Advancing Science with DNA Sequence

JGI
DOE JOINT GENOME INSTITUTE
US DEPARTMENT OF ENERGY
OFFICE OF SCIENCE

Challenges in whole-genome annotation of pyrosequenced eukaryotic genomes

3rd IBC, April 17, 2009

Alan Kuo* and Igor Grigoriev
DOE Joint Genome Institute
*akuo@lbl.gov

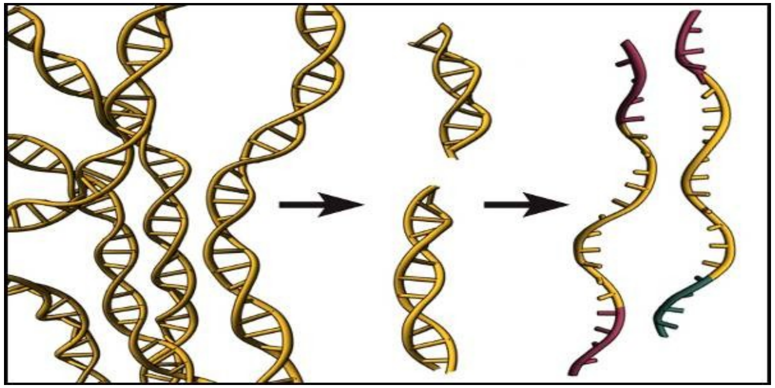
This work was performed under the auspices of the US Department of Energy's Office of Science, Biological and Environmental Research Program, and by the University of California, Lawrence Berkeley National Laboratory under contract No. DE-AC02-05CH11231, Lawrence Livermore National Laboratory under Contract No. DE-AC52-07NA27344, and Los Alamos National Laboratory under contract No. DE-AC02-06NA25396.

Introduction

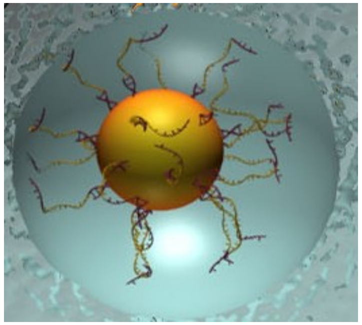
- Pyrosequencing technologies such as 454 and Solexa sequence DNA at much higher rate and lower cost than traditional Sanger technology.
- 454 is now mature enough to be used for **eukaryotic** genome sequencing and assembly.
- What will be the effect on **annotation**??

- A simple experiment to assess assemblies that use 454 reads.
- Successful production annotation of 2 assemblies that use 454.

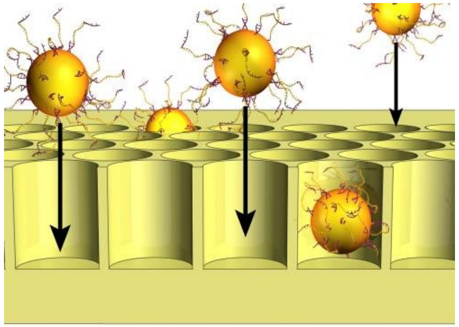
454 technology



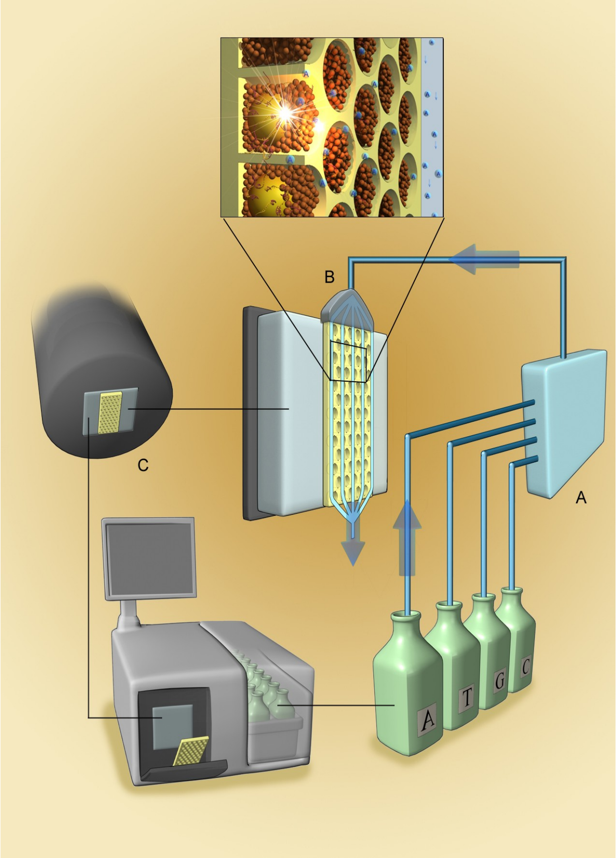
1. prepare adapter-ligated ssDNA library



2. clonally amplify on 28 µm beads



3. Load beads and enzymes in PicoTiterPlate™



4. Sequence by synthesis on the 454 Instrument

Images from Stephen Kingsmore, NCGR

454 vs. Sanger

	Sanger	454
Mbp per run	0.3	100
US$ per kbp	$1.0	$0.1
Read length (nt)	800	240
Paired ends distance (kb)	40	3
Error rate (%)	0.1	0.5

High coverage and depth of coverage

Poor assembly of repetitive regions

Many small gaps

Frameshifts genes

Data from Stephen Kingsmore, NCGR

Homopolymer stutter

Real sequence

CGCA**CCCCC**TCATATAAG

R T P S Y K

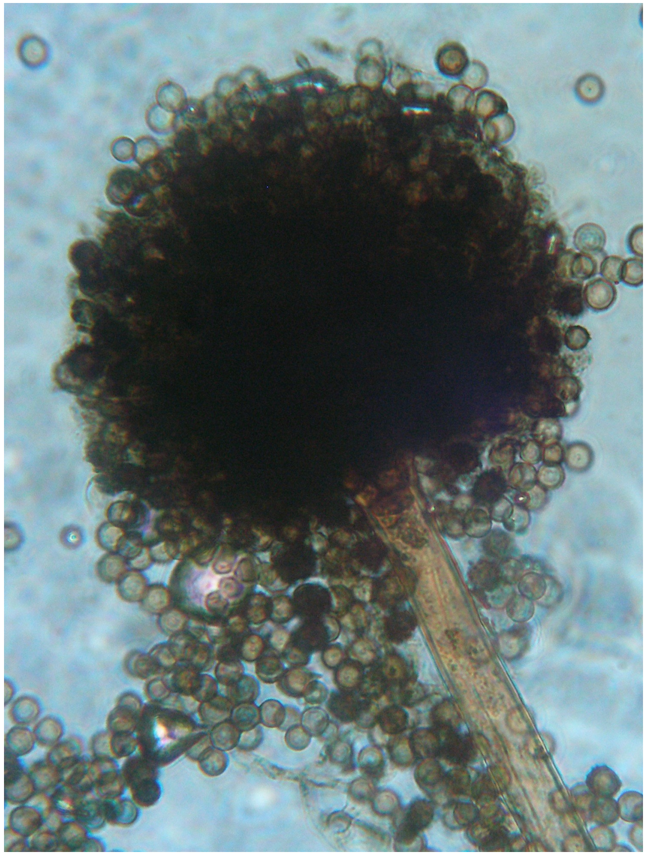
454 error

CGCA**CCCCCC**TCATATAAG

R T P **L** **I** *****

What is the effect of the stutter on automatic annotation?

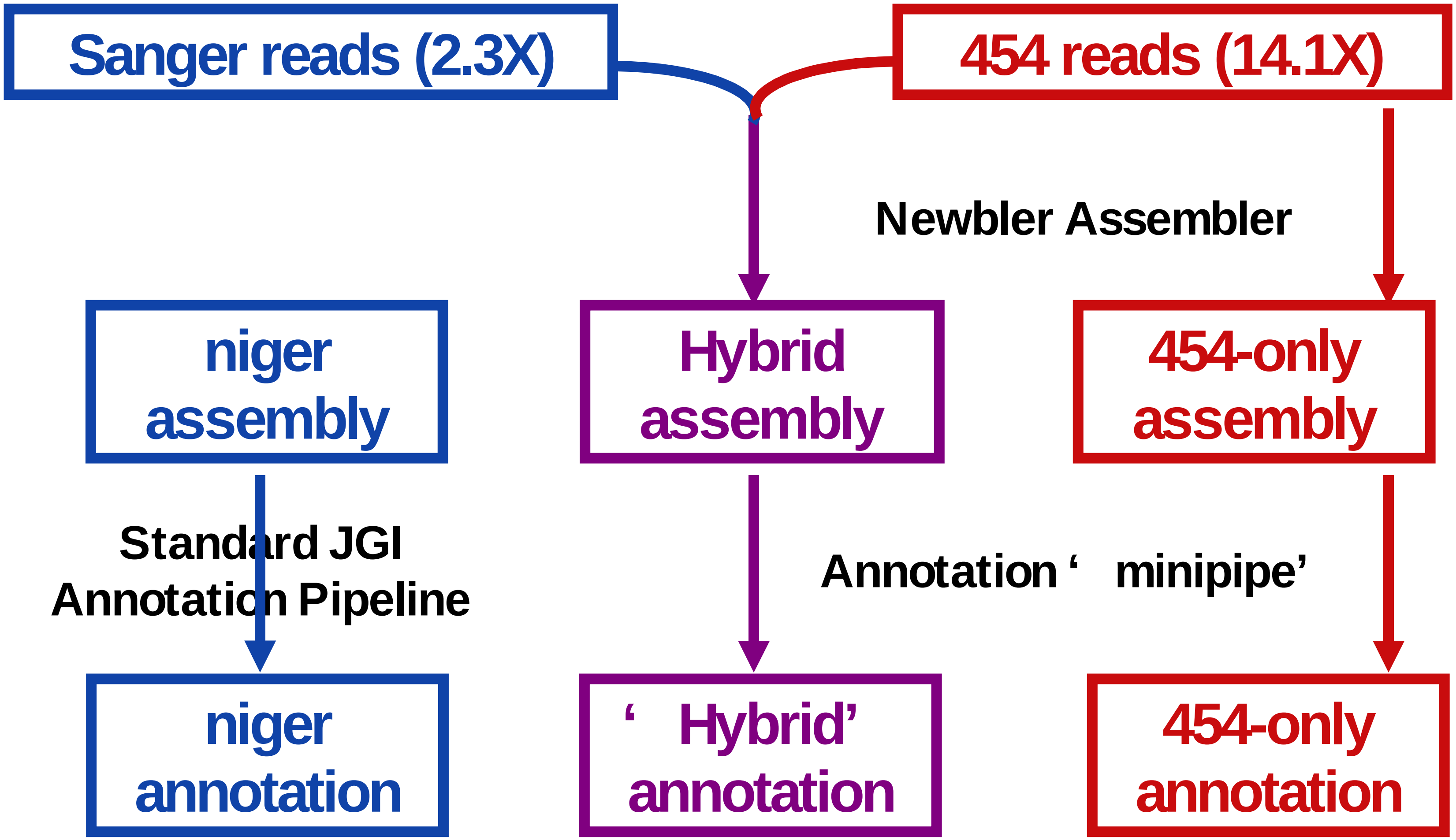
The test bed



- *Aspergillus carbonarius*
- Ascomycote fungus
- Small (< 40Mbp)
- Haploid
- Well-known close relative: *Aspergillus niger* genome sequenced and annotated by JGI 2006

Photo from Scott Baker, PNNL

Experimental design



Sanger reads (2.3X)

454 reads (14.1X)

Newbler Assembler

niger assembly

Hybrid assembly

454-only assembly

Standard JGI Annotation Pipeline

Annotation ‘ minipipe’

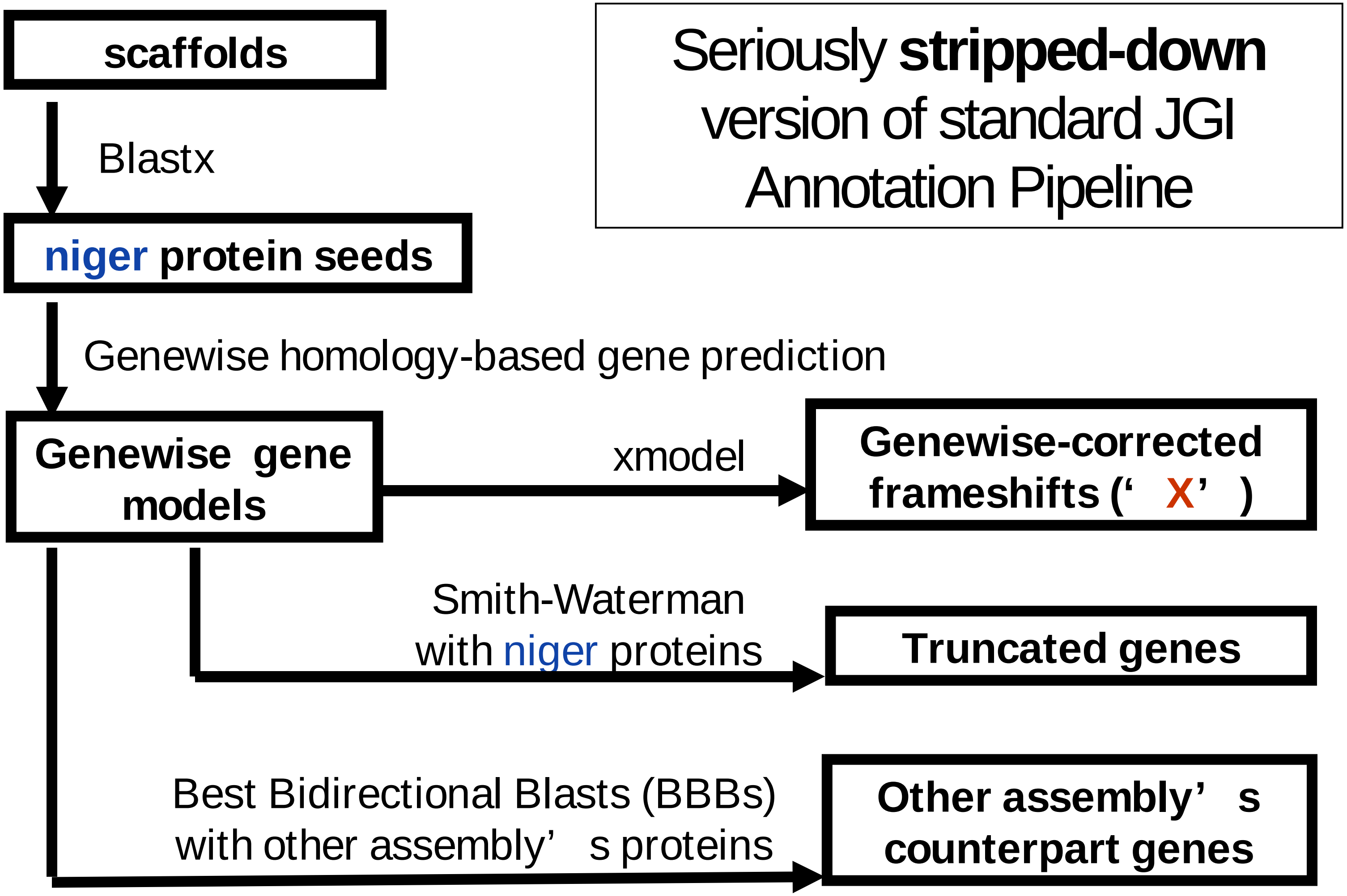
niger annotation

‘ Hybrid’ annotation

454-only annotation

What is a ‘ minipipe’ ?

Seriously **stripped-down** version of standard JGI Annotation Pipeline



scaffolds

Blastx

niger protein seeds

Genewise homology-based gene prediction

Genewise gene models

xmodel

Genewise-corrected frameshifts (‘ X’)

Smith-Waterman with niger proteins

Truncated genes

Best Bidirectional Blasts (BBBs) with other assembly’ s proteins

Other assembly’ s counterpart genes

Aspergillus assemblies

	niger	Hybrid	454-only
Assembly size (Mbp)	37.2	34.9	32.2
# scaffolds	143	873	78
N50/L50 (# / Mbp)	6 / 2.0	8 / 1.8	10 / 0.9
Total gap space (Mbp)	2.4 (6%)	2.5 (7%)	0.5 (2%)
# gaps		556	1482
Ave. gap size (nt)		4420	367
Std. dev. gap size (nt)		6068	294

minipipe results

	Hybrid	454-only
# Genewise models	9730	9595
Model density (# / Mbp)	279	297
Models with 'X' (frameshift)	1048 (11%)	1161 (12%)
# aligned niger proteins	10494 (94%)	10406 (93%)
# niger proteins < 80% covered (truncated)	2710 (16%)	4115 (27%)

Sanger fixes 454 errors

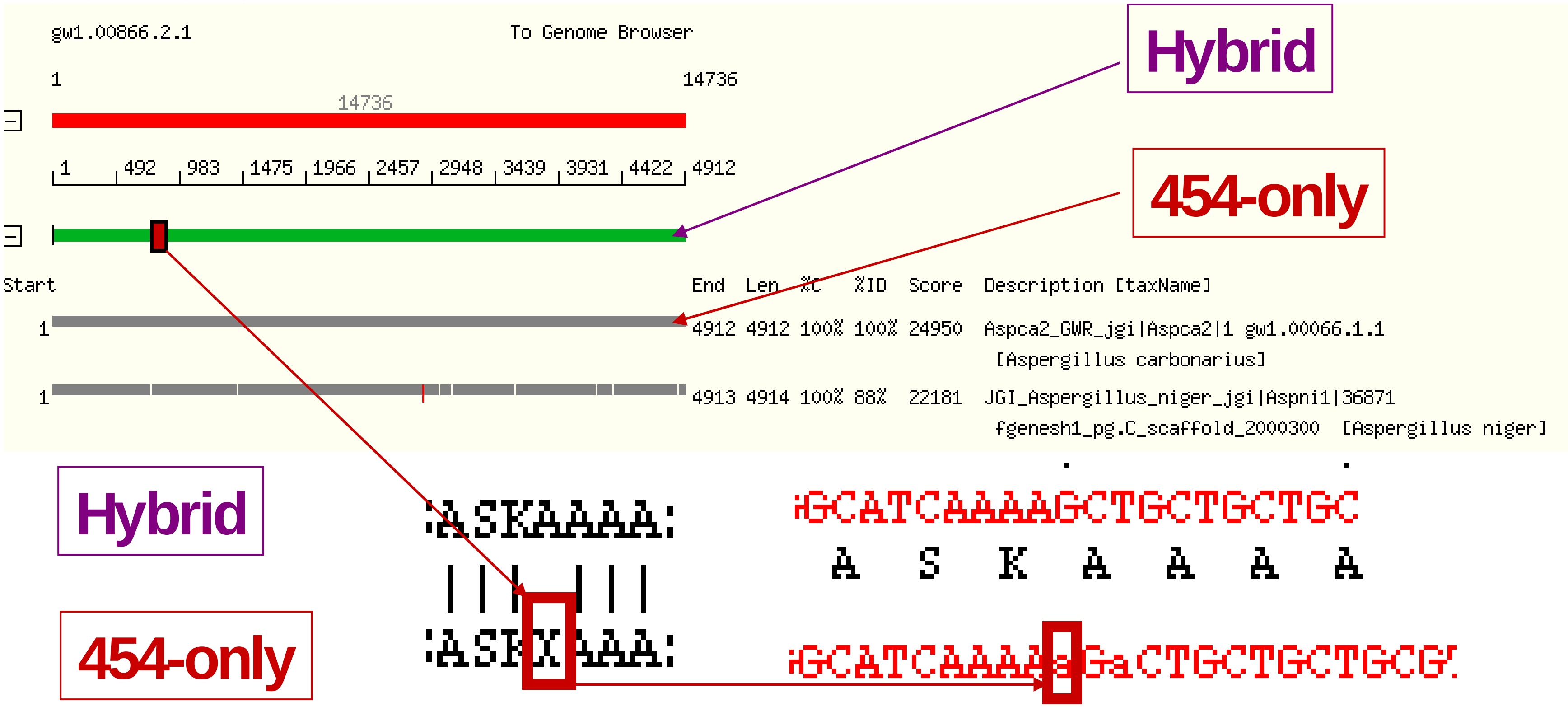
BBBs between the 2 annotations		454-only	
		w/o ‘ X’	w/ ‘ X’
Hybrid	w/o ‘ X’	8359	238
	w/ ‘ X’	34	912

Genes w/ IS corrected by Sanger

Uncorrected genes + real pseudogenes

An error, but not in the hybrid

JGI Portal view



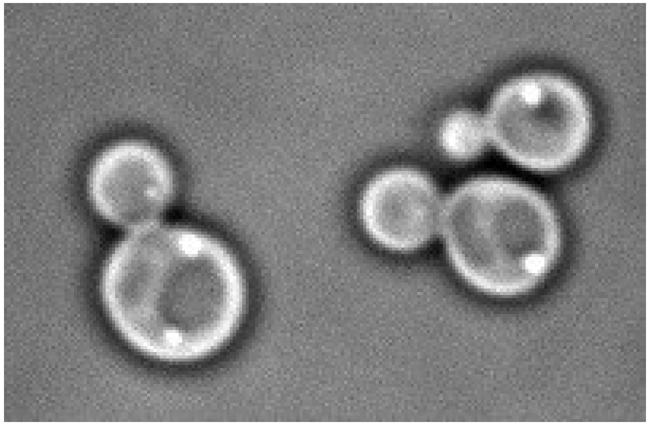
```
gw1.00866.2.1                              To Genome Browser

1                                                              14736
                              14736

 1    492   983   1475  1966  2457  2948  3439  3931  4422  4912

Start                                                          End  Len  %C   %ID  Score  Description [taxName]
    1                                                          4912 4912 100% 100% 24950  Aspca2_GWR_jgi|Aspca2|1 gw1.00066.1.1
                                                                                           [Aspergillus carbonarius]
    1                                                          4913 4914 100% 88%  22181  JGI_Aspergillus_niger_jgi|Aspni1|36871
                                                                                           fgenesh1_pg.C_scaffold_2000300  [Aspergillus niger]
```

protein view

gene view

Production annotation of hybrid assemblies

- Yeasts *Candida tenuis* and *Spathaspora passalidarum*
- do xylose -> ethanol
- Tiny haploid genomes, few introns
- Well-known close relative: *Pichia stipitis* genome released by JGI 2006



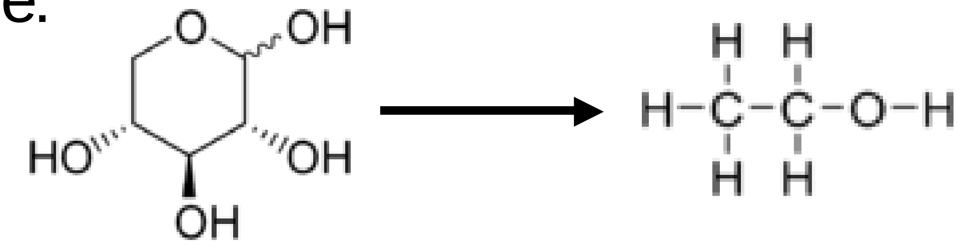


Photo from Dana Wohlbach, U. WI

454 vs. Sanger, round 2

	Sanger	Old 454	New 454
Mbp per run	0.3	100	450
US$ per kbp	$1.00	$0.10	$0.02
Read length (nt)	800	240	450
Paired ends distance (kb)	40	3	20
Error rate (%)	0.1	0.5	??

Data from Joann Mudge, NCGR

DOE JOINT GENOME INSTITUTE
US DEPARTMENT OF ENERGY
OFFICE OF SCIENCE

Quality of yeast assemblies

	Pichia	Spatha	tenuis
Assembly size (Mbp)	15.4	13.3	10.7
# scaffolds	9	47	25
N50/L50 (# / Mbp)	4 / 1.8	4 / 1.7	3 / 1.2
Total gap space (Mbp)	0.0 (0%)	0.3 (2%)	0.2 (2%)
‘X’ rate (frameshifted)	2.0%	3.4%	2.4%
% Pichia proteins aligned		95.7%	94.6%
% Pichia aligned proteins <80% covered (truncated)		3.2%	5.8%

Production yeast annotations

	Pichia	Spatha	tenuis
# genes	5841	5726	5452
Gene density (# / Mbp)	378	431	507
Avg. gene length (nt)	1627	1472	1459
Avg. protein length (aa)	493	478	477
# exons / gene	1.4	1.3	1.2
% genes w/ Pfam	62.4	71.3	73.4
% genes w/ SwissProt	88.3	91.9	92.3

Unexceptional -- A GOOD THING!

Conclusion

- 454 techonology poses challenges to both assembly and annotation.
- Hybrid assembly helps resolve many of these challenges, including correction of many 454 sequence errors.
- The JGI Annotation Pipeline successfully annotated 2 yeasts of bioenergy significance.
- Hybrid assemblies of small eukaryotic genomes can be suitable substrates for production annotation.

Acknowledgments

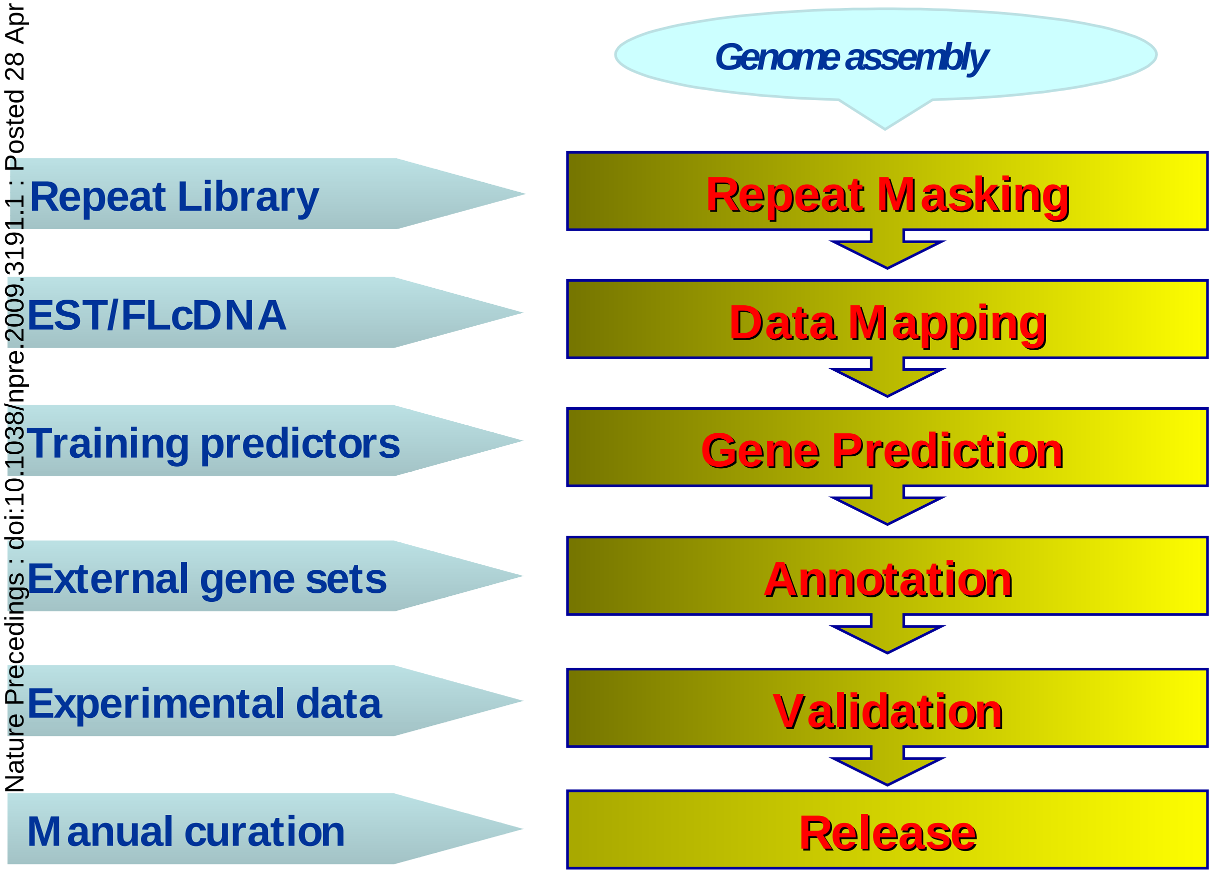
JGI

- **Assembly**
  - **Alla Lapidus**
  - **Brian Foster**
- **Annotation**
  - **Andrea Aerts**
  - **Asaf Salamov**
  - **Frank Korzeniewski**
  - **Xueling Zhao**
- **Informatics staff**

Our collaborators

- Aspergillus
  - Scott Baker
  - Giancarlo Perrone
- Yeasts
  - Audrey Gasch
  - Dana Wohlbach
  - Tom Jeffries
- NCGR
  - Stephen Kingsmore
  - Joann Mudge

JGI Annotation Pipeline



Genome assembly

Repeat Library → Repeat Masking

EST/FLcDNA → Data Mapping

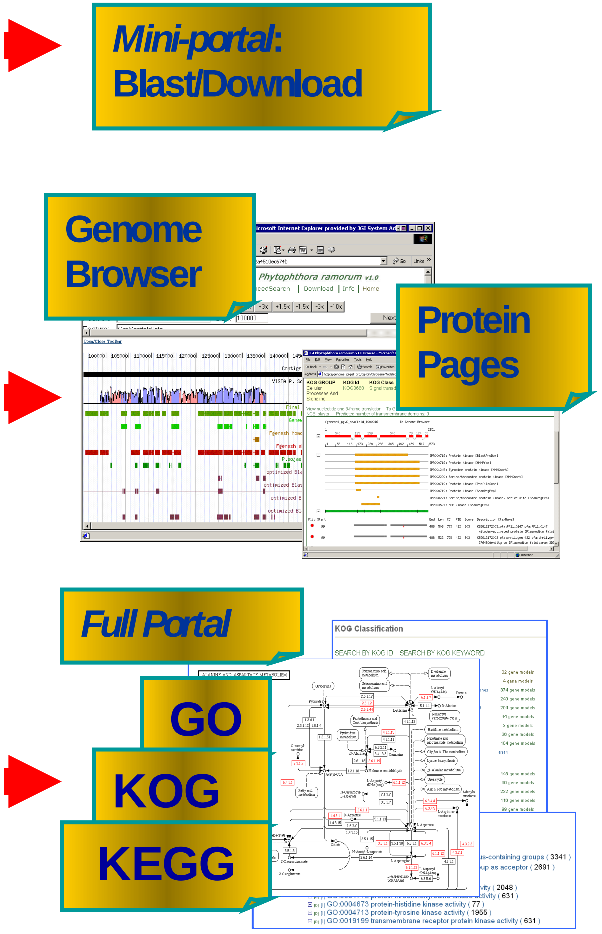
Training predictors → Gene Prediction

External gene sets → Annotation

Experimental data → Validation

Manual curation → Release

Automated Pipeline



Mini-portal: Blast/Download

Genome Browser

Protein Pages

Full Portal

GO

KOG

KEGG

Phytophthora ramorum v1.0

KOG Classification

SEARCH BY KOG ID SEARCH BY KOG KEYWORD

GO:0004673 protein-histidine kinase activity (77)

GO:0004713 protein-tyrosine kinase activity (1955)

GO:0019199 transmembrane receptor protein kinase activity (631)

Staged Portal Release