



One Million Reads of Maize

Jarrold Chapman, Asaf Salamov, Uffe Hellsten,
J. Chris Detter, Tijana Glavina-del Rio, Susan Lucas, Daniel Rokhsar



ABSTRACT

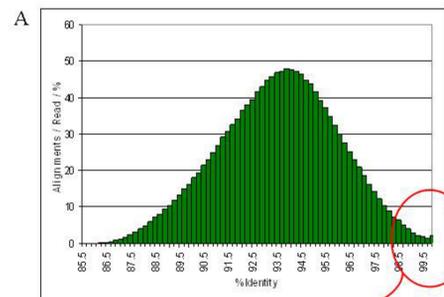
Conventional wisdom holds that the maize genome, with its high density of long (10-13kb) repetitive elements, is not a good candidate for whole genome shotgun sequencing. To explore the structure of the maize genome and assess the feasibility of various sequencing strategies, the DoE Joint Genome Institute sequenced over one million whole genome paired shotgun reads from the maize B73 inbred line from plasmids and fosmids of various insert sizes. Analysis of this million-read sample (~0.37X sequence coverage of the genome) in the context of other public maize and sorghum sequences provides an emerging picture of the repetitive structure and evolutionary history of the *Zea mays* genome. The recent tetraploidization is evident, and the pattern of repetitive activity over the past ~10 million years can be analyzed to show that a whole genome shotgun approach could successfully capture ~90% of the maize genome.

THE DATA SET

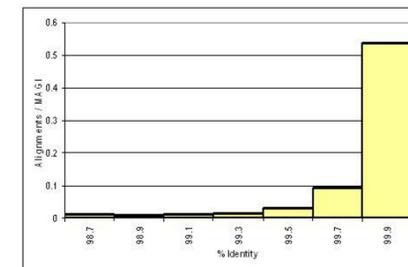
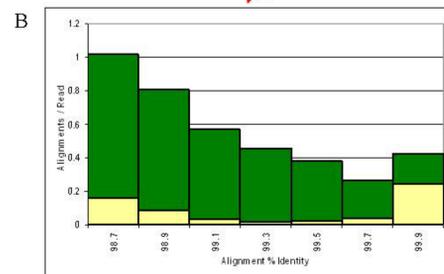
In total, 1.02 million reads from five libraries (four plasmid and one fosmid) were sequenced. By comparing a small fraction of chloroplast contaminant reads to the known organellar sequence, the average shotgun read length at 99.9% accuracy for this dataset is estimated to be 920 bp. This constitutes an ~0.37X coverage of the ~2.4 Gb genome.

Library ID	Insert (kB)	N (kReads)
APSZ	2.8 ± 0.3	454
APTA	6.8 ± 0.9	449
APTB	35 ± 4	16
APZT	4.8 ± 0.5	93
ASBH	4.1 ± 0.4	7
Total		1,019

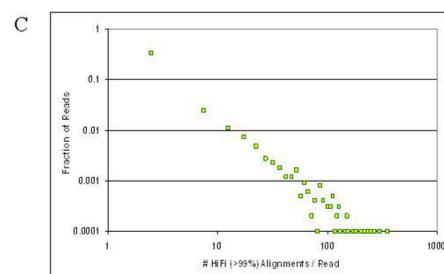
ALIGNMENTS OF SHOTGUN SEQUENCES



A. Histogram of alignment percent-identity for all alignments between a set of 10,000 sample reads and the complete 1.0 million-read set, computed using BLASTN with word size 24, E-value cutoff 10^{-199} , and mismatch penalty -5. These parameters limit alignments to those longer than ~350 bp and above ~85% identity. The main peak centered at 94% identity represents alignments between different repeats. The mean number of alignments per sample maize shotgun read is ~300. The small upturn at 100% includes both “true” overlaps and high-fidelity alignments between recently diverged repeats.

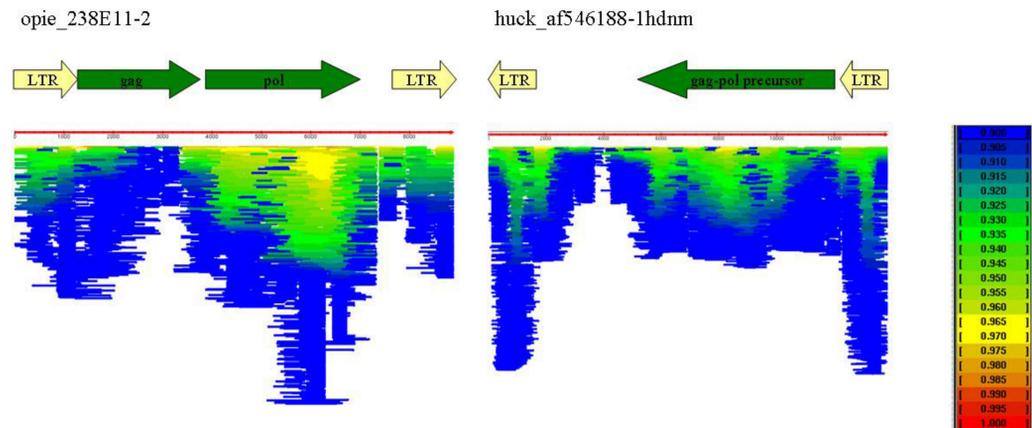


B. Detail showing only high-fidelity alignments. Upper histogram (green) shows all alignments; lower histogram (gold) shows only alignments to reads that have fewer than three high-fidelity alignments. Because at ~0.37X fewer than 0.5% of reads are expected to have three or more true overlaps, this effectively removes repeat-induced alignments. Right panel shows alignment % identity distribution to putatively unique maize assembled gene islands (MAGIs – ISU v. 3.1).



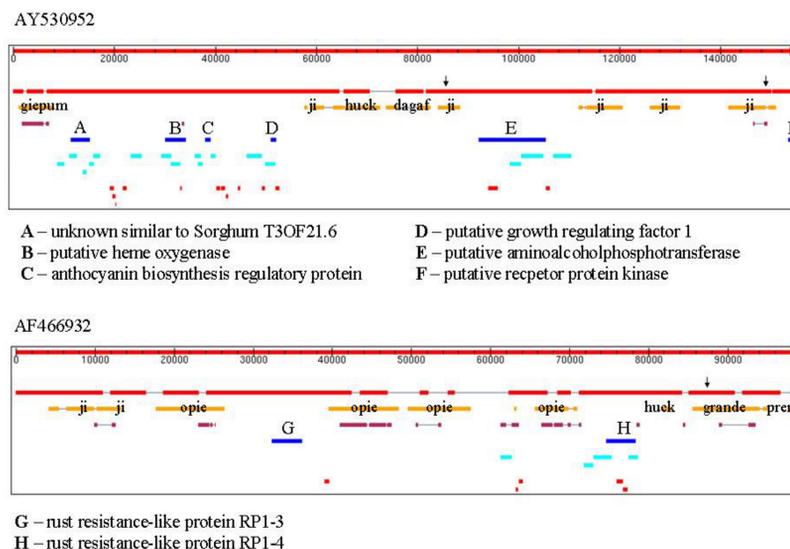
C. Number of reads with given number of high-fidelity (>99%) alignments, on a log-log scale. Analysis shows that at most ~10% of the maize genome has an unexpectedly large number of high-fidelity alignments.

REPEAT DIVERSITY



Alignments of WGS reads to two common maize long terminal repeat (LTR) retrotransposons. Random samples of ~2700 and ~5700 alignments of at least 300 bp to exemplars of the (copia-type) opie and (gypsy-type) huck LTR retrotransposons, respectively, are depicted. Alignments are color-coded by nucleotide %-identity. The vast majority of aligning reads are less than 97% identical to the exemplars with 0/2736 and 1/5701, respectively, more than 99% identical to the sample elements. WGS reads are allowed to align more than once per element. While 50-60% of the maize genome is composed of LTR retrotransposon-related sequences (with huck and opie among the most prevalent), we estimate that recently diverged (>99% identical) copies comprise less than ~7% of the genome.

ASSEMBLY SIMULATION



Simulated WGS assembly compared with two finished maize BACs. Thirteen “finished” maize BACs were concatenated into a ~2 Mb “mini-genome” and sampled *in silico* to simulate 6X sequence coverage in 6-8 kb plasmid clones and 15X clone coverage in 32-40 kb fosmid clones. The virtual reads were combined with the 1.02 million WGS sequences and assembled with JAZZ using stringent settings. The mini-genome was reconstructed into three major scaffolds demonstrating the feasibility of megabase-scale scaffolding in maize via a WGS approach. Results of the reconstruction are depicted for two BACs above. For each BAC, the top line shows assembled contigs (all in one scaffold); second line shows repeats identified by BLAST homology; third line shows high fidelity repetitive (HFR) regions (covered by three or more WGS reads at ≥99% identity); fourth line shows gene spans; bottom two groupings show alignment with maize assembled gene islands (MAGIs) (turquoise) and MAGI singletons (red). Arrows indicate sequence gaps of less than 250 bp that are too small to resolve on this scale.