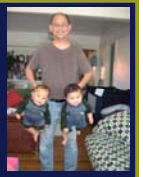
De Novo Genome Sequencing: Genome Annotation as a Tool to Assess Short Read Assemblies

Igor Grigoriev*, Alan Kuo, Edward Kirton, Erika Lindquist, Darren Platt, Paul Richardson
DOE Joint Genome Institute, Walnut Creek, CA, 94598, USA.
*Correspondence: ivgrigoriev@lbl.gov



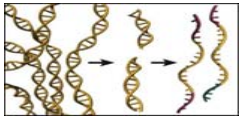
Abstract

New sequencing technologies that produce large numbers of short reads promise to make whole genome sequencing cheaper and faster. Assembly of huge numbers of short reads is computationally challenging, especially for large eukaryotic genomes,. At the same time utility of short reads for assembly of larger eukaryotes still needs to be shown. Quality of these assemblies can be assessed using genome annotation, where genome and EST assemblies are critical inputs. Two applications for 454 pyrosequencing – genome assembly and EST sequencing – are discussed here. We annotated short-read and hybrid assemblies of the genome of *Phytophthora capsici*, a widespread plant pathogen. Short-read ESTs were employed for annotation of a fungus, *Aspergillus niger*.

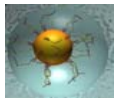
ESTs provide an invaluable resource for gene discovery in the process of genome annotation. Assembled full-length (FL) genes and splice site signals derived from EST-genome alignments are used for training gene predictors. Assembly of a FL gene requires either fewer but longer Sanger reads or significantly higher 454 coverage. At the same time 454 outperforms Sanger in tagging a much broader set of expressed genes with at least one short read to support expression of predicted gene loci.

Several assemblies of *P. capsici* genome that were built from Sanger-only, 454-only, and hybrid data were annotated using a short version of the standard JGI Annotation Pipeline tuned to build homology-based gene models based on the proteins predicted in genomes of two other phytophthoras – *P. sojae* and *P. ramorum* [Tyler et al, Science 2006]. Predicted genes in three *P. capsici* assemblies were analyzed for frameshifts resulting in internal stop codons and for completeness of predicted gene models. While we observed a high rate of frameshifts due to sequencing errors and polymorphisms, Sanger and 454 data complement each other and hybrid assemblies may provide a solution for sequencing smaller eukaryotes.

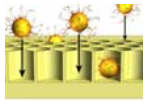
454 sequencing



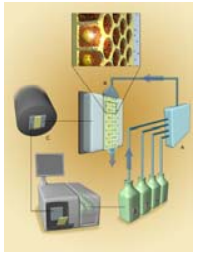
1. prepare adapter-ligated ssDNA library



2. clonally amplify on 28 µm beads



3. Load beads and enzymes in PicoTiterPlate™

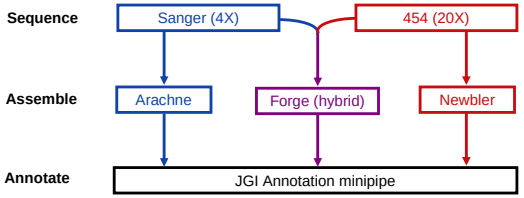


4. Sequence by synthesis on the 454 Instrument

Pros and Cons

	Sanger	454(GS20)	
Cost per bp	$0.01	$0.0002	Cheaper sequencing
Basepairs per run	307 kbp	25 Mbp	Increased depth of coverage
No. reads per run	384	250000	
Read length	800 nt	105 nt	Poor assembly of repetitive regions
Error rate	0.1%	1%	Frameshifts in predicted genes

Genome Assemblies



Gene models were predicted in 3 different assemblies of *Phytophthora capsici* using Genewise [Birney & Durbin, 2000] based on *Phytophthora sojae* and *Phytophthora ramorum* protein seeds.

Assembly Comparison

	Sanger (Arachne)	454 (Newbler)	Sanger+454 (Forge)
Assembly size (Mbp)	86.5	40.8	107.8
Gap size (Mbp)	30.9 (36%)	N/A	44.7 (41%)
Number of scaffolds	6,169	14,346	1,406
N50/L50 (num. / kbp)	83 / 262.0	2634 / 3.9	28 / 438.6
Number of genes	17,253	15,407	16,871
Complete (>90% target*)	13,921 (81%)	10,231(73%)	14,161 (84%)
Genes with internal stops (broken)	4,863 (28%)	9,935 (64%)	5,648 (33%)

*targets are proteins of the other *Phytophthora* sp.

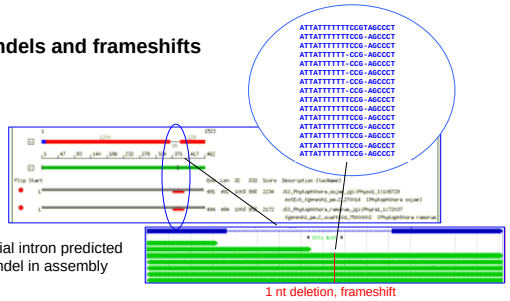
Hybrid assembly shows better statistics and fewer partial models.

Fixed and broken gene models

		Arachne good	Arachne broken	Forge good	Forge broken
Newbler	Good	29%	3%	26%	4%
	Broken	51%	17%	48%	22%
Forge	Good	65%	8%		
	Broken	13%	14%		

Sanger reads allow repairing a large number of frameshifts found in broken models predicted using Newbler assembly.

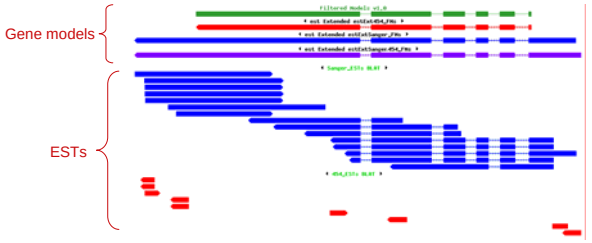
Indels and frameshifts



An artificial intron predicted around indel in assembly

1 nt deletion, frameshift

ESTs



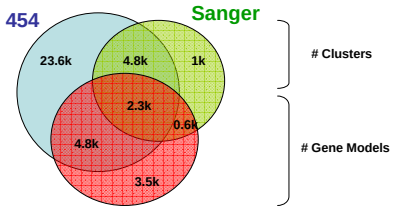
ESTs were aligned to genome assembly using Blat [Kent, 2002], compared to overlapping gene models, which then were extended using available ESTs.

EST Comparison

	Sanger	454	454+Sanger
Number of ESTs (% aligned)	31,578(98.9)	386,512(97.8)	418,090(98.3)
Average EST length	689	104	145
Number of spliced ESTs (%)	18,522(58.7)	104,786(27.1)	123,308(29.5)
Average number of EST HSP	2.2	1.4	1.5
Number of supported genes (%)	2,943(19.4)	7,162(47.1)	7,752(69.1)
Number of extended genes(%/%) **	2,892(98/26)	4,966(69/44)	6,149(79/55)
Genes supported over full length (%/%)**	1,272(43/11)	1,436(20/13)	1,577(20/14)
Average gene length	2,089	1,671	1,836
Average gene length supported by ESTs	89%	62%	62%
Average number of exons	4.1	3.2	3.3

** - fraction of supported/all (11,200) genes

Combined 454 and Sanger ESTs provide support for more gene models and better coverage for each of them.



Gene Models

Acknowledgments

Joann Mudge and **Stephen Kingsmore** at the National Center for Genome Resources
Lei Du at 454 Life Sciences provided the all-454 Newbler assembly
Jeremy Schmutz at JGI provided the all-Sanger Arachne assembly

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.