Whole-Genus Seuqencing: 300 Aspergilli

Alan Kuo[*1], Robert Riley[1], Alicia Clum[1], Asaf Salamov[1], Scott Baker[2], Blake Simmons[2], Mikael Andersen[3], Igor Grigoriev[1]

[1]DOE Joint Genome Inst., Walnut Creek, CA,
[2]Joint BioEnergy Inst., Emeryville, CA
[3]Technical U. of Denmark, Kgs. Lyngby, Denmark

March 2015

The work conducted by the U.S. Department of Energy Joint Genome Institute is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231

DISCLAIMER

Whole-Genus Sequencing: 300 Aspergilli

Alan Kuo*[1], Robert Riley[1], Alicia Clum[1], Asaf Salamov[1], Scott Baker[2], Blake Simmons[2], Mikael Andersen[3], Igor Grigoriev[1]

*email: akuo@lbl.gov, [1]DOE Joint Genome Inst., Walnut Creek, CA, [2]Joint BioEnergy Inst., Emeryville, CA, [3]Technical U. of Denmark, Kgs. Lyngby, Denmark

DOE JOINT GENOME INSTITUTE
US DEPARTMENT OF ENERGY
OFFICE OF SCIENCE

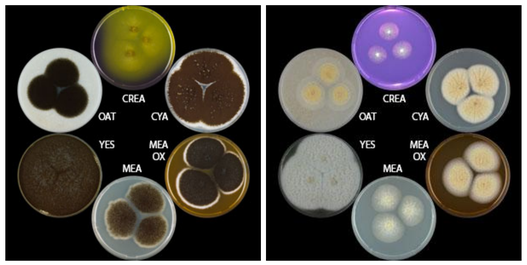
ABSTRACT

***Aspergillus* is a ubiquitous and phenotypically diverse genus of filamentous Ascomycota, many of which play key roles as fermenters in food production, platforms for biotechnology and industrial production of enzymes and chemicals, plant and opportunistic animal pathogens, and agents of agricultural toxigenesis and biomass conversion for bioenergy. As part of a DOE Joint BioEnergy Institute initiative to characterize the entire genus, the JGI plans to sequence, assemble, and annotate the genomes of each of the ~300 species of the genus *Aspergillus*. To accomplish this massive task in a timely manner without sacrificing quality, we have sought to streamline our existing processes as well as explore alternative technologies, especially assembly and annotation of long PacBio sequencing reads. Over the past year we have released on MycoCosm the genomes of an additional 24 *Aspergillus* sp. with preliminary analyses of their phylogenies, secretomes, and secondary metabolism. The next tranche of 92 species is expected soon.**

Why whole-genus sequencing?

The 300+ species of Aspergilli have been identified and classified based on morphology, and on profiles of secondary metabolites, of which *Aspergillus* sp. produce a greater variety than any other fungi.

Fig. 1: Two of the 24 species used in this study, each with distinct morphology and secondary metabolism on 6 standard growth media.



A. eucalypticola *A. steynii*

Advances in genomics allow increasing numbers of aspergilli to be analyzed, and now we are in the process of sequencing every species in the genus. An entirely-sequenced genus will provide a comprehensive platform for taxonomy and phylogeny, as well as functional studies including but not limited to secreted proteins and secondary metabolites.

To optimize sequencing strategies for such a grand scale, JGI is exploring ways to streamline our standard processes. We are testing on some of the first 24 *Aspergillus* sp. we have received for this project. We can also provide some preliminary results for these 24 genomes.

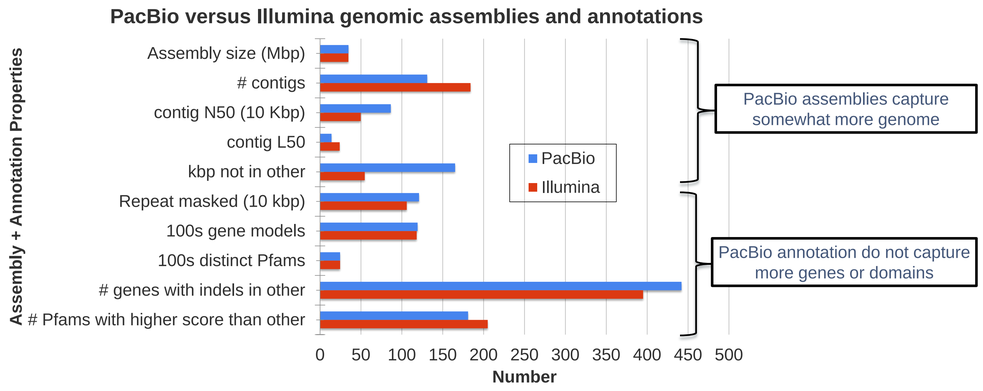
Credits

Ellen Lyhne (DTU)
cultures, photos
Martin Kogle (DTU)
DNAs, RNAs
Matt Nolan (JGI)
genome assembly
Anna Lipzen (JGI)
transcriptome assembly
Frank Korzeniewski (JGI)
SMC pipeline
Roman Nikitin (JGI)
SMC portal

Explore alternative sequencing technology

For 6 *Aspergillus* sp. we did both PacBio sequencing + assembly, as well as JGI standard Illumina sequencing + assembly, followed by the standard JGI Annotation Pipeline. We then compared the resulting assemblies and annotations.

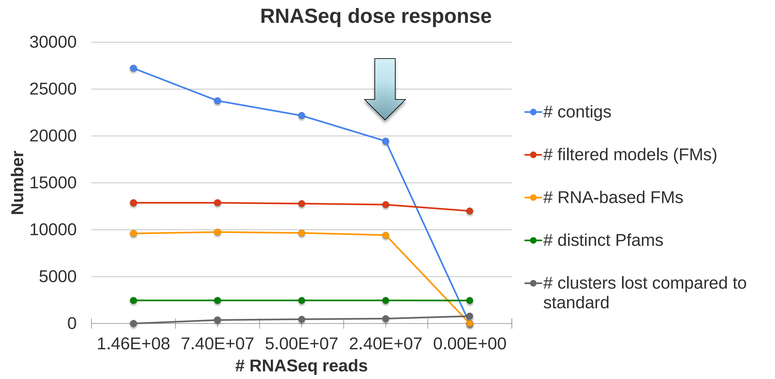
Fig. 2: Representative results with *A. eucalypticola*. PacBio did not provide a significant advantage over Illumina,



Explore reductions in RNASeq

For 4 *Aspergillus* sp. we reduced the number of RNASeq reads used for JGI standard transcriptomic assembly and for JGI Annotation Pipeline. We then compared the resulting annotations.

Fig. 3: Representative results *with A. steynii*. Quality began to fall at 1/6th the standard no. of reads.



Preliminary results

Fig. 4A: ML species tree using 2k conserved genes. Most subgenera are confirmed.

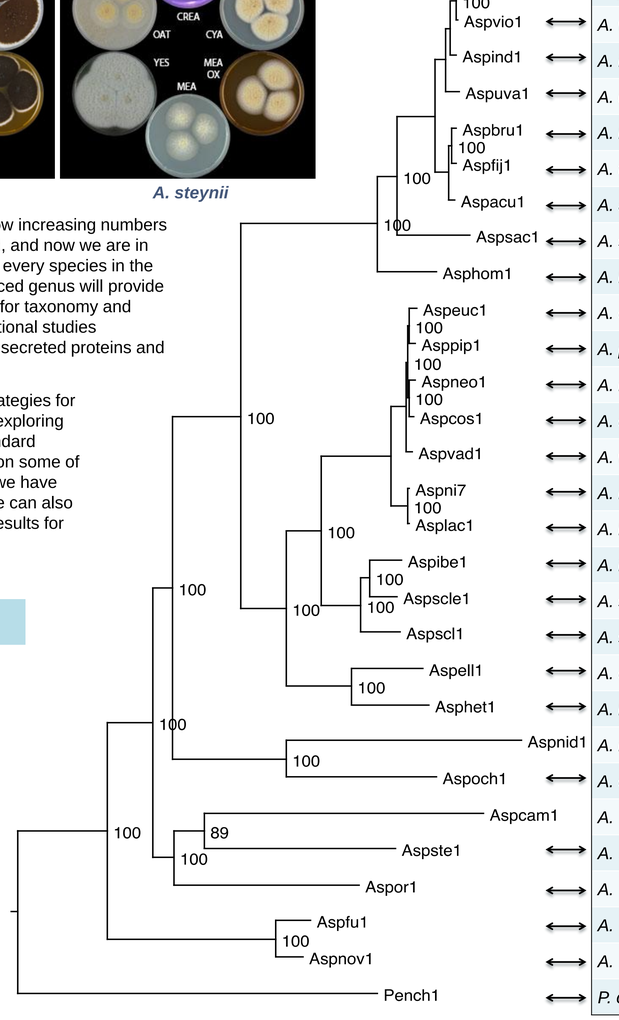


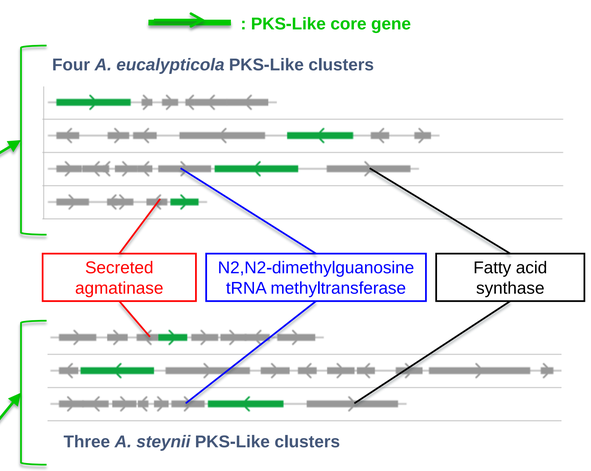
Fig. 4B: Predicted gene complements and secreted proteins.

Species	Genome (Mbp)	1000s genes	# secreted proteins
A. japonicus	36.1	12.0	1374
A. violaceofuscus	36.0	12.1	1377
A. indologenus	38.6	12.1	1382
A. uvarum	35.9	12.0	1367
A. brunneoviolaceus	37.5	12.1	1396
A. fijiensis	36.5	12.0	1366
A. aculeatinus	36.5	12.0	1387
A. saccharolyticus	31.1	10.1	1061
A. homomorphus	34.1	11.4	1205
A. eucalypticola	34.8	11.9	1310
A. piperis	35.3	12.1	1373
A. neoniger	35.4	11.9	1318
A. costaricaensis	36.9	12.0	1330
A. vadensis	35.7	12.1	1342
A. niger	34.9	11.9	1300
A. lacticoffeatus	35.9	13.1	1388
A. ibericus	33.4	11.7	1321
A. sclerotiicarbonarius	37.6	12.6	1401
A. sclerotioniger	36.7	12.3	1295
A. ellipticus	42.9	12.9	1351
A. heteromorphus	35.6	11.1	1183
A. nidulans	30.5	10.7	1142
A. ochraceoroseus	27.7	8.9	839
A. campestris	28.3	9.8	1058
A. steynii	37.8	13.2	1400
A. oryzae	37.9	12.0	1333
A. fumigatus	29.4	9.8	1026
A. novofumigatus	32.4	11.5	1299
P. chrysogenum	31.3	11.4	1083

Fig. 4C: Predicted secondary metabolite clusters (SMCs). Statistically significant gains are blue, losses are red.

PKS	PKS-Like	NRPS	NRPS-Like	HYBRID	DMAT	TC
19	3	21	22	1	4	1
17	4	19	20	3	6	3
27	3	17	24	5	6	2
21	3	17	21	3	4	2
24	3	21	21	2	4	4
26	3	21	25	3	5	4
28	3	19	20	2	6	3
11	6	14	15	2	3	0
24	5	16	22	3	4	4
26	4	11	19	6	2	6
30	5	14	20	6	2	8
29	6	13	16	4	3	9
32	7	17	21	5	3	7
27	5	13	22	5	2	4
27	4	16	16	9	2	6
23	6	16	16	9	2	7
15	6	7	18	6	2	3
25	6	14	22	6	3	3
21	5	9	21	5	3	5
28	7	15	14	2	2	4
16	3	12	16	3	4	3
22	5	9	13	1	5	1
10	2	9	3	2	4	1
12	2	11	11	5	5	2
29	3	17	22	9	9	4
23	5	14	15	2	9	2
14	2	9	7	0	2	0
24	2	16	5	3	4	2
18	3	9	14	2	0	4

Fig. 4D: Some SMCs are conserved across the genus



Conclusions, so far

Methodological optimization, so far

- PacBio sequencing of *Aspergillus* genomes was not substantially better than Illumina.
- RNASeq was reduced 6-fold without adverse effects on *Aspergiillus* annotation.
- SMC pipeline and portal were developed and tested.

Biological results, so far

- 24 *Aspergillus* sp. across multiple subgenera were assembled and annotated.
- Tentative phylogeny was inferred using whole-genome methods.
- Secreted proteins were identified.
- SMCs were cataloged and clade-specific gains and losses were identified.

What's next

More streamlining of methods

- One example: test pipelines with fewer gene predictors.

More analyses

- Target other functionally significant proteins: proteases, transporters, CAZymes.
- Identify gains and losses in the above categories of genes

More genomes

- Next tranche of 92 species is in sequencing stage.

The work conducted by the U.S. Department of Energy Joint Genome Institute is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.