From:	Scott Allen Jackson
To:	Steven Cannon
Cc:	Richard Michelmore; Scheffler, Brian; rajeev varshney; Peggy Ozias-Akins; Corley Holbrook; Lutz Froenicke (lfroenicke@ucdavis.edu); Jeremy Schmutz; David Bertioli; Howard Valentine; Xin LIU; Mark Burrows; Guo, Baozhu; Soraya Bertioli; David Bertioli; Schnell, Ray; Victor Nwosu; Rich Wilson; Agarwal, Gaurav (ICRISAT-IN); Brian Abernathy; Ran Hovav; Kale, Sandip (ICRISAT-IN); Pandey, Manish K (ICRISAT-IN); Sudhansu Dash; Ethalinda Cannon; Nathan Weeks; Andrew D. Farmer; Longhui Ren; Huang, Wei [AGRON]
Subject:	Re: Diploid annotations for review - revisions, updates; "consensus" gene models?
Date:	Thursday, September 18, 2014 8:14:46 PM

I suggest we vote by email. Once that is determined, i think we can move forward to release it with the usual large genome data sets caveats. (my votes are below, but you can just reply to me and Steve, if you wish, and we can tally.)

Questions:
1. Maker or Glean as primary annotation.
2. Do we make the other annotation available via peanutbase.

My votes:
1. Maker, because of the support for the annotation pipeline in the US and ability to redo annotations as we need.
2. Yes, available as a track/download via peanutbase.

scott

On Sep 18, 2014, at 1:08 PM, Cannon, Steven <Steven.Cannon@ARS.USDA.GOV> wrote:

> PGC group,
>
> As a follow-up to the discussion below, we have taken some time to assess the characteristics and qualities of the two sets of gene models: MAKER-P (from Andrew Farmer) and GLEAN (from BGI). It turned out to be a non-simple comparison, so the results are in the attached report.
>
> Here's the executive summary:
>
> Each gene model set has complementary strengths. There is value in retaining and making both available. There is also merit, however, in recommending one annotation set as the primary reference annotation for the genomes, in order to encourage consistency in publications and analyses. On balance, we recommend the MAKER-P set over GLEAN for use as the primary reference annotation, but we also recommend making both available to researchers, in order to take advantages of benefits of each annotation set.
>
> Of the five criteria evaluated, both methods generally performed well. The only characteristic that strongly separates the two methods is in the gene structural specification: the GLEAN models lack 5' and 3' UTR features. Among the other criteria, two criteria result in a toss-up (comparison against gene families and assessment of annotation quality), one weakly favors GLEAN (transcript support, by ~2 percentage points), and one weakly favors MAKER-P (lengths of

transcripts and peptide sequences, by ~5 percentage points).

The gene model sets are indeed complementary: each identifies more than 10,000 gene models that the other misses, out of ~40,000 genes identified in by each process in each species – with overlaps of ~27,000 in common (the numbers differ by species; see the details in Section 5). We don't recommend combining the sets, since the methods, model names, and gene structural characteristics differ between the two gene sets.

On the whole, I came away from the evaluation feeling good about both sets of gene models. They both seem pretty solid – for automated gene models from a draft assembly.

A few questions now:
- We've made a recommendation, but I think it ought to be up to the group to discuss, maybe refine, and then decide – by vote, consensus, etc.
- When we do settle on a decision (e.g. a primary reference set and secondary - or whatever the decision is), that moves us closer to releasing these publicly. So: when and how. Early November, in time for the Savannah meeting? Delay until the diploid paper is accepted? Something in-between?

A few other details: during the evaluation, we realized that the MAKER set lacked files of CDS sequences (we had provided only full-transcript sequences before). We have added the CDS sequences now – and also made a minor change to the GFF encoding of the CDS and UTR features (making the IDs identical to their parent features, rather than unique). The updated annotation sets are here:

[redacted]

Will be interested to hear the discussion.

Steven

From: Scott Allen Jackson <sjackson@uga.edu>
Date: Wednesday, August 27, 2014 at 9:24 AM
To: Steven Cannon <steven.cannon@ars.usda.gov>
Cc: "Varshney, Rajeev (ICRISAT-IN)" <r.k.varshney@cgiar.org>, "Lutz Froenicke (lfroenicke@ucdavis.edu)" <lfroenicke@ucdavis.edu>, Jeremy Schmutz <jschmutz@hudsonalpha.org>, "Scheffler, Brian" <Brian.Scheffler@ARS.USDA.GOV>, Richard Michelmore <rwmichelmore@ucdavis.edu>, Peggy Ozias-Akins <pozias@uga.edu>, David Bertioli <djbertioli@gmail.com>, Howard Valentine <hvalentine@peanutsusa.com>, Xin LIU <liuxin@genomics.org.cn>, "mburow@tamu.edu" <mburow@tamu.edu>, Soraya Bertioli <Soraya.Bertioli@embrapa.br>, David Bertioli <david.bertioli@pq.cnpq.br>, "Schnell, Ray" <Ray.Schnell@effem.com>, Victor Nwosu <victor.nwosu@effem.com>, Rich Wilson <rfwilson@mindspring.com>, Brian Abernathy <bla@uga.edu>, "Guo, Baozhu" <Baozhu.Guo@ARS.USDA.GOV>, "Holbrook, Corley" <Corley.Holbrook@ARS.USDA.GOV>, Ran Hovav

<ranh@volcani.agri.gov.il>, Sudhansu Dash <sdash@iastate.edu>, Ethalinda Cannon <ekcannon@iastate.edu>, Nathan Weeks <weeks@iastate.edu>, Andrew Farmer <adf@ncgr.org>, Longhui Ren <lhren@iastate.edu>, "Huang, Wei [AGRON]" <weih@iastate.edu>, "Kale, Sandip (ICRISAT-IN)" <S.Kale@cgiar.org>, "Pandey, Manish K (ICRISAT-IN)" <M.Pandey@cgiar.org>, "Agarwal, Gaurav (ICRISAT-IN)" <Gaurav.Agarwal@cgiar.org>
Subject: Re: Diploid annotations for review - revisions, updates; "consensus" gene models?

I think transcriptome comparisons would be useful, then we could make a preferred version.
-scott

On Aug 27, 2014, at 10:16 AM, Cannon, Steven <Steven.Cannon@ARS.USDA.GOV> wrote:

> We could identify one of the annotations as "preferred." So far, I don't have a sense for which one is better, except that the MAKER set had a few more transposon-like sequences in A. ipaensis (before the July 21 clean-up). Comparison with the transcriptome sequences would help. Also, some of the gene family analysis that we're doing here. But we might not know which one is better until mostly after the fact.
>
> Or, could do some more selective demoting and promoting — sort of a semi-hand-made collection. That's a lot of work though.
>
> - Steven
>
> ---
>
> **From:** Scott Allen Jackson <sjackson@uga.edu>
> **Date:** Wednesday, August 27, 2014 at 9:09 AM
> **To:** Steven Cannon <steven.cannon@ars.usda.gov>
> **Cc:** "Varshney, Rajeev (ICRISAT-IN)" <r.k.varshney@cgiar.org>, "Lutz Froenicke (lfroenicke@ucdavis.edu)" <lfroenicke@ucdavis.edu>, Jeremy Schmutz <jschmutz@hudsonalpha.org>, "Scheffler, Brian" <Brian.Scheffler@ARS.USDA.GOV>, Richard Michelmore <rwmichelmore@ucdavis.edu>, Peggy Ozias-Akins <pozias@uga.edu>, David Bertioli <djbertioli@gmail.com>, Howard Valentine <hvalentine@peanutsusa.com>, Xin LIU <liuxin@genomics.org.cn>, "mburow@tamu.edu" <mburow@tamu.edu>, Soraya Bertioli <Soraya.Bertioli@embrapa.br>, David Bertioli <david.bertioli@pq.cnpq.br>, "Schnell, Ray" <Ray.Schnell@effem.com>, Victor Nwosu <victor.nwosu@effem.com>, Rich Wilson <rfwilson@mindspring.com>, Brian Abernathy <bla@uga.edu>, "Guo, Baozhu" <Baozhu.Guo@ARS.USDA.GOV>, "Holbrook, Corley" <Corley.Holbrook@ARS.USDA.GOV>, Ran Hovav <ranh@volcani.agri.gov.il>, Sudhansu Dash <sdash@iastate.edu>, Ethalinda Cannon <ekcannon@iastate.edu>, Nathan Weeks <weeks@iastate.edu>, Andrew Farmer <adf@ncgr.org>, Longhui Ren <lhren@iastate.edu>, "Huang, Wei [AGRON]" <weih@iastate.edu>, "Kale, Sandip (ICRISAT-IN)" <S.Kale@cgiar.org>, "Pandey, Manish K

(ICRISAT-IN)" <M.Pandey@cgiar.org>, "Agarwal, Gaurav (ICRISAT-IN)" <Gaurav.Agarwal@cgiar.org>
Subject: Re: Diploid annotations for review - revisions, updates; "consensus" gene models?

I'm not sure about option 3, Steven. In rice we had two competing annotations for a while that caused all sorts of problems (though they came from different groups as opposed to one here). It seems that having two annotations just complicates things and analyses will have to be done twice. However, I'm not sure if I would prefer the Union or the INtersection. Thoughts?

scott

On Aug 27, 2014, at 10:00 AM, Cannon, Steven <Steven.Cannon@ARS.USDA.GOV> wrote:

> Hello again -
>
> At the cost of adding to everyones' in-boxes again, I thought I would share a response to this question about the gene models:
> "The gene models are predicted using MAKER and GLEAN and as we understand there should be a final gene model which should be based on a set of most confident common genes predicted by these two pipelines."
>
> Here is my response, for consideration. Possibly someone in the group has a better alternative. And if not, then we'll have this as a consensus practice for analyses of the gene models …
>
> ==========
> No one in the Consortium has made a single unified gene set. It turns out this is not trivial. It looks like there is good stuff in each annotation set that the other misses (analysis from Sudhansu below**). A few possible approaches:
>
> 1) We could take the intersection (sort of) by designating one set (e.g. MAKER) as primary and discarding all genes in the primary set that don't overlap with the secondary set. This would get rid of some probable false positives but would also get rid of some true positives.
>
> 2) We could take the union (sort of) by designating one set as primary and keeping all of those genes and adding those from the secondary set that don't overlap with the primary set. This would result in a hybrid annotation set, composed of genes called by two different methods.

3) Live with two sets of gene models. Conduct down-stream analyses on both (MAKER and GLEAN). It might be useful (at the expense of more stuff for a reader to ingest) to report results on the intersection with respect to each (i.e. MAKER with overlap by GLEAN and GLEAN with overlap with MAKER).

To me the third option seems the best. We could provide the "intersection" lists as a supplement though (MAKER models with overlap by GLEAN and GLEAN models with overlap with MAKER).

So I think what I would suggest is: do the analyses on each gene set above. We will provide the intersection lists that you can apply after the fact if you want to explore the arguably higher-confidence genes (e.g. if you have result Y on 1000 MAKER genes, you have result Y' on the 665 MAKER genes that overlap with GLEAN genes) - but this shouldn't hold up current analyses.

- Steven

** Repeating some analysis from Sudhansu (this is prior to the current [Aug 24] set):
Approximately two thirds of the MAKER and GLEAN gene models correspond (overlap). Specifically …
- In Aradu 26533 GLEAN genes (out of 37842 total, 70%) overlap with 25203 MAKER genes (out of 38149 total, 66%) in 27099 cases of overlap.
- In Araip 26910 GLEAN mRNA models (out of 39303 total, 68.5%) overlap with 25738 MAKER gene models (out of 42883 total, 60%) in 27629 cases of overlap.

Each method seems to provide additional apparently "real" that the other misses. Specifically …
- There are significant numbers of "good" genes (with high AHRD scores) in both the "MAKER-only" and "GLEAN-only" sets. AHRD assigns a "quality" score ranging from 0 to 4 stars, with 4 being the best. In Aradu, about 2.5% of the MAKER only genes (331/12946) are 4-star and about 3.5% (393/11309) of the GLEAN-only genes are 4-star.

There are also frequently differences in gene fragmentation and structure (informal observation from looking at the browser).
==========

From: <Varshney>, "Varshney, Rajeev (ICRISAT-IN)"

<r.k.varshney@cgiar.org>
Date: Tuesday, August 26, 2014 at 9:40 PM
To: Steven Cannon <steven.cannon@ars.usda.gov>, "Lutz Froenicke (lfroenicke@ucdavis.edu)" <lfroenicke@ucdavis.edu>, Jeremy Schmutz <jschmutz@hudsonalpha.org>, "Scheffler, Brian" <Brian.Scheffler@ARS.USDA.GOV>, Richard Michelmore <rwmichelmore@ucdavis.edu>, "Ozias-Akins, Peggy (GCP)" <pozias@uga.edu>, David Bertioli <djbertioli@gmail.com>, Howard Valentine <hvalentine@peanutsusa.com>, Scott Jackson <sjackson@uga.edu>, Xin LIU <liuxin@genomics.org.cn>, "mburow@tamu.edu" <mburow@tamu.edu>, Soraya Cristina De M Leal Bertioli <Soraya.Bertioli@embrapa.br>, David Bertioli <david.bertioli@pq.cnpq.br>, "Schnell, Ray" <Ray.Schnell@effem.com>, Victor Nwosu <victor.nwosu@effem.com>, Rich Wilson <rfwilson@mindspring.com>, Brian Abernathy <bla@uga.edu>, "Guo, Baozhu" <Baozhu.Guo@ARS.USDA.GOV>, "Holbrook, Corley" <Corley.Holbrook@ARS.USDA.GOV>, Ran Hovav <ranh@volcani.agri.gov.il>, Sudhansu Dash <sdash@iastate.edu>, Ethalinda Cannon <ekcannon@iastate.edu>, Nathan Weeks <weeks@iastate.edu>, Andrew Farmer <adf@ncgr.org>, Longhui Ren <lhren@iastate.edu>, "Huang, Wei [AGRON]" <weih@iastate.edu>
Cc: "Kale, Sandip (ICRISAT-IN)" <S.Kale@cgiar.org>, "Pandey, Manish K (ICRISAT-IN)" <M.Pandey@cgiar.org>, "Agarwal, Gaurav (ICRISAT-IN)" <Gaurav.Agarwal@cgiar.org>
Subject: Re: Diploid annotations for review - revisions, updates

Thanks very much, Steven. I agree with your approach and suggestion.

Dear all: I would like to use this message to update you all on the other activity that we have undertaken. With an objective to identify the markers easily assayable for breeding applications, we have also searched genomes (AA and BB) for insertion and deletions. Please see the message below and attached file. We are in process of identify chromosome/ genome specific indel markers that can be easily scored for genetics and breeding applications.

Again this stuff can be included in Marker section (along with SSRs) in the genome MS. We can have a few supplementary tables on statistics as well as primer sequence.

Many thanks and kind regards

Rajeev

Begin forwarded message:

From: sandip kale <sandipmkale@gmail.com>
Subject: InDel identification between A and B genomes of peanut
Date: August 25, 2014 at 4:53:13 PM GMT+5:30
To: Manish Pandey <manishgenetics@gmail.com>
Cc: Gaurav Agarwal <gaurav.iari@gmail.com>, "Varshney, Rajeev (ICRISAT-IN)" <R.K.Varshney@CGIAR.ORG>

Hello Sir,

Attached herewith the summary of InDels identified between peanut A and B genomes.

The following steps were used for InDel identification

1.The MUGSY software was used for Indel identification assuming these two genomes are closely related
2. The output was parsed to obtain InDel sizes using perl script (the perl script was procured from Dr. Jingjing)

Total 1045015 insertions and 953715 deletions were obtained, out of which, 269974 insertions and 245250 deletions were present on same chromosomes of A and B genomes while rest were present on different chromosomes.

	Total	Same_chromosome	Different_chromosome
Insertions	1045015	269974	775041
Deletions	953715	245250	708465

We would like to discuss the results and further plans with Rajeev sir and you

Kindly let us know the your availability

Thanking you

With best

Sandip and Gaurav

On Aug 27, 2014, at 4:17 AM, Cannon, Steven <Steven.Cannon@ARS.USDA.GOV> wrote:

> PGC group,
>
> Another update. As Ethy was submitting scaffolds to GenBank (still underway; waiting for some information from BGI, and then more QC), we learned of several redundant scaffolds — that is, scaffolds present in both the pseudomolecules and in the remaining "unplaced" scaffold set, for both of the species. There were eleven redundant scaffolds in A. duranensis and in A. ipaensis.
>
> Beyond these twelve scaffolds, we identified a number of other very low-quality scaffolds in the "unplaced" set, which I would also like to remove. These have no genes, and contain < 2000 bases of non-N sequence or are > 80% Ns.
>
> Although it would be OK for the GenBank submission to diverge from the 1.0 assembly (leading eventually to a new assembly version), I think it will be best to correct (remove) these duplicated scaffolds from the "unplaced" scaffolds in the current assembly. The change doesn't affect any of the pseudomolecules. My strong preference is therefore to leave the overall assembly version as 1.0, but to make a dated update to the "all-scaffolds" file and to the files with unplaced scaffolds.

A small additional complication is that the scaffold removals will also affect the gene models (903 MAKER genes in A. duranensis and 354 MAKER genes in A. ipaensis, and similar changes for GLEAN). However, I think this also should be OK, since we (this group) should be the only set of people who have the gene models (i.e. they haven't been made public yet).

Analyses that are sensitive to total assembly sequence or the full gene set may need to be re-run (although the total sequence changing will be less than 0.1%, and no changes in the pseudomolecules). I would not expect aggregate or summary statistics to change.

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

Please let us know if you have any questions or concerns, or if you find any problems.

Steven

From: <Cannon>, Steven Cannon <steven.cannon@ars.usda.gov>
Date: Monday, July 21, 2014 at 5:13 PM
To: Lutz Froenicke <lfroenicke@ucdavis.edu>, Jeremy Schmutz <jschmutz@hudsonalpha.org>, "Scheffler, Brian" <Brian.Scheffler@ARS.USDA.GOV>, "<rwmichelmore@ucdavis.edu>" <rwmichelmore@ucdavis.edu>, Peggy Ozias-Akins <pozias@uga.edu>, David Bertioli <djbertioli@gmail.com>, Howard Valentine <hvalentine@peanutsusa.com>, "Jackson, Scott" <sjackson@uga.edu>, "Michelmore, Richard" <rwmichelmore@ucdavis.edu>, "Liu, Xin" <liuxin@genomics.org.cn>, "mburow@tamu.edu" <mburow@tamu.edu>, "Bertioli, Soraya" <Soraya.Bertioli@embrapa.br>, "Bertioli, David" <david.bertioli@pq.cnpq.br>, "Schnell, Ray" <Ray.Schnell@effem.com>, "Nwosu, Victor" <victor.nwosu@effem.com>, Richard Wilson <rfwilson@mindspring.com>, "Varshney, Rajeev (ICRISAT-IN)" <r.k.varshney@cgiar.org>, Brian Abernathy <bla@uga.edu>, "Guo, Baozhu" <Baozhu.Guo@ARS.USDA.GOV>,

"Holbrook, Corley" <Corley.Holbrook@ARS.USDA.GOV>, Ran Hovav <ranh@volcani.agri.gov.il>
Cc: Sudhansu Dash <sdash@iastate.edu>, Ethalinda Cannon <ekcannon@iastate.edu>, Nathan Weeks <weeks@iastate.edu>, Andrew Farmer <adf@ncgr.org>, Longhui Ren <lhren@iastate.edu>, "Huang, Wei [AGRON]" <weih@iastate.edu>, Steven Cannon <steven.cannon@ars.usda.gov>
Subject: Diploid annotations for review - revisions, updates

PGC group,

Here is an update on annotation work from the IA and NM groups.

Briefly:
- Some MAKER gene models have been demoted
- New functional descriptions for both MAKER and GLEAN models
- New browser tracks, with functional descriptions
- New analysis/comparison of the MAKER and GLEAN (BGI) models …
-… which suggests (I think) that both the MAKER and GLEAN models should be used/promoted
- Question for the group about when and how to release these (i.e. remove password protection).
Any accompanying news releases? Time this release with any events (perhaps the November meeting)?

==========
What's new:

- In the MAKER annotation, we "demoted" 1164 Aradu and 1553 Araip genes to the "lowqual_or_TE" set. This is
Longhui Ren's work here, resulting from sleuthing to find why there were more Araip models than for Aradu.
It seems that – as David B has mentioned – there has been a greater proliferation and diversification of transposable
elements in A. ipaensis. The repeat masking in the MAKER gene modeling was apparently not as

thorough or
as unbiased as for the GLEAN (BGI) masking, so MAKER picked up some additional “genes” in A. ipaensis.
Even so, there are some apparently significant differences in the gene complements of the two genomes.

- For both the MAKER and GLEAN annotations, the functional descriptions have been updated, in the *.AHRD.* files.
These annotations were gathered using the "Automated Assignment of Human Readable Descriptions”
(AHRD) tool, by Andrew Farmer, using the following search targets: Arabidopsis v10, Medicago v4.0,
soybean v. Glyma.Wm82.a2.v1, and InterProScan 5.3-46.0 (targeting UniProt90 (2014) and Gene Ontology (2014)).
There are two variants of the *.AHRD.* files: the full AHRD results in *.AHRD.csv, and an abbreviated,
two-column format in *.AHRD.slim .

- The README files have been updated accordingly.

- The browser tracks for the MAKER and GLEAN models now have functional descriptions.

(The links below are ones you have seen before, but the contents have been updated. If you are working with the annotation sets below (*.tar.gz), please download these again. Although the gene models are the same, the annotations are new, as is the separation into “good” vs. “demoted” sets.)

[redacted]

[redacted]

==========

Some analysis about the similarities and differences between MAKER and GLEAN (Sudhansu):

Approximately two thirds of the MAKER and GLEAN gene models correspond (overlap). Specifically …

- In Aradu 26533 GLEAN genes (out of 37842 total, 70%) overlap with 25203 MAKER genes (out of 38149 total, 66%) in 27099 cases of overlap.
- In Araip 26910 GLEAN mRNA models (out of 39303 total, 68.5%) overlap with 25738 MAKER gene models (out of 42883 total, 60%) in 27629 cases of overlap.

Each method seems to provide additional apparently “real” that the other misses. Specifically …

- There are significant numbers of “good” genes (with high AHRD scores) in both the “MAKER-only” and “GLEAN-only” sets. AHRD assigns a “quality” score ranging from 0 to 4 stars, with 4 being the best. In Aradu, about 2.5% of the MAKER only genes (331/12946) are 4-star and about 3.5% (393/11309) of the GLEAN-only genes are 4-star.

There are also frequently differences in gene fragmentation and structure (informal observation from looking at the browser).

==========

Please let us know if have any questions or spot any problems!

Steven and group

<peanut_gene_model_evaluation.docx>