The pan-genome of *Brachypodium distachyon*, capturing the full genetic complement of a plant species

John Vogel

Outline

- Pan-genome concept
- *Brachypodium distachyon* as a model grass
  - Natural diversity
- *B. distachyon* pan-genome
  - Multiple methods
  - Size
  - Characterization of non-reference genes
  - Population studies

Pan-genome: *E.coli* example

non-redundant sequence space of a species or genus

- Each line differs from the pan-genome by either substitution or deletion
- We want non-coding as well as coding sequence, CNS elements...etc.

Model plant: *Brachypodium distachyon*

- Model for biomass grasses
- JGI flagship
- Mutant resources
- Experimentally tractable
  - Small, self fertile
  - Small genome (272Mb)
  - Short generation time, 8 weeks
  - Easily transformed
  - 'Finished' genome

Plant pan-genomics

- Wild soybean (*Glycine soja*) (Li et al. 2014 Nat. Biotech.)
  - De-novo assembly of 7 diverse accessions
  - 80% of genes were core
  - Relatively little novel sequence in the pan-genome—only 30Mb larger than any single accession
  - Limitations: low number of accessions, poor assemblies, the lack of a proper reference genome (they used *G. max*)
- Arabidopsis (Cao et al. 2014 Nat. Gen.)
  - Primarily used a reference-based approach with only targeted de-novo assembly.
  - Identifying variants was the focus not characterizing the pan-genome
- Maize (Hirsch et al. 2014 Plant Cell v26)
  - Analyzed 503 transcriptome assemblies
  - Only 20% of genes were core
  - Limitations: differential expression, assembly quality, very complex genome

Natural diversity

Feruloylated arabinan LM12

Photosynthetic rate

Bob Furbank, CSIRO, Australia
B. distachyon genome sequence

Jeremy Schmutz, Jane Grimwood, Jerry Jenkins, Shengqiang Shu, JGI/Hudson-Alpha

- Assembly V3:
  - An essentially finished sequence: 10 scaffolds and 34 contigs
  - Contig N50 of 52 Mb with 100% of sequence in chromosomes
  - Error rate <1 base per 100kb
  - 34,310 gene loci

1,303,793 pan

Essential for molecular breeding
Missed by reference based approaches

449,125
2
258
An essentially finished sequence: 10 scaffolds and 34 contigs

5
Assembly V3:

SNPs
Create a pan
Error rate <1 base per 100kb

409
665,595
644
209

Bd2
Bd3
Shu, JGI/Hudson

# bp per Contig N50 of 22 Mb with 100% of sequence in chromosomes

May be a large percentage of diversity

426
Genome sequence

What about novel sequence?

- Missed by reference based approaches
- May be a large percentage of diversity in any species
- Essential for molecular breeding strategies
- Create a pan-genome that captures whole species diversity

B. distachyon pan-genome a test case

De-novo assembly

Gene-based pan-genome

Sequence-based pan-genome

Annotate individually

Identify novel genomic sequence

Cluster genes at the nt level

Cluster predicted proteins

Annotate

Number genes in pan-genome

Number of genes in pan-genome

Number of genes in pan-genome

Assembly and annotation quality

median genome size 268Mb (reference 272Mb) mean N50 = 68,516

Using synteny to order contigs

Illumina platform
>30-70x coverage
>750bp paired ends from 250 bp fragments

**Sequence-based pan-genome**
- Very conservative approach
- Iteratively add stretches >600 bp with no previously added 21 mers
- Adds 160 Mb genomic sequence
- Adds 15,000 genes

**Gene-based pan-genome**
- CDOS sequences from genes are clustered by an orthoMCL-like algorithm and then collapsed at 90 percent protein identity
- 61,155 pan-genome clusters
- 26,125 more genes in the pan-genome

**Pan-genome compartments**
- Cloud was excluded from most analyses
- About 28% of cloud genes are expressed in leaves

**High-confidence pan-genome**
- 37,886 high confidence (non-cloud) pan-genome clusters
- Comparison to reference DNA
- 7,283
- 13,408
- 2,557
- 30,751

**Functional attributes**
- Expression
- Expression pattern
- Evolutionary rate
- Narrow: shell
- Broad: core
Population analysis

Conclusions

- One plant genome may only contain about 50% of the genes in a species
- Dispensable genes are expressed lower, have higher variation in expression, are expressed more narrowly across tissues, evolve faster
- The pan-genome (and individual genome assemblies) will accelerate breeding by increasing accuracy and resolution of sequence-based breeding strategies

Acknowledgements

Vogel Lab
- Sean Gordon
- Lifeng Liu
- Joseph Knelman
- Amy Cartwright
- Andrew Katz
- Ryan Hill

JGI
- Dan Rokhsar
- Kerrie Barry
- Wendy Schackwitz
- Joel Martin
- Shengqiang Shu

HudsonAlpha
- Jeremy Schmutz
- Jerry Jenkins
- Jane Grimwood

Collaborators

- David Des-Marais, Harvard
- Pilar Catalan, Univ. of Zaragoza, Spain
- Bruno Contreras Moreira, CSIC, Spain
- Rob Hasterok, U. of Silesia, Poland
- Tom Juenger, U. of Texas, Austin
- Luis Mur and John Doonan Univ. of Aberystwyth, UK
- Daniel Woods and Rick Amasino, UW, Madison

- Ana Caicedo, U. Mass., Amherst
- Eric Lyons, U. of Arizona
- Hikmet Budak, Sabanci Univ. Turkey.

Funding