Exploring Variation Detection Within a Wide Range of Bioenergy-Relevant Species Via Short Read Technology

Wendy Schackwitz, Joel Martin, Sirisha Sunkara, Mary Ann Pedraza, Maria Shin, David Hillman, Anna Lipzen, Crystal Wright, Feng Chen, Len A. Pennacchio
U.S. Department of Energy Joint Genome Institute, Walnut Creek, CA 94598 USA

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

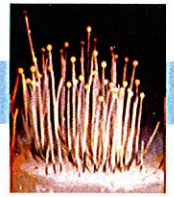
U.S. DEPARTMENT OF ENERGY
Office of Science

Plants

Microbes

+

Fungi

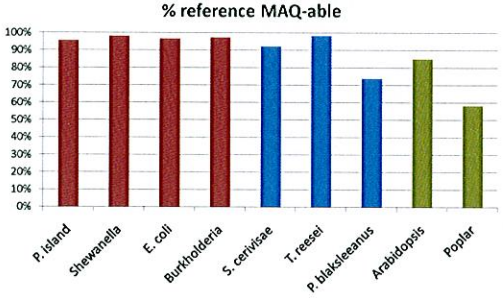


Variation Detection

	Sample	Genome Size (Mb)	Depth	bp covered (Mb)	%ref MAQ-able	% simSNP	# SNP hap/homo	SNP/BP	# SNP hets	% simIn dels	# indels hap/homo	# indels het
Canonical	*Dehalococcoides sp.*	1.3	1566X	1.3	92%	89%	0	0	9	96%	346	348
	Desulfovibrio vulgaris	3.7	112X	3.7	97%	98%	0	0	0	85%	0	0
	Shewanella amazonensis	4.3	176X	4.3	98%	98%	0	0	0	86%	1	2
	Rhodopseudomonas palustris	5.3	113X	5.3	98%	99%	0	0	0	83%	0	0
	Verminephrobacter eiseniae	5.6	52X	5.6	98%	97%	0	0	0	77%	7	7
	Burkholderia cepacia	7.5	62X	7.5	97%	99%	0	0	0	75%	0	0
	Pyrobaculum islandicum	1.8	500X	1.8	96%	98%	0	0	0	NA	NA	NA
	E. coli Rad CB1012	4.6	23X	4.6	97%	96%	71	1/64,788	0	NA	NA	NA
	E. coli Rad CB1013	4.6	47X	4.6	97%	97%	47	1/97,872	2	NA	NA	NA
	E. coli Rad CB1014	4.6	57X	4.6	97%	97%	54	1/85,185	2	NA	NA	NA
	E. coli Rad CB1015	4.6	33X	4.6	97%	97%	57	1/80,701	1	NA	NA	NA
	E. coli Rad CB1024	4.6	57X	4.6	97%	97%	64	1/71,875	1	NA	NA	NA
	E. coli Rad CB1025	4.6	45X	4.6	97%	97%	43	1/106,976	2	NA	NA	NA
	S. cerivisae BY4742	12.2	25X	12.1	92%	92%	361	1/33,795	17	NA	NA	NA
	S. cerivisae EPY330	12.2	20X	12.1	92%	93%	519	1/23,507	40	NA	NA	NA
	Trichoderma reesei	33.4	33X	33.1	98%	97%	410	1/81,463	1210	49%	56	376
	Phycomyces blaeksleeanus	51.0	40X	50.8	72%	72%	73,997	1/687	5745	NA	NA	NA
Canonical	*Arabidopsis thaliana* Col-0	119.5	13X	117.9	85%	73%	829	1/144,150	458	NA	NA	NA
	Arabidopsis thaliana Shadara	119.5	24X	112.9	85%	73%	435,130	1/258	13995	NA	NA	NA
	Arabidopsis thaliana Bay	119.5	20X	113.7	85%	71%	343,295	1/331	9529	NA	NA	NA
	Poplar BESC-79	427.6	33X	411.7	59%	67%	1,098,163	1/375	1539795	43%	184,020	198,744
	Poplar 93-968	427.6	24X	412.6	59%	61%	844,197	1/489	1,089,779	24%	70,805	49,722
	Poplar BESC-52	427.6	19X	409.6	59%	67%	994,608	1/412	1,281,829	40%	152,052	143,099

Better Biofuels!

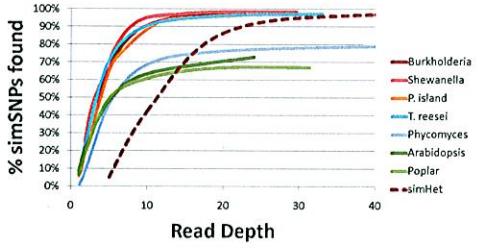
Short Reads Can Capture a Wide Range of Genome Complexity



Conclusion:

- A significant percentage of all genomes are accessible using 35bp reads.

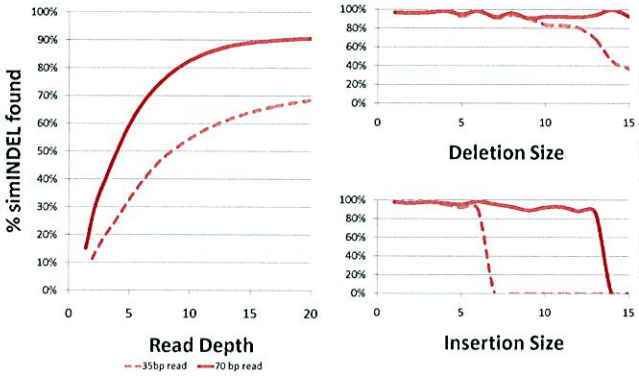
Simulated SNP Analysis Determines Required Depth



Conclusion:

- ~15X depth is sufficient to capture the majority of haploid or homozygous variants.
- ~20x depth is sufficient to capture the majority of heterozygous variants.

Longer Reads Find More Indels



Conclusion:

- 70bp read length give significantly more indel detection.
- Majority of short indels can be found, longer indels can not be found.

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