were pooled and subjected to the Nextera library prepartion protocol for Illumina sequencing. Indices were added during the process to enable the libraries to be pooled for a MySeq 300-bp paired-end run. The output data was sorted by indices and then fastq files were produced using the Illumina BaseSpace default applications. Both de novo assembly and assembly against the Chinese chloroplast reference was carried out with BaseSpace tools. These data sets were then uploaded to the Kbase environment for further analysis and visualization.

Tree70PF_CP

v3 - KBaseGenomeAnnotations.Assembly-4.1

Assembly Summary | Contigs

KBase Object Name	Tree70PF_CP
Number of Contigs	4220

Proposed Structure and Collaboration Process

- Each major section will be a text box.
- Add your section (order can be rearranged), give it the section type label (‘Introduction’) – suggest you put your initials there as well.
- If you find information to add to someones section, preface it with (Comment - initials) - information – Reference (End Comment)

- Additional tags: Question, Answer
- SAVE!

Formatting the text:

- Kbase expects that you will use a markup language, either html or LaTex, to format the text – that is, to make it look nice.
- HTML is much easier and also very useful if you want to make your own Web pages, so we are going to use it here.

Formatting and including pictures

- This interface expects you to use a markup language, such as html or LaTex. I use this for the class Web pages that I post, so there is some sample material that I can use.
- For example, if I want Introduction to be in bold, I put brackets around the word:
- First, explain what markup language you are using: <html>
  - Then use the markup symbols for the presentation style you want: means bold. So I can format the work Introduction using <html>Introduction
- Try this with the heading of your section, then type a few more words – what happens?
  - Everything will be in bold – so I have to tell it when to stop:
  - <html>Introduction</html>
- Feel free to play around with this – just make sure you don't change the content.

Embedding graphics (an image or picture)

- <html>
- <body>
- Introduction
-
- <figure>
- <img src= "https://webpages.uncc.edu/~jweller2/pages/SummerCamp2016/SummerCamp2016_Pictures/LowRes_16June/LeafHopper.jpg" style="width:48px; height:98px;">
- </figure>
-
- </body>
- </html>

Where to look up styles:

- There are lots of free Web tutorials and lists – try
  - https://www.w3schools.com/tags/

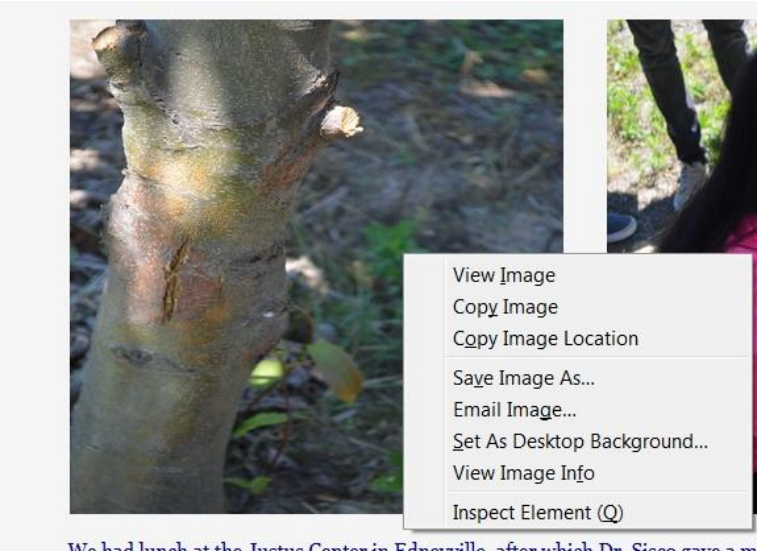
- For example, if you want colored text you can use the style attribute
- <h1 style="color:blue;text-align:center"> This is a header</h1>
- This is a paragraph.
- A leaf hopper on one of our samples. can be used to color just part of a text statement
- <figure>
  - <img src= "https://webpages.uncc.edu/~jweller2/pages/SummerCamp2016/SummerCamp2016_Pictures/LowRes_16June/LeafHopper.jpg" style="width:48px; height:98px;">
  - <figcaption> Fig 1. A leaf hopper on one of our samples. </figcaption>
- </figure>
- Italics is   content

You can take images from the class Web pages as follows

Go to the class home page:
http://webpages.uncc.edu/~jweller2/pages/SummerCamp2016/SummerCamp2016_Home.html

Go to Pictures (for example):
http://webpages.uncc.edu/~jweller2/pages/SummerCamp2016/SummerCamp2016_Pictures.html

Find the picture you want, hover the mouse over it and right-click, select Inspect element



<img src="SummerCamp2016_Pictures/LowRes_16June/FungalLesions5.JPG" width="300" height="300" align="right;">

Embedding a Table

- Of course, you might want to save it as an image and insert it as a figure, but you can also format one this way:
- <table>
  - <tr>
    - <th> column 1 </th>
    - <th> column 2 </th>
  - </tr>
  - <tr>
    - <th> column 1 </th>
    - <th> column 2</th>
  - </tr>
- </table>
- tr is a row, th is a column, and you could specify a single cell using <td>
- There are ways to make the table display with borders and line colors – you can look these up if you want to get fancy

Tables

- The tree plot – seq sample ID-hybrid type
- Make a table of the following information (Sample Origin: Sample Description: Illumina Sample ID: Nextera Primer Index):
  - 01: 100% American from Crowder's Mt. SP:240916_01:N701(TAAGGCGA)
  - 02: Pryor Farm Tree 4: 50% American female /50% Japanese male, male sterile:240916_02:N702(CGTACTAG)
  - 03: 100% Chinese (probably) from a Charlotte NC: 240916_03:N703(AGGCAGAA)
  - 04: Pryor Farm Tree 43, 100% American:240916_04:N704(TCCTGAGC)
  - 05: Pryor Farm Tree 70, 50% American female, 50% Chinese male, male sterile:240916_05:N705(GGAATCCT)
  - 06: Pryor Farm, 50% Old NC10 American female (chinkapin cp), 50% Chinese male, male fertile:240916_06:N706(TAGGCATG)
  - Make a figure legend that explains what is in each column, and add the comment that the common end on the DNA fragments is Nextera S501 (TAGATCGC)

- The Primer Pairs used for long-range PCR (will likely go in Materials and Methods) are probably easiest to present in an image, but until I provide a source, it will have to be a table.

C mollissima Chloroplast _length = 160799)		primers to create 15000bp overlapping amplicons			
Working stock index	5'-->3' sequence	Start	End	Amplicon length	Order Name
1f	AAG CAG AAG TGA TGT GGA TT	155490	7889	13198	Cm_CP1F
1r	GGT AAG AGG TAA GAT GAG AGC				Cm_CP1_R
2f	AGA CAG CCG CAT ATA TTG AA	7634	22119	14485	Cm_CP2F
2r	TGG TCG TGT ATT AGC AGA TG				Cm_CP2R
3f	AGA TTG ACC CGC GAT AAT AC	21879	35631	13752	Cm_CP3F
3r	GAT GAC TTA CGC CTT ACC AT				Cm_CP3R
4f	TTA TCG TTC CTG AAT GGT CC	35414	48808	13394	Cm_CP4F
4r	CAG GAA GGT TGG CTA GAA AT				Cm_CP4R
5f	AGG GAT CAA TTT CTA GTC GC	48532	62074	13542	Cm_CP5F
5r	AAG ATT GGG CCG AGT TTA AT				Cm_CP5R
6f	CTA TTT TGT GCC GAA GCA AT	61244	76206	14962	Cm_CP6F
6r	CCA TCA ACC TGC TAG TTC TT				Cm_CP6R
7f	CCT CTT AGT CCG TTG TTG AA	75399	89557	14158	Cm_CP7F
7r	AAC CCA AGG TCA TCA TTC TC				Cm_CP7R
8f	AAG CGG TGA GTT GGT TAT TA	88674	102532	13858	Cm_CP8F
8r	TAT GCG GTG CTA ACG ATT TA				Cm_CP8R
9f	AGA GAT TCG TTG TTC CTG AC	102010	115821	13811	Cm_CP9F
9r	TTC TTC CGT TTC TGA GGT TT				Cm_CP9R
10f	ATG AAA TGG GTT GGA TTG GT	115603	128928	13325	Cm_CP10F
10r	ATT GGG AAG TAC AAT GGC AA				Cm_CP10R
11f	TGA GAT CGG CTA AAA CAA GG	128100	142594	14494	Cm_CP11F
11r	CAA CGT ATC TTC ACA GAC CA				Cm_CP11R
12f	GTG AAG TAA ATC ATC GCA CC	142016	155734	13718	Cm_CP12F
12r	AAC GGA GGA ACA ATA TCA CC				Cm_CP12R
Cm_IGStrnT_F	CAT TAC AAA TGC GAT GCT CT	51758		810	
Cm_IGStrnT_R	TCT ACC TAT TTC GCC ATA TC	52568			

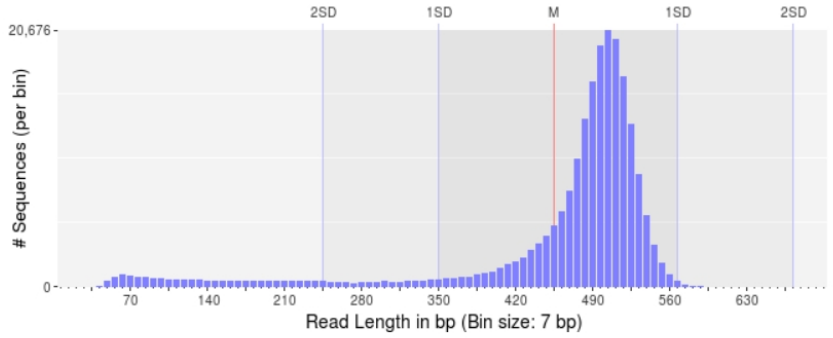
Remember this?
Explain in the legend.

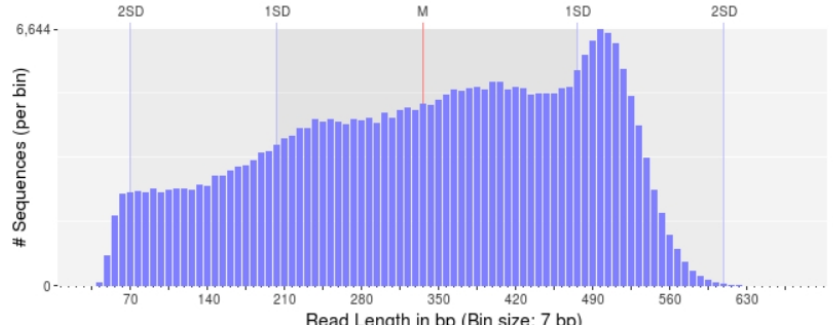
What could I use to make a graphic showing now the amplicons overlap?

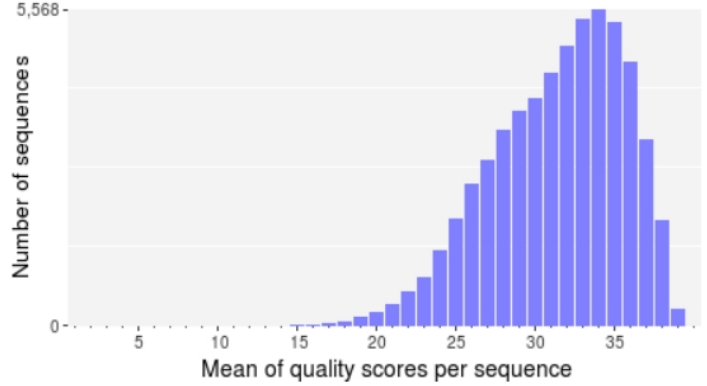
- We don’t have a visualization tool for this on Kbase.
- DNAPlotter does have a Windows version and a MacOS version (as well as the unix type most often used in Bioinformatics), so this is something that can be installed on the types of computers you most likely have at school or at home. http://www.sanger.ac.uk/science/tools/dnaplotter
- I will create the input file and send it out – those of you who want to try it with the software can do so – I’ll save it as an image file so we can insert it in the document.
- Note: this program can ALSO be used to add the features (genes) to a graphical image of the whole genome, when we figure out what their stop and start positions are, and the gene labels.

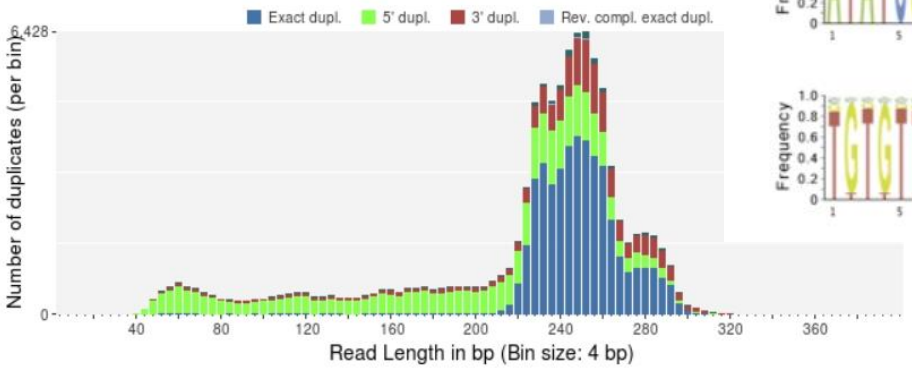
Types of summary graphics we will want – some tables, some graphical (AKA visualizations)

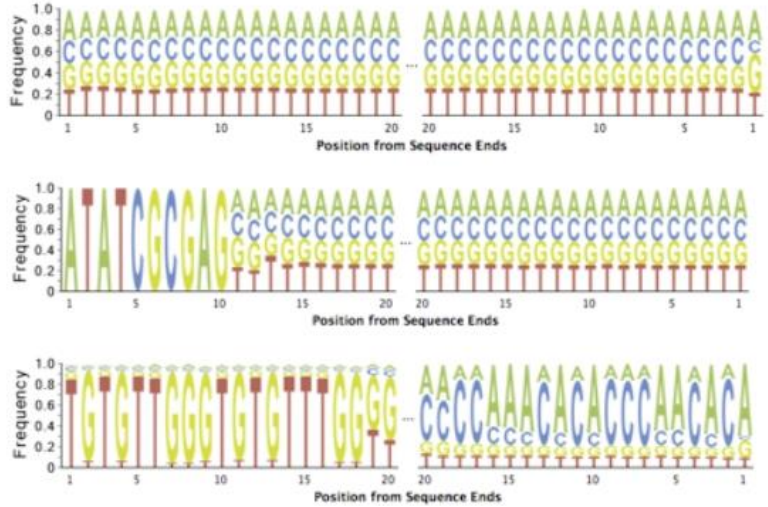
- Number of contigs, quality scores, duplicate sequences, where tags from our long PCR might be. Some of this is reported in the Illumina statistics for the runs.
  - Another program that can be used as a Web service (you upload your data to the site) is Prinseq (described here: http://prinseq.sourceforge.net/manual.html)











To trust the data you need both quality and coverage (30X when there is a reference, 100X when it is de novo)

- A graphic might look like this image (the genome browser IGV was used, it can take contig files and include the quality score) – we might just use the visualization when we assemble the longer contigs, to help make the decision about whether we should go ahead or not:

Types of summary graphics we will want – some tables, some graphical (AKA visualizations)

- Gaps in alignments, gaps relative to reference – one possible solution is
  - “GapBlaster a graphical gap filler for Prokaryotic Genomes” (PLOS ONE, 2016, Pablo HCG de Sa et al., doi.org/10.1371.journal.pone.0155327) , which allows you to enter a set of contigs and a reference, and visually inspect it.
- Annotations - graphical on genome, or a table with start and stop locations and the gene name or abbreviation.
  - There are several tools in Kbase for this – pick one that looks at prokaryotic genomes. They do a sequence comparison by aligning our sequence with known genes – if there is enough similarity a match will be declared. Proper start and stop signals will be identified and labeled.

Examples and Tools in Kbase

- What does Kbase have available?
- Narratives – Tutorial or Shared
- Apps (functional pipelines)

KBase

Your Dashboard

Dashboard

Catalog

Account

Your Narratives

New Narrative

Search Your Narratives

Chestnut Chloroplasts

- Assemble Contigs from Reads
- 4 markdown cells

8

saved 13 hours ago by kimk98

Tutorial Narratives

Search Tutorials

Ecoli RNA-seq Analysis - Update in Proc

- Align Reads using HISAT2 - v2.0.4
- Assemble Transcripts using StringTie
- Create Differential Expression Matrix
- Create RNA-seq Sample Set
- Identify Differential Expression using
- Load Single-End Reads From Web
- View CummeRbund Plots

Arabidopsis RNA-seq Analysis using Ne

- Align Reads using HISAT2 - v2.0.4
- Assemble Transcripts using StringTie
- Create Differential Expression Matrix
- Create RNA-seq Sample Set
- Load Single-End Reads From Web
- 15 markdown cells

The PlantSEED Resource in KBase

- Annotate Plant Transcripts with Meta
- Build Metabolic Model
- Run Flux Balance Analysis
- 6 markdown cells

Community Modeling Protocol: Mult

- Run Flux Balance Analysis
- merge_to_community_model
- run_flux_balance_analysis
- 11 markdown cells

RNA-seq Analysis Tutorial using Original Tuxedo Suite in Arabidopsis

> NOTE: This tutorial is view-only, allowing you to see, but not alter, the input and output of the KBase apps used in this workflow. To run the steps yourself in a new Narrative using your own data or different parameters, create a new Narrative using the "New Narrative" button at the top left and follow all the steps given in this Narrative. If you just want to read this Narrative, you still can see the data objects generated in the workflow by using the "Controls" link at the top left. For more information, please see the Narrative Interface User Guide.

RNA-seq Primer

RNA-seq uses next-generation sequencing to account for all the transcripts in the biological sample at a particular time and therefore, can be used for a variety of purposes such as transcriptome assembly, gene discovery/annotation, and detection of differential transcript abundances between tissues, developmental stages, genetic backgrounds, and environmental conditions. Generally, RNA-seq is mainly used to compare gene expression between different conditions, such as control and treatment and find out which genes are up- or down regulated in each condition. Similarly, samples can be obtained from a group of wild and mutant genotype to get the candidate genes responsible for the genetic differences that might explain the observed phenotype differences.

KBase provides a number of tools for discovery of candidate genes starting with importing of raw sequencing reads and genome, reads processing using reads QC tools like fastQC and Trimmomatic, and then aligning reads to the reference genome and produces a transcriptome assembly, lists of differentially expressed genes and further downstream analysis to refine regulated genes and transcripts.

The KBase RNA-seq Service provides data analysis tools (Apps) that are based on the original and new Tuxedo suites to get the normalized full and differential expression matrix of the reads obtained from **Illumina or Solid platform** using the **reference genome**.

The original Tuxedo suite consists of **Bowtie2/TopHat2** to align the reads, **Cufflinks** to assemble the transcripts, **Cuffdiff** to identify the differentially expressed genes and **CummeRbund** to visualize the differentially expressed genes obtained from Cuffdiff as 2D plots[1,2,3].

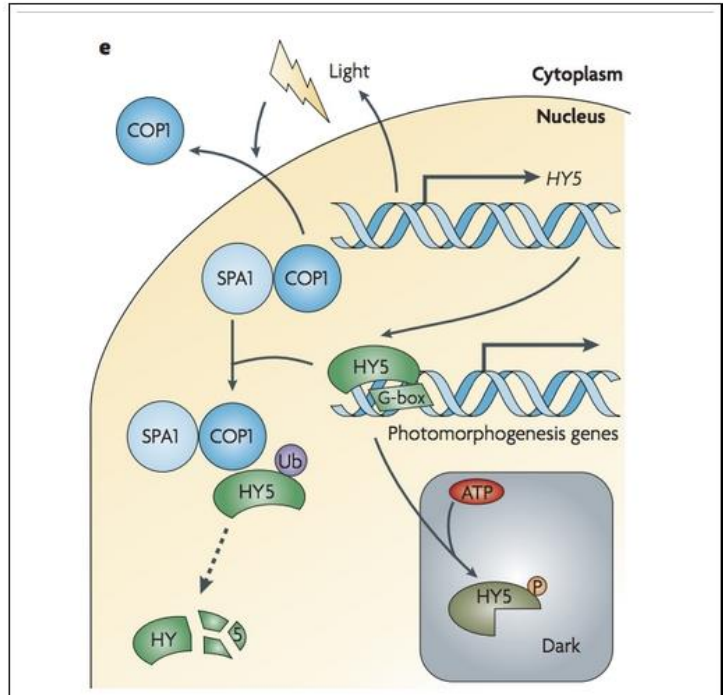
The new Tuxedo suite uses **HISAT2** instead of (TopHat2) to align reads and **StringTie** instead of Cufflinks to assemble the transcripts and **Ballgown** instead of Cuffdiff to identify the

Arabidopsis Transcriptome Analysis based on Wild and Mutant Studies

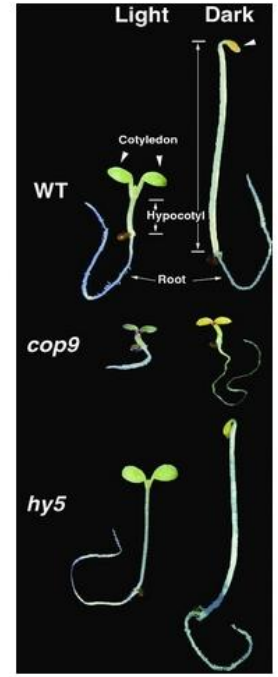
Description

Light is one of the most important stimulus to plant development. In Arabidopsis, the seedling hypocotyl has emerged as an exemplar model to study light control of cell expansion. Seedlings grown in the light show a short hypocotyl with green and expanded cotyledons whereas dark grown seedlings show a long hypocotyl with yellow and unopened cotyledons. LONG HYPOCOTYL 5 (HY5) is a basic leucine zipper transcription factor (bZIP TF) that binds directly to the promoters of photomorphogenic genes and promote their expression. Mutations in the HY5 gene in seedlings cause lateral root formation, longer hypocotyl and affects greening in hypocotyls [6,7]. It shows that HY5 gene plays role in the coordination of light signaling and appropriate gene expression and is responsible for the regulation of fundamental developmental processes such as cell elongation, cell proliferation and chloroplast development [8,9,10] suggesting that HY5 could act as a key modulator of signal transduction pathways that links a wide varity of stimulus responses and developmental processes in the seedlings to coordinate development.

For this narrative, the published study [11] on RNA sequencing to examine HY5 mediated gene expression changes between wild-type and hy5 mutant plants is used.



General mechanism of transcription factor regulation by light



Wild, cop9 and hy5 mutant types

Data Description

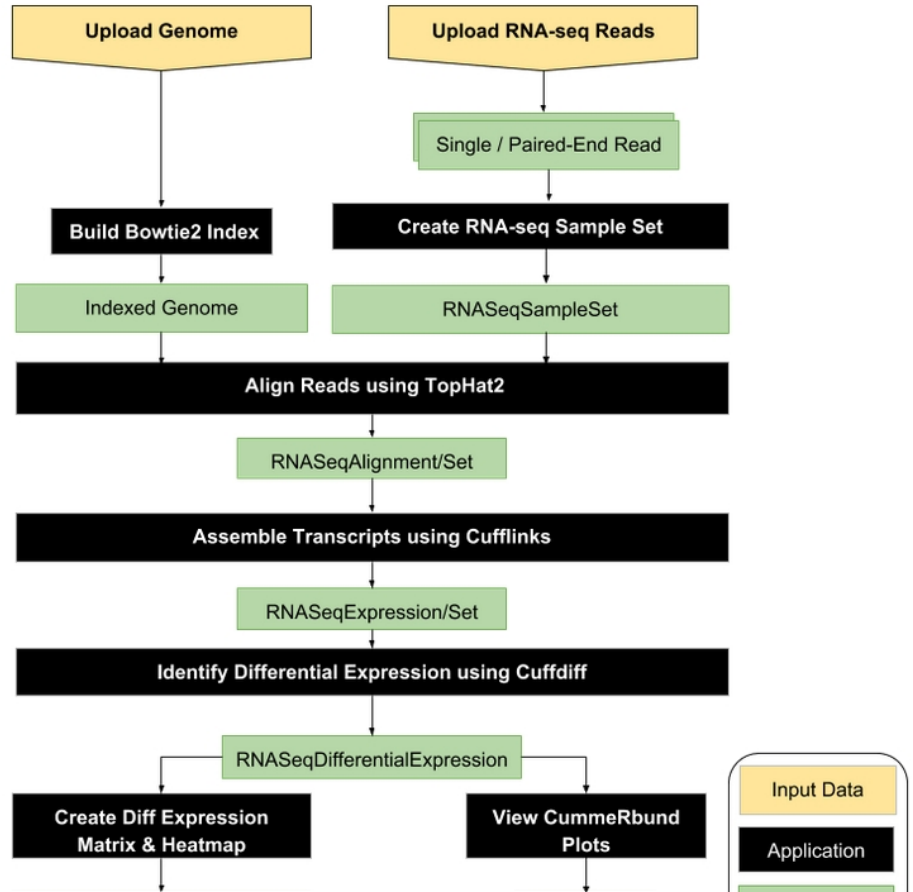
This Narrative uses a small sample dataset from NCBI's Sequence Read Archive (SRP003951, Zhang *et al.* 2011) to identify differential gene expression between wild-type (WT) and mutant-type (HY5) samples in *Arabidopsis thaliana* using KBase's RNA-seq pipeline

In this Tutorial, the RNA-seq workflow uses Tophat2 -> Cufflinks -> Cuffdiff.

This tutorial uses single-end Illumina sequencing, but similar analysis steps can be done for paired-end sequencing as well.

- **Step 1:** Import of Reference Genome
- **Step 2:** Build Bowtie2 Index
- **Step 3:** Load Single-End Reads from Web
- **Step 4:** Create RNA-seq Sample Set
- **Step 5:** Align Reads to Reference Genome using Tophat2
- **Step 6:** Assemble Transcripts using Cufflinks
- **Step 7:** Create Differential Expression Matrix using Cuffdiff
- **Step 8:** View CummeRbund Plots
- **Step 9:** View Interactive Volcano Plot
- **Step 10:** Create Differential Expression Matrix and HeatMap using Cuffdiff

---



Annotate Assembly with Prokka v1.12

ProkkaAnnotation v.1.0.0

by rsutormin

★ 1 | R B D | ↗ 425 | ✔ 92.0% | 17m 54s

Annotate Assembly with Prokka annotation pipeline.

This is a KBase wrapper for the Prokka annotation pipeline.

Prokka is a software tool for the rapid annotation of prokaryotic genomes. A typical 4Mbp genome can be fully annotated in less than 20 minutes.

For rRNA prediction this app currently uses Barrnap (written by the author of Prokka and recommended if you prefer speed over absolute accuracy).

Prokka version: 1.11

Barrnap version: 0.7

Inputs

Assembly *KBaseGenomeAnnotations.Assembly*
Assembly object (set of contigs)

Outputs

Dashboard
Catalog
Account

back to the Catalog

Assemble with SPAdes

kb_SPAdes v.0.0.12

by gaprice

★ 2 | R | 331 | ✔ 70.1% | 1h 55m

Assemble reads using the SPAdes assembler.

This is a KBase wrapper for the SPAdes genomic reads assembler.

SPAdes is designed for assembling reads. We strongly recommend not to use it for large and medium-size eukaryotic genomes.

Operational notes:

- Currently the wrapper only supports Illumina, IonTorrent, PacBIO CLR and PacBIO CCS in FASTQ format, either uncompressed or gzipped.
- The --careful flag is always used, except for metagenomic assemblies where it is not allowed.
- Metagenome and Plasmid assembly can only be run on one reads library.
- Metagenome assembly can only be run on a paired end library.
- Illumina and IonTorrent reads can not be mixed in the same assembly.
- PacBIO CLR needs to be run with at least one accompanying Illumina or IonTorrent library.
- The k-mer parameter is autodetected by SPAdes.
- The PHRED parameter is autodetected by EAUtils.

If you need support for command line options not exposed in the wrapper please contact KBase Help.

SPAdes version: 3.10.0

Assemble with Velvet

AssemblyRAST v.0.0.5

by fangfang

★ 2 | R B D | 142 | ✔ 85.9% | 49m 30s

Assemble short microbial reads using the Velvet assembler.

This is a KBase wrapper for Velvet.

Velvet is a classic de-bruijn graph based assembler. Velvet works by efficiently manipulating de Bruijn graphs through simplification and compression. It eliminates errors and resolves repeats by first using an error correction algorithm that merges sequences together. Repeats are then removed from the sequence via the repeat solver that separates paths which share local overlaps. Velvet is fast and robust, although it's no longer in active development.

Known limitations: Velvet assembly quality is known to be sensitive to the hash length.

Inputs

Read Library *KBaseAssembly.SingleEndLibrary*, *KBaseFile.SingleEndLibrary*, *KBaseAssembly.PairedEndLibrary*, *KBaseFile.PairedEndLibrary*
Read library

Outputs

Output ContigSet name *KBaseGenomeAnnotations.Assembly*
Enter a name for the assembled contigs data object

Parameters

Minimal contig length
Minimum length of contigs to output, default 300

Advanced assembly parameters
Enter extra assembly parameters as strings (hash_length=29 auto_insert=False)

Assemble Contigs from Reads

AssemblyRAST v.0.0.5

by cbun, fangfang

★ 3 | R B D | 801 | ✔ 82.9% | 3h 29m

Assemble DNA reads into a set of contigs (an Assembly object).

This app can be used to perform an automatic genome assembly using the latest computational tools. Single or multiple assemblers can be invoked to compare results. Resulting assemblies are automatically processed via a collection of analysis tools developed by both KBase and the research community. The app attempts to select the best assembly (the smallest number of contigs, the longest average contig length) to suggest to the user.

Several assembly workflows or "recipes" are available. These workflows have been tuned and tested to fit certain dataset types or desired analysis criteria such as throughput or rigor. The compute engine's flexible nature also enables the rapid design and emulation of other popular protocols.

Additionally, custom workflows can be designed and executed in "pipeline" mode without having to compose complicated scripts. Workflows can be composed with combinations of quality filtering or trimming, error correction, adapter removal, assembly, scaffolding, or post-processing.

Assembly Recipe Descriptions:

Automatic Assembly:

- Provides a nice balance between "fast pipeline" and "smart pipeline"
- Runs BayesHammer on reads
- Assembles with Velvet, IDBA and SPAdes
- Sorts assemblies by ALE score

Fast Pipeline:

- Assembles with A6, Velvet and SPAdes (with BayesHammer for error correction)
- Results are sorted by ARAST quality score

Smart Pipeline:

- Runs BayesHammer on reads, KmerGenie to choose hash-length for Velvet
- Assembles with Velvet, IDBA and SPAdes
- Sorts assemblies by ALE score
- Merges the two best assemblies with GAM-NGS

Tree70PF_CP

v3 - KBaseGenomeAnnotations.Assembly-4.1

Assembly Summary | Contigs

KBase Object Name	Tree70PF_CP
Number of Contigs	4220
Total GC Content	41.50%
Total Length	1,196,293 bp

Objects

Created Object Name	Type	Description
contigs	Assembly	Assembled contigs

Summary

```
============= Raw Contigs ============
QUAST: All statistics are based on contigs of size >= 500 bp, unless otherwise noted (e.g., "#
contigs (>= 0 bp)" and "Total length (>= 0 bp)" include all contigs).

Assembly                    spades_contigs  velvet_contigs
# contigs (>= 0 bp)         6007            127509
# contigs (>= 1000 bp)      96              0
Total length (>= 0 bp)      2866753         9977420
Total length (>= 1000 bp)   168436          0
# contigs                   960             2
Largest contig              19768           507
Total length                670701          1010
GC (%)                      36.10           41.39
N50                         589             507
N75                         522             503
L50                         321             1
L75                         630             2
# N's per 100 kbp           0.15            0.00


========== Filtered Contigs ==========
ContigSet saved to: jwweller:1490796672910/contigs
Assembled into 5595 contigs.
Average Length: 496.970688114 bp.
Contig Length Distribution (# of contigs -- min to max basepairs):
```

Summary

```
============= Raw Contigs ============
QUAST: All statistics are based on contigs of size >= 500 bp, unless otherwise noted (e.g., "#
contigs (>= 0 bp)" and "Total length (>= 0 bp)" include all contigs).

Assembly                    spades_contigs  velvet_contigs
# contigs (>= 0 bp)         6007            127509
# contigs (>= 1000 bp)      96              0
Total length (>= 0 bp)      2866753         9977420
Total length (>= 1000 bp)   168436          0
# contigs                   960             2
Largest contig              19768           507
Total length                670701          1010
GC (%)                      36.10           41.39
N50                         589             507
N75                         522             503
L50                         321             1
L75                         630             2
# N's per 100 kbp           0.15            0.00


========== Filtered Contigs ==========
ContigSet saved to: jwweller:1490796672910/contigs
Assembled into 5595 contigs.
Average Length: 496.970688114 bp.
Contig Length Distribution (# of contigs -- min to max basepairs):
```