SCIENTIFIC REPORTS

OPEN

SUBJECT AREAS:
RNA SEQUENCING
GENE EXPRESSION PROFILING
GENOME INFORMATICS
AGRICULTURAL GENETICS

Received
8 April 2013

Accepted
26 February 2014

Published
31 March 2014

Correspondence and requests for materials should be addressed to Z.W. (zhongwang@lbl.gov)

A near complete snapshot of the *Zea mays* seedling transcriptome revealed from ultra-deep sequencing

Jeffrey A. Martin[2], Nicole V. Johnson[2], Stephen M. Gross[2], James Schnable[3], Xiandong Meng[2], Mei Wang[2], Devin Coleman-Derr[2], Erika Lindquist[2], Chia-Lin Wei[2], Shawn Kaeppler[4], Feng Chen[2] & Zhong Wang[1,2]

[1]Genomics Division, Lawrence Berkeley National Laboratory, Berkeley, California, USA, [2]Department of Energy, Joint Genome Institute, Walnut Creek, CA 94598, USA, [3]Department of Plant and Microbial Biology, University of California, Berkeley, CA, 94720, USA, [4]Department of Agronomy and Great Lakes Bioenergy Research Center, University of Wisconsin, 1575 Linden Drive, Madison, WI 53706, USA.

RNA-sequencing (RNA-seq) enables in-depth exploration of transcriptomes, but typical sequencing depth often limits its comprehensiveness. In this study, we generated nearly 3 billion RNA-Seq reads, totaling 341 Gb of sequence, from a *Zea mays* seedling sample. At this depth, a near complete snapshot of the transcriptome was observed consisting of over 90% of the annotated transcripts, including lowly expressed transcription factors. A novel hybrid strategy combining *de novo* and reference-based assemblies yielded a transcriptome consisting of 126,708 transcripts with 88% of expressed known genes assembled to full-length. We improved current annotations by adding 4,842 previously unannotated transcript variants and many new features, including 212 maize transcripts, 201 genes, 10 genes with undocumented potential roles in seedlings as well as maize lineage specific gene fusion events. We demonstrated the power of deep sequencing for large transcriptome studies by generating a high quality transcriptome, which provides a rich resource for the research community.

The recent development of massively-parallel RNA sequencing (RNA-seq) technologies[1] has provided much greater depth than traditional EST sequencing, which has enhanced the discovery capabilities of transcriptomic studies. The increased throughput of RNA-seq has superseded previous sequencing technologies in their ability to quantify gene expression[2], as well as identify gene isoforms and alternative splice variants[3]. RNA-seq has also facilitated the assembly of a transcriptome without a reference genome, or *de novo* assembly, which offers a practical solution for rapid and accurate transcriptome generation for organisms without reference genomes (reviewed in Martin and Wang[4]).

Despite the success of RNA-seq, a typical experiment for a single developmental stage or tissue often has a sequence depth less than 20 million uniquely mapped reads per sample due to the relatively high cost of deep sequencing. At this depth, many lowly expressed transcripts escape detection, which include transcripts derived from transcription factors that are often a primary interest of expression profiling studies[5]. Limited sequence depth can also impede *de novo* assembly where sequencing coverage of at least 30-fold is needed to assemble a full-length transcript from short reads[6]. Consequently, many transcripts with medium expression levels are only partially assembled due to the high unevenness in sequencing coverage resulting in incomplete gene annotation. It is important to address these issues since the success of RNA-seq experiments depends on the quality and comprehensiveness of transcript identification: however, the completeness aspect of transcriptomic studies has not been systematically addressed.

At 2.3 gigabasepairs (Gbp), the *Zea mays* (maize) genome represents a good model for a plant transcriptomic study as there are many challenges associated with gene identification and annotation[7]. The maize genome is highly repetitive with the vast majority of the genome (85%) consisting of transposable and repetitive elements[8], which increases the chance of mis-mapping short reads and complicates *de novo* transcriptome assemblies[9]. Additionally, the presence of two maize subgenomes[10], resulting from a tetraploidization event approximately 12 million years ago[11], has led to an overabundance of paralogous genes, or ohnologs, that are commonly found in plant genomes. In *de novo* assemblies, these highly similar gene pairs are often *de novo* assembled into a single gene further complicating transcriptome analysis.

The goal of this study was to explore the impact of read depth on the comprehensiveness of assembly and annotation of a transcriptome by deeply sequencing a maize seedling mRNA sample. The majority of maize genes (57–66%) are expressed at this developmental stage[12], which makes the maize seedling an ideal candidate for evaluating the power of ultra-deep RNA sequencing for transcriptome analysis. First, we conducted a comparative analysis of a shallow and ultra-deep sequence datasets to systematically evaluate the biological functions that were missed in shallow sequencing. We then assembled a near complete transcriptome via a hybrid approach by combining *de novo* and reference-based assembly strategies. Lastly, we analyzed the assembled transcript isoforms in our final assembly and were able to uncover several unique features of this maize transcriptome.

Results

Increased read depth detects expression from most genes. To better understand the capabilities of ultra-deep RNA sequencing for the annotation of large transcriptomes, we generated an ultra-deep sequence dataset from a maize seedling mRNA sample (Materials and Methods). Briefly, we constructed several short (180–250 bp) and long (500 bp) insert libraries to capture the variable lengths of the transcripts and sequenced them using Illumina sequencing platforms. These libraries together yielded nearly 2.8 billion paired-end (PE) reads and 341 Gb of sequence (Table 1). This depth is equivalent to 148× total genome coverage where only 10% of the maize genome is covered by more than 2 reads, while the rest is largely not transcribed in this sample.

To evaluate the ability of deep sequencing to detect the presence of rare transcripts, we compared our deep dataset to a typical RNA-seq experiment with shallow sequencing datasets. We generated 10 simulated shallow sets by randomly sampling 20 million read pairs from the total set of 2.8 billion reads and used the number of high confidence set of annotated maize transcripts (5b)(http://ftp.maizesequence.org/) detected by each dataset as a comparative measure. After aligning the reads from each set to the 5b transcripts, we detected 52,826, or 83%, 5b maize transcripts with at least two uniquely mapped fragments in the total set while we only detected 45,382 (71%) transcripts, on average, in the shallow sets (Supplementary Table S1). As expected, the median transcript abundance of the deep set (FPKM = 0.66) was significantly lower than the median abundance of the shallow sets (average FPKM = 3.01), which demonstrates the increased capability of a deeply sequenced dataset to detect lowly expressed transcripts.

Despite the additional amount of resources required to generate a 2.8 billion read dataset, the additional information gained by deeper sequencing can be an invaluable asset to any study. Therefore, we assessed the sensitivity to detect expressed genes at different sequencing depths to characterize the levels of transcript abundance captured in each dataset. To mimic datasets of variable depths, we generated 13 datasets by randomly selecting paired-end reads, ranging from 20 to 3 billion reads, from the deep dataset. For each dataset, we identified concordant read pairs that aligned to the 5b maize transcripts and calculated the normalized expression values (FPKM) of each expressed transcript (>two uniquely mapped fragments)(Methods). As shown in Figure 1, we observed a steady increase in detected transcripts as sequence depth increased from 20 million to 3 billion reads (All). The abundant transcripts (>0.8 FPKM) were consistently detected beyond 20 million; more moderately abundant transcripts (0.2–0.8 FPKM) were detected from 20 to 100 million that eventually became saturated; while detected rare transcripts (<0.2 FPKM) kept increasing but slowed past 1 billion reads. At the final depth (3.0 billion), nearly 90% of maize transcripts were detected (Supplementary Table S1). These results suggest that deep sequencing is essential for a comprehensive view of the transcriptome, especially for detecting rare transcripts.

To evaluate the importance of the rare transcripts missed from shallow datasets, we conducted a gene ontology (GO) enrichment analysis using the Biological Network Gene Ontology tool, BINGO[13]. To emulate shallow datasets, we simulated 10, 20-million-read datasets and compared the transcripts detected in each shallow set to the set of transcripts detected in the deep dataset. We used BINGO to characterize the types of rare transcripts that were missed by the shallow sets, but detected in the deep set. We found 20 enriched GO terms consisting of 16 biological process and 4 molecular function terms that were consistently enriched in the deep dataset (in all 10 experiments). All terms were associated with transcription factor activities (except 1 with cell wall metabolism, Table 2). Given that transcription factors are typically low in abundance, these results demonstrate the importance of ultra-deep sequencing for studying these low-level genetic elements and genes important in developmental regulation.

A hybrid transcriptome assembly strategy. We used a hybrid assembly strategy to assemble the maize seedling transcriptome via a step-wise, assembly-then-align approach[4] (Figure 2). In the first step, four paired-end libraries (assembly group 1 in Table 1) were *de novo* assembled using Rnnotator[6] for an initial set of assembled contigs. As part of this process, single 250 b reads from one of the four libraries were constructed by computationally joining two 150b overlapping reads and used for the assembly. In the second step, reads from assembly group 2 (Table 1) were aligned to the contigs from step 1 to extend these contigs while reads that failed to align were independently assembled into contigs. Contigs from the above two steps were then combined to form a comprehensive *de novo* assembly. In the final step, the *de novo* assembled contigs were aligned to the current maize reference genome (B73 maize RefGenv2) and fragmented contigs from the same gene were merged. This merging step reduced the number of contigs from 187,045 to 126,708 while increasing the median contig length from 440 bp to 869 bp. The resulting assembled contigs were considered to be our "final" transcriptome assembly and used for all subsequent analyses.

Table 1 | Summary of RNA-seq datasets used in this study

Platform	Insert size (bp)	Stranded	Read length	Number of reads	Data size (Gb)	Assembly group*
GAII	180	no	76	191,125,326	14.5	1
GAII	250	yes	151	253,935,324	38.3	1
GAII	500	yes	151	166,641,198	25.2	1
HiSeq	250	yes	100	489,875,122	49.0	1
HiSeq	250	yes	100	752,239,100	75.2	2
HiSeq	250	yes	150	927,571,432	139.1	2
Total				2,781,387,502	341.4	
Pac Bio	>1000	no	1,425	144,177	0.3	

*Data from assembly group 1 were used in the initial de novo assembly while assembly group 2 were used for assembly improvement.

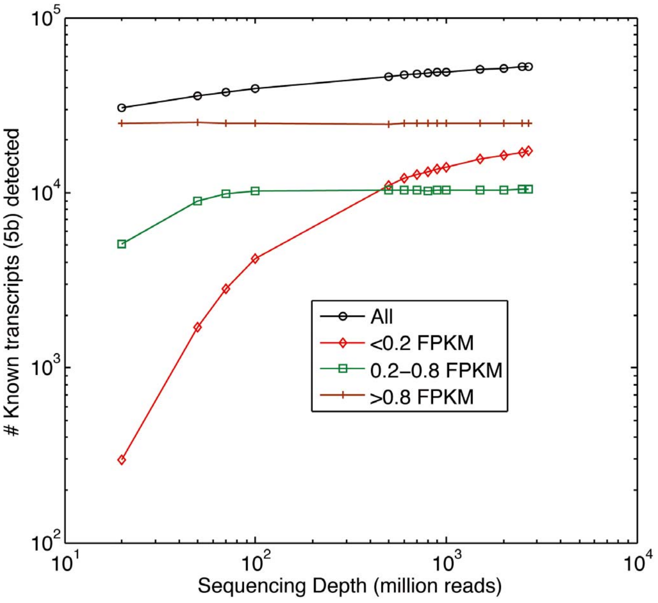
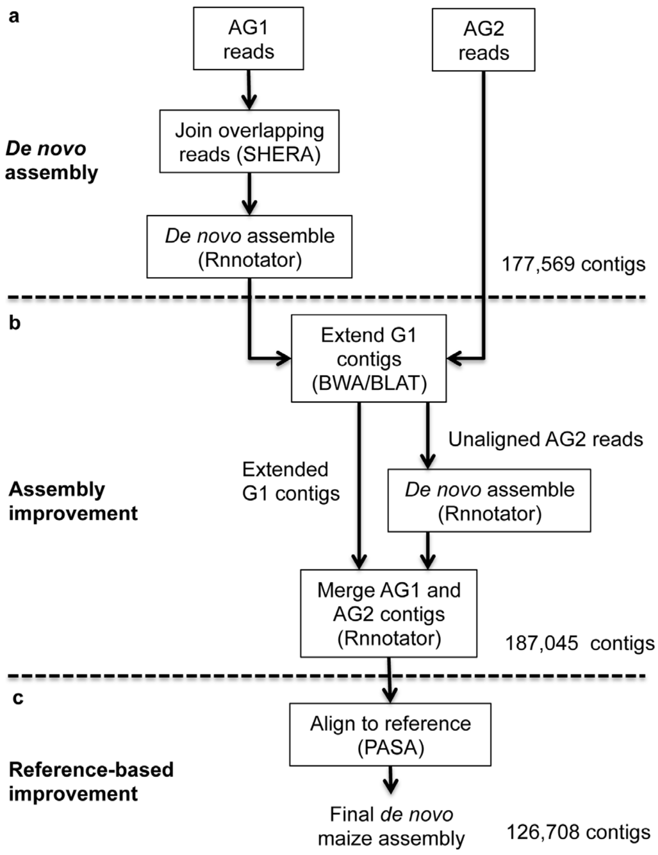


Figure 1 | Rare transcripts are continually detected with ultra-deep sequencing. Saturation curves were generated to illustrate the transcript detection capabilities at various sequencing depths. The sequence depth of 13 datasets is plotted on the x-axis and the number of 5b maize transcripts detected on the y-axis. The normalized expression values (FPKM) of the detected transcripts were categorized by low, moderate, and high expression, which are represented by red, green, and brown lines, respectively, while the black line includes all levels of expression. Most moderately to highly expressed gene transcripts were detected with shallow sequencing, while about half of the lowly expressed transcripts (<0.2 FPKM) were detect at 100 million reads. Rare transcripts were also continually detected after 750 million reads while a miniscule increase in transcripts with moderate to high expression (>0.2 FPKM) was observed.



a
AG1 reads
AG2 reads
De novo assembly
Join overlapping reads (SHERA)
De novo assemble (Rnnotator)
177,569 contigs
b
Extend G1 contigs (BWA/BLAT)
Unaligned AG2 reads
Extended G1 contigs
Assembly improvement
De novo assemble (Rnnotator)
Merge AG1 and AG2 contigs (Rnnotator)
187,045 contigs
c
Align to reference (PASA)
Reference-based improvement
Final *de novo* maize assembly
126,708 contigs

Figure 2 | *De novo* assembly of the maize seedling transcriptome. (a) 3.8 billion Illumina paired end reads were generated using the Genome Analyzer II (GAII) and HiSeq. A subset of reads (Table 1, assembly group 1), denoted as AG1 reads, was used to create an initial *de novo* assembly using Rnnotator resulting in 177,569 assembled contigs. (b) The assembly was improved by using a combination of BLAT and BWA to align a second set of reads (Table 1, assembly group 2), or AG2 reads, to extend the contigs. Unaligned AG2 reads were *de novo* assembled and combined with the improved AG1 contigs for a total of 187,045 contigs. (c) All assembled contigs were aligned to the reference maize genome assembly to further extend truncated or join adjacent transcripts for a final *de novo* assembly consisting of 126,708 contigs.

Table 2 | GO terms enriched in transcripts unique to ultra-deep sequence dataset

		Description
BP	GO:0065007	biological regulation†
	GO:0050789	regulation of biological process†
	GO:0009889	regulation of biosynthetic process†
	GO:0031326	regulation of cellular biosynthetic process
	GO:0031323	regulation of cellular metabolic process†
	GO:0050794	regulation of cellular process†
	GO:0010468	regulation of gene expression†
	GO:0010556	regulation of macromolecule biosynthetic process†
	GO:0060255	regulation of macromolecule metabolic process†
	GO:0019222	regulation of metabolic process†
	GO:0051171	regulation of nitrogen compound metabolic process†
	GO:0019219	regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolic process†
	GO:0080090	regulation of primary metabolic process†
	GO:0051252	regulation of RNA metabolic process†
	GO:0045449	regulation of transcription†
	GO:0045449	regulation of transcription, DNA-dependent†
MP	GO:0030528	transcription regulator activity†
	GO:0003700	transcription factor activity†
	GO:0043565	sequence-specific DNA binding†
	GO:0030599	pectinesterase activity‡

BP = biological process, MP = molecular process.
†GO terms related to transcription factor activities.
‡GO terms related to cell wall metabolism.

A systematic quality evaluation of the assembled maize transcriptome. We evaluated the overall quality of this transcriptome assembly by applying several quality metrics[4] to assess the accuracy, completeness, and contiguity compared to the current maize genome assembly and annotations. Overall, we obtained a highly accurate assembly with 93.4% of the contigs having extremely accurate alignments (>95% identity) to the maize genome (Figure 3a). When compared to the high-confidence set of annotated maize transcripts (5b), 78.2% of the annotated transcripts, or 98.7% of transcripts with detected expression, were fully covered by one or more contigs as indicated by the completeness criterion (Figure 3b). Of these, 81.1% (or 63.4% of the total) were represented by a single contig covering the entire annotated transcript, as indicated by the contiguity metric (Figure 3c).

The paleopolyploid maize genome complicates transcriptome assembly due to the prevalence of duplicated gene pairs, or ohnologs, with high sequence similarities that may result in artificial gene fusions. Therefore, we introduced another assembly quality metric to differentiate and assess the assembly of ohnolog gene pairs. Using a dataset of high-confidence ohnologous gene pairs[10], we selected 141 ohnologs with read coverage >30× over 80% of their length and grouped the pairs by the percent sequence identity (<87.5%, 87.5%,

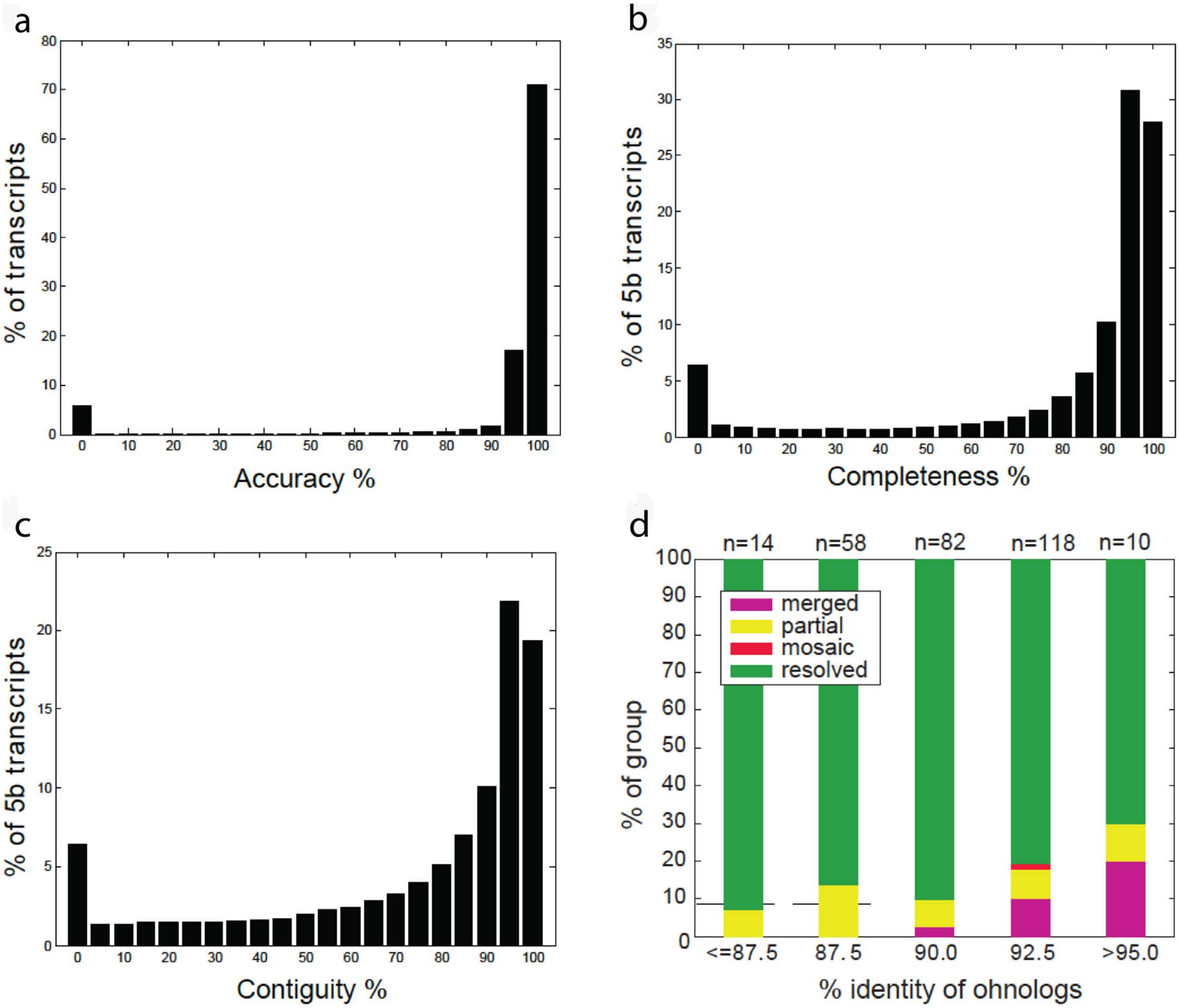


Figure 3 | Quality assessment of the de novo maize assembly. (a) The accuracy of the assembly was determined by the percentage of assembled transcripts (x-axis) that aligned to the maize reference with sequence identity (>95%). (b) When compared to the high quality set of maize annotations (5b), the percent length of the annotations represented by the corresponding assembled contig determined completeness of the assembly while the proportion of annotations covered by a single assembled transcript demonstrated the assembly's contiguity (c). (d) The isoforms of 141 high-confidence ohnologs were investigated and categorized by the differentiation status in the assembly. Ohnolog pairs were grouped by sequence identity (x-axis) and the percentage of isoforms with varying degrees of differentiation were calculated (y-axis) where isoforms were either merged with its paralogous pair (merged), partially assembled (partial), inaccurately assembled (mosaic), or correctly assembled (resolved).

90%, 92.5%, >95%)(Figure 2d) for a confident reference set for evaluation. With the exception of a few merged ohnolog pairs (16 at identity >=90%) and two pairs assembled into a mosaic transcript (2 at 92.5% identity), the majority (123 out of 141, 87.2%) of the pairs were unambiguously resolved.

We also assessed the ability of the high quality assembly to detect biologically relevant genes by investigating the identification of a set of well-studied "classical" maize genes[14]. We were able to assemble approximately 380 (93.6%) of the genes to near completion (Supplementary Table S2) and detect new expression patterns for 24 of the genes. Of these, 10 genes previously had undocumented expression in seedlings (Supplementary Table S3) while the expression of 14 genes was typically restricted to meristems and developing leaf primordia (Supplementary Table S4).

Since the above quality measurements do not evaluate our ability to assemble alternative spliced isoforms, which are a prevalent feature of higher eukaryote transcriptomes, we conducted an additional study using Pacific Biosciences (PacBio) sequencing technologies[15]. We generated 300 Mb of long reads from the same mRNA sample for a set of alternative spliced transcripts (accession: SRA053579) (Table 1). Given that most *de novo* assembly errors typically occur at the exon junctions, we were confident that the quality of the assembly could be effectively inferred based on the PacBio reads. We obtained a high quality set of long reads using several filtering criteria (Methods and Figure 4a). First, we corrected the errors within the PacBio reads using Illumina reads[16] and obtained 50,130 high quality long reads (>500b). Among these, 23,380 reads that spanned at least two exons were further selected as a reference set for the validation of spliced variants. With the small amount of quality PacBio data, we only used these data to validate isoforms contained within these regions.

When aligned to the final transcriptome assembly, 98.8% of these long PacBio reads formed alignments without any gaps, suggesting

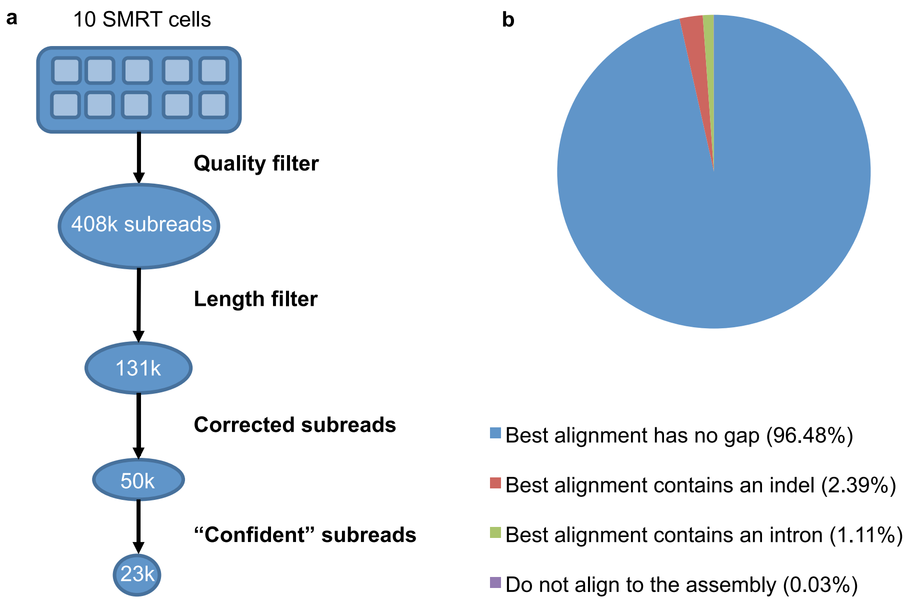


Figure 4 | PacBio sequence data validates the exon structure of *de novo* assembly. (a) Sequence reads were generated using the Pacific BioSciences (PacBio) sequencing platform. The reads were filtered by quality, length, and error corrected for a set of confident subreads that were used to validate alternative splicing in the *de novo* assembly. (b) The Illumina reads were aligned to the PacBio subreads and categorized by reads without a gap, those containing a gap (indel or intron), and those that did not align. The PacBio reads were also aligned to the assembled contigs where subreads were located within, extended, or contained a contig.

that the majority of the assembled transcripts contained the correct splicing information (Figure 4b). Furthermore, almost all of the confident PacBio subreads were aligned to the assembled transcriptome, which is consistent with the accuracy of this assembly.

Based on all of the above quality metrics, we could confidently conclude that we generated a high quality maize seedling transcriptome assembly, and therefore, carried out further analyses to investigate the unique transcriptome features.

Novel transcripts expand the current transcriptome annotations. Compared to the above-assembled transcriptome, we found that many genes within the current maize annotations in the maize version 5a working gene set (5a) were truncated or completely missed (Supplementary Table S5, Supplementary Figure S1). In 299 cases, the open reading frame of a maize gene was significantly extended by our new annotation, as the original ones contained truncated coding sequences (CDS) that were missing start or stop codons. Collectively, these improvements to the maize annotation impacted 1% of the currently annotated maize protein coding genes.

Within our set of assembled transcripts, we found that 16,507 aligned to maize pseudomolecules with high sequence homology (>90%), but did not overlap any genes in the 5a annotations, which is the unfiltered, more comprehensive transcriptome annotation (http://ftp.maizesequence.org/). Among these, 7,249 (or 66%) had significant BLASTX hits against other grass species (Sorghum, rice or Seteria) where 212 contained full-length open reading frames (ORFs). This suggests that they are likely true, previously unannotated maize transcripts and not contamination.

In addition, we discovered 201 transcripts that did not align to the RefGenv2 maize genome assembly or maize annotations (Supplementary Table S6). With further interrogation of the transcript isoforms, we found that most of transcript isoforms (148) were likely to contain protein coding regions, and were not contaminants, because they encode proteins that share sequence homology with closely related plant species (see Methods and Supplementary Table S7 and S8). One of the predicted proteins is derived from a previously undefined locus, Locus10189, which encodes a novel N6-adenosine methyltransferase (MT-A70) protein family. Interestingly, it only shares 19.8% amino acid sequence identity with the only other known MT-A70 protein in maize encoded by locus GRMZM2G116563, which further supports the newly discovered maize proteins by ultra-deep sequencing.

Lastly, we compared the splice junctions of the 5a transcripts to our assembled gene set and detected 23,043 new transcript isoforms where 4,842 contained full-length ORFs that were not repetitive. These newly defined isoforms contributed an additional 8% to the total number of known transcript isoforms in maize that adds yet another set of genetic elements undetected by the current available RefGenv2 maize annotations.

Maize lineage specific transcriptome features compared to other grasses. Using our improved set of maize annotations, we conducted a comparative study to evaluate the syntenous genomic regions of related species, *Oryza sativa* (rice) and *Sorghum bicolor* (sorghum), to identify unique transcriptome features of the maize lineage[17]. We first examined gene fusion events after the divergence of these species and identified 36 pairs of genes annotated as two separate genes in rice and sorghum that were represented as a single gene in our assembly (Supplementary Table S8). Since genes with overlapping transcripts from the same strand could be assembled into single transcripts ("fused assemblies") and the overlapping regions could have much higher sequencing coverage, we excluded possible fused assemblies and identified four high-confident maize specific gene fusion events. Two of the fusion events resulted in fused protein products while the other two did not contain a complete open reading frame. The validity of these gene fusions was supported by the presence of paired-end reads that spanned the gene fusion junctions and by RT-PCR of each of the four fusions (Supplementary Table S9, Supplementary Figure S2).

Figure 5a shows a particularly interesting gene fusion detected in an assembled transcript (Locus2145v1rpkm80.17). It encodes a previously unannotated chromatin remodeling SNF2_N-domain protein, SNF2_N, which spanned the gene fusion junction. Therefore, we suspect that the current annotations in maize, sorghum bicolor, and rice (*Oryza sativa*) are likely inaccurate, as multiple genes were annotated in all three species (Supplementary Figure S3–6).

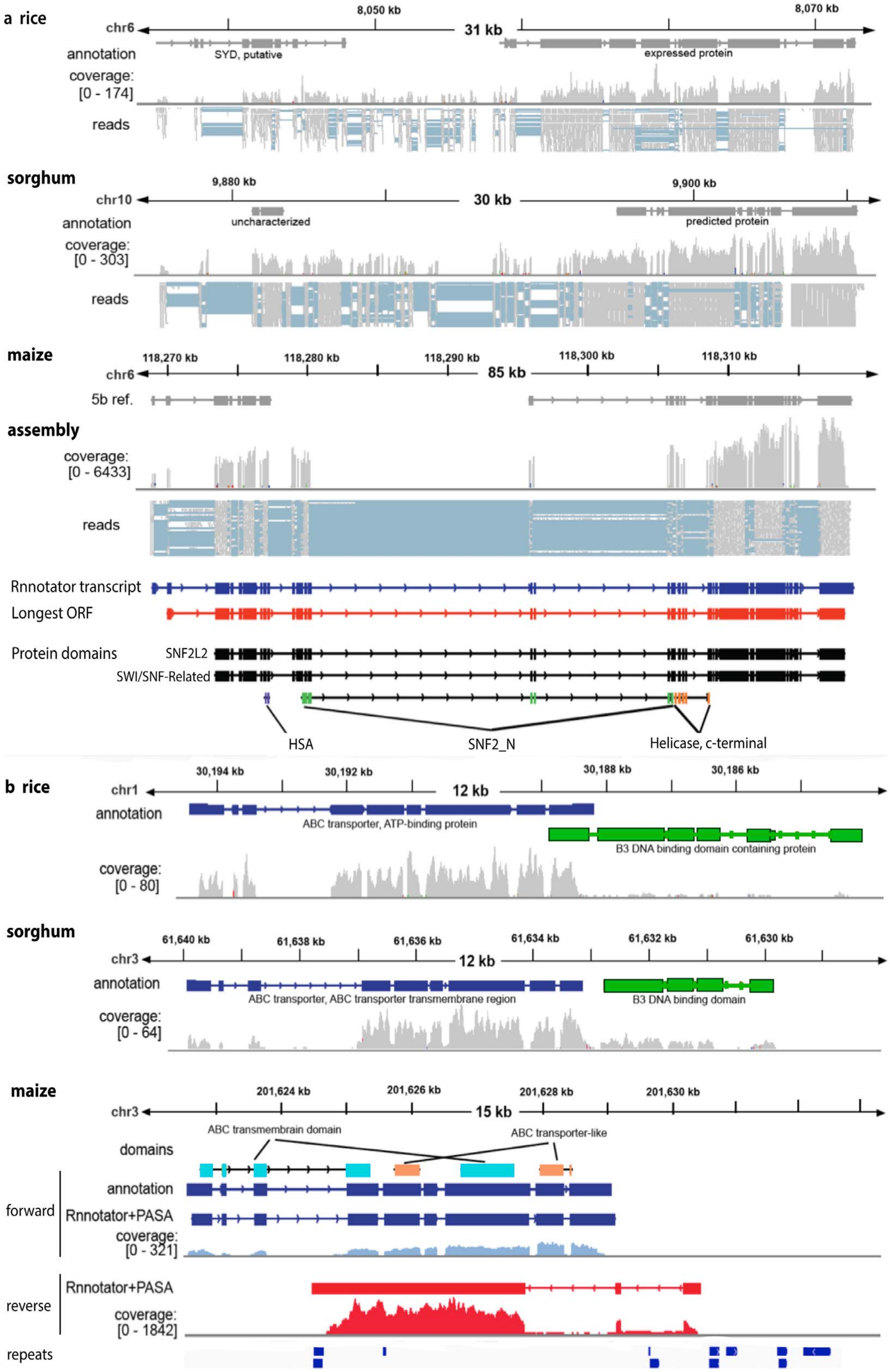


Figure 5 | The *de novo* assembly detects new maize annotations and reveals incomplete annotations in closely related grasses. (a) The current annotations for two proteins, HSA and helicase, is shown for rice and maize. The read coverage in rice shows a clear separation of the two proteins, but in the *de novo* assembly, an assembled transcript spanned the region. The open reading frame (ORF) located in the transcript contained an additional protein, SNF2_N, that was previously unannotated in maize. (b) A syntenic region identified in rice, sorghum and maize that contained significant anti-sense transcription with sense (blue) and anti-sense (red) expression levels shown alongside the strand-specific reads assembled in Rnnotator. The repeats located in the region are displayed in the last panel. A portion of the second protein, B3 DNA binding domain, identified in rice and sorghum is lost in maize. The presence of repeats, which are good transcription facilitators, upstream of the antisense transcript is suggestive of a transcriptionally active region.

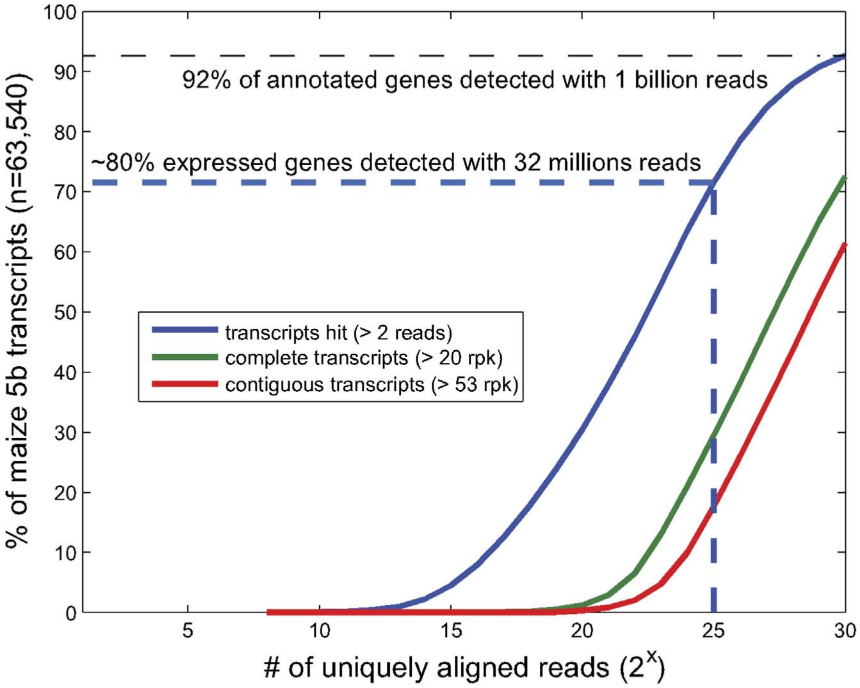


Figure 6 | Proportion of 5b maize annotations assembled at various read depths. Reads that uniquely aligned to maize transcripts were assessed at different read depths. The read depth is shown on the x-axis and the percentage of 5b transcripts hit by the reads on the y-axis. The blue line represents 5b transcripts hit by at least 2 unique reads, the green line are transcripts completely assembled (>80% complete), and the red line are contiguous transcripts, as defined by the completeness and contiguity metrics. The dotted lines represent the point at which the vast majority of genes are expressed.

Given that anti-sense transcription plays an important role in transcriptional regulation across various organisms[18,19], we surveyed the entire maize transcriptome for the presence or absence of significant maize lineage specific, anti-sense transcription sites. By evaluating the orientation of Illumina reads aligned to the 5b gene annotations, we determined that 1,127 of the 39,656 genes (2.8%) in the 5b annotation had significant anti-sense transcription. Another 732 transcripts from the 5b annotation exhibited extreme degrees (>90%, mean = 98%) of anti-sense transcription, suggesting that the strand of these genes are mis-annotated in the 5b annotation or are strong indicators of a regulatory region.

Of the extreme anti-sense transcription sites, we found one syntenic region between maize, sorghum and rice to be particularly interesting. There are two, largely separated transcripts on opposite strands in rice and sorghum, but the transcript encoding a B3 DNA binding domain protein appears to be partially deleted in maize (Figure 5b). This deletion in maize likely led to the addition of a new transcript from the opposite strand that overlaps most of the transcript encoding an ABC transporter. Our assembled transcript, Locus19225v1rpkm4.69, extends the current RefGenv2 annotation, which is denoted as a novel coding region, but there is no known protein domains contained in the region. The presence of upstream repetitive elements, which are known to be drivers of transcription, suggests that this gene could be transcriptionally active, but additional experiments are needed to understand the molecular consequences of this putative lineage-specific gene deletion and addition event in maize.

Discussion

In summary, we generated an ultra-deep sequenced data set from a single mRNA sample and systematically evaluated the impact of sequence depth on expression detection. We found that many transcripts that encode proteins with regulatory roles are not detected in RNA-Seq experiments with typical sequencing depth. By applying a hybrid assembly approach we compiled a high quality, comprehensive annotation of the maize transcriptome and added many novel transcripts and isoforms to the current maize annotations. These findings highlight the utility of ultra-deep RNA-seq for comprehensive transcriptome annotation in complex large plant genomes, especially for those that lack robust reference genomes.

By increasing the read depth of our RNA-seq dataset, we were able to assemble and annotate a more comprehensive maize transcriptome. The insights gained in this study are not limited to plant genomes, but could be generically applied to any large and complex transcriptome. For example, many researchers are unaware of the required sequencing depth for a transcriptome study to achieve their specific scientific goals. In the case of maize seedling transcriptome, we evaluated the relationship between the number of uniquely aligned reads and transcript detection/assembly (Figure 6). At a depth achieved by most RNA-Seq studies with less than 32 million reads, nearly 80% of the genes can be detected for expression. However, if the scientific goal is to *de novo* assemble full-length transcripts using this depth and the reference genome is unknown, or very incomplete, only 20–30% of the expressed transcripts can be assembled. With our ultra-deep sequence dataset, we demonstrated that with 1 billion uniquely aligned reads, it is possible to assemble close to 93% of the transcripts with over 60% of the transcripts being completely assembled and highly expressed. From this rarefaction curve, we can also predict that additional sequencing (e.g., adding another billion of uniquely aligned reads) is less likely to be very useful for detection of gene expression, but is still useful for *de novo* assembly of full length transcripts. Therefore, an optimal sequencing depth for a large transcriptome has to be determined by balancing the scientific need and cost of sequencing.

Despite the improvements made to numerous maize annotations, the scope of genetic elements that we could study was limited to polyadenylated elements due to the experimental methods used, namely the RNA selection and purification protocols. For an all-inclusive transcriptome analysis, total RNA should be used to obtain an in-depth look at other genetic features, such as regulatory mechanisms involved in transcription and translation processes, which were not captured in this study. Deep sequencing is a technique that is promising for such studies: however, the experimental protocols for the proper RNA processing protocols have yet to be firmly established. Once these procedures have been solidified, these, in combination with ultra-deep sequencing, would provide a powerful tool for the in-depth exploration of any transcriptome.

An additional consideration for a future analysis would be to conduct RNA-seq on additional tissue types and/or developmental stages for improved transcript detection. Since we only sequenced a single maize developmental stage, we were restricted to looking at general gene expression patterns by comparing the deeply sequenced and simulated shallow datasets. Generating multiple types of datasets could, potentially, also lessen the need to generate such an extremely deep dataset, but this consideration is dependent on the genetic elements of interest and their predicted expression levels.

Another advantage of sequencing multiple tissue types and/or developmental stages is to help address the question researchers often have about the utility of increasing read depth when the library complexity tends to increase concurrently. By using Rnnotator we addressed this issue during the pre-processing stage of the *de novo* assembly pipeline where duplicated PCR fragments were consolidated by generating a single consensus sequence based on the duplicated PE reads. For most libraries, we found that the fraction of PCR duplications were very small (~1.7%), so this was not an overwhelming issue for this study. In lieu of using Rnnotator, the sequencing of several different tissues or developmental stages would not only help to alleviate this issue, but also facilitate the generation of more comprehensive annotations of a transcriptome. In the case of the maize transcriptome, it would be possible to discover new transcripts not found in this study.

Before the cost of sequencing is dramatically reduced even further, alternative RNA-seq technologies may be considered to maximize the value of sequencing. For example, targeted RNA sequencing is an efficient way to reduce the overall sequencing costs to interrogate a small set of transcripts of interest[20]. With this technology, it is possible to increase the read coverage of the low abundance transcripts so they can be assembled, thereby enabling the study of their genomic and transcriptional features, such as RNA-editing, splicing, and potential gene fusions[21,22]. If the goal is to study genes with low expression levels at the genome scale, normalization during RNA-seq library construction may be an effective way to reduce the representation of highly expressed transcripts while enriching for the low abundant transcripts[23]. While promising, both of these techniques may introduce biases as more experimental steps are introduced. Nevertheless, the benefits of using ultra-deep sequencing for transcriptome analysis alone are plentiful and we predict that this type of sequencing will grow in usefulness as the cost of sequencing decreases.

Methods

Plant materials and RNA extraction. Four *Zea mays* (maize) B73 seedlings were grown as described in Hansey et al[12] and harvested at the V3 stage (3rd juvenile leaf has emerged). Total RNA from above-ground tissues of the seedling was purified using AmpPure SPRI beads (Beckman Coulter). Poly-adenylated RNA was isolated from purified using the Absolutely mRNA Purification Kit (Stratagene) until less than 5% rRNA remained.

Illumina and PacBio sequencing. Experimental methods used for library preparation and sequencing are described in the Supplementary Methods online.

RT-PCR validation. 1 ug of total RNA was DNAse treated and converted to cDNA using Invitrogen's Superscript III kit with Oligo dT primers. Maize gDNA and cDNA were PCR amplified using primers using the annealing temperatures listed in Supplementary Table S10. The following PCR conditions were used: 94°C for 3 minutes, followed by 30 cycles of 94°C for 1 min, 50–62°C (depending on the primer pair) for 1 min, 72°C for 3 minutes, with a final extension of 72°C for 10 minutes.

Comparative analysis of an ultra-deep and shallow sequence datasets. We evaluated the ability of ultra-deep sequencing to detect rare transcripts by comparing this set (Table 1) with 10 shallow sequence datasets generated by randomly sampling 20 million read pairs from the deep dataset. PASA[24] (Program to Assemble Spliced Alignments) was used to align the reads from each dataset to the high confidence set of annotated maize transcripts (5b) from the current maize assembly (http://maizesequence.org, RefGenv2) and transcripts with quality alignments were selected. For PASA, we used a maximum intron length of 50 kb and default parameters otherwise.

Using SAMtools and custom perl scripts, we identified expressed transcripts (>2 uniquely mapped fragments) detected in the deep and shallow datasets, calculated their FPKM[2], and compared the deep transcripts with those in the shallow sets. Using transcripts only detected in the deep dataset, we evaluated the information gained by deep sequencing, or lost with shallow sequencing, by conducting a Gene Ontology (GO) analysis using BiNGO[13] (version 2.44). We retrieved GO and maize transcript annotations from the Phytozome database (http://phytozome.net/) and identified significantly enriched biological process and molecular function GO terms using the Hypergeometric test (FDR[25] corrected p-value $< .05$).

Evaluating the detection of expressed transcripts relative to sequence depth. We generated 13 sequence datasets by sampling 20 m, 50 m, 70 m, 100 m, 500 m, 700 m, 800 m, 900 m, 1b, 2b, and 3b (m = million, b = billion) paired-end reads from the 2.8 billion read dataset that were then mapped to the 5b maize transcripts. The transcript abundance in each dataset, determined by FPKM, was used to evaluate the detection capabilities of the datasets.

Assembly of the maize seedling genome. We generated a *de novo* assembly using Illumina sequence data (Table 1) that was divided into two datasets: assembly group 1 (AG1) and assembly group 2 (AG2), where AG1 was used for the initial assembly and AG2 for improving the AG1 assembly. Overlapping pairs of 150 bp paired-end reads in the 250 bp insert libraries were computationally joined using the Short Read Reducing Aligner (SHERA)[26] resulting in 250 bp single-end reads. Only joined reads with confidence values > 0.5 and at least a 10 bp overlap were included in the assembly. Rnnotator was used to generate a *de novo* assembly of the joined and AG1 sequence data using default parameters. Using BWA and BLAT[27], we aligned AG2 reads to the assembled contigs and extended contigs with overlapping reads. AG2 reads with poor alignments were *de novo* assembled for a second set of contigs using Rnnotator. These contigs were combined with the improved AG1 contigs and used as input into Rnnotator at the merging step to further improve the *de novo* assembly. Using PASA, we aligned the final assembled contigs to the RefGenv2 maize assembly to extend overlapping and adjacent contigs and remove redundancy for a final set of uniquely mapping contigs.

Quality and accuracy assessment of the *de novo* assembly. With the RefGenv2 maize assembly as a reference, we evaluated the quality of the *de novo* assembly for accuracy, completeness, and contiguity using the criteria defined in Martin and Wang[9]. Briefly, the accuracy measures the proportion of correctly assembled bases among expressed reference transcripts while completeness and contiguity measure the portion of a reference transcript covered by an assembled transcript or by a single, longest-assembled transcript.

To evaluate the accuracy of assembling gene pairs with high sequence identities, we used a set of high confidence ohnologs[10], consisting of 1,750 paralogous ohnolog gene pairs, that was filtered for pairs with <30× read coverage over 80% of the length of the gene. We aligned the ohnolog pairs to the assembled contigs using BLAT and evaluated the accuracy and contiguity of the corresponding contig regions using the previously described metrics.

Validation of alternative splice sites using PacBio data. Given the accuracy of PacBio sequence reads to capture the exon structure of genes, we generated a high quality set of PacBio subreads (Figure 4a) to validate alternative splice sites in the *de novo* assembly. We used BLAT to align the PacBio and Illumina datasets to the RefGenv2 maize assembly and retained reads with >75% similarity. The PacBio subreads were error corrected using pacBioToCA[16] with the following settings: length = 500, partitions = 100, -t 8, pacbio.spec params: ovlHashBlockLength = 20,000,000, and ovlRefBlockSize = 5,000,000. Confident corrected subreads that aligned to the maize genome with high sequence identity (>95%), did not contain indels, and spanned at least two exons were used for analysis.

Improving the current maize annotations using the *de novo* assembly. We assessed the ability of our ultra-deep sequence dataset to improve gene annotations by comparing our assembled transcripts, with the longest open reading frames (ORFs), to the 5b maize annotations. We identified annotations extended by our contigs by aligning each annotation's assembled transcript to its corresponding coding sequence (CDS) via BLAT. Alternative splicing was confirmed by comparing the splice junctions of the current maize version 5a working gene set annotations (http://ftp.maizesequence.org/current/working-set/) to our improved gene annotations. Lastly, we used Interproscan[28] to scan isoform transcripts for protein domains and RepeatMasker (http://www.repeatmasker.org) to check these regions for repetitive sequence.

Identification of novel transcriptome features. To identify novel transcripts, we investigated assembled transcripts that did not align to the reference maize assembly. For transcripts with less than 10% sequence identity to the RefGenv2 maize assembly, we used BLASTX[29] to determine protein sequence homology against other species. Transcripts with significant (e-value $< 1e^{10}$) BLASTX hits against the NCBI non-redundant protein database (nr) were filtered and considered novel based on the following criterion: 1) transcript length $>$ 300 bp, 2) not labeled as pre-mRNA, 3) contained a full-length ORF), and 4) top BLASTX hit was against a grass species (*Zea mays*, *Sorghum bicolor*, *Oryza sativa*, *Setaria italica*). Transcripts containing full-length ORFs, denoted by the presence of a start and stop codon, were considered to contain potential coding regions and those with protein domains were considered novel transcript isoforms. We identified enriched protein annotations by normalizing groups of genes with similar protein identifiers over the set of transcripts with protein domains, calculating the actual and expected counts for each group as compared to the entire set, and using the ratio of the two values. Transcripts with the largest ratios were considered to be novel transcripts predicted to encode proteins.

We evaluated the correctness of annotations for other closely related grass species by comparing the assembled maize transcripts to the sorghum genome using LASTZ[30]. Locally duplicated genes in sorghum were condensed using the local duplicate finder within SynMap[31]. An initial set of candidate fusion transcripts were identified where portions of the same maize transcript matched against two or more sorghum genes, which were either immediately adjacent on the sorghum genome or separated by less than three genes. We manually screened these candidate genes in GEvo[32] and compared them to the syntenic orthologous regions in *Setaria italica* and *Oryza sativa* to filter out false positives caused by incorrect annotations within the sorghum genome (typically due to extremely large introns in sorghum resulting in a single gene from other grass species being split into two gene models in sorghum). Since genes with overlapping untranslated regions (UTRs) can also assemble into single transcripts, we identified genes with even read coverage across the putative fusion boundary and considered fusion transcripts with overlapping 500b inserts to be high confidence gene fusions, which were validated by RT-PCR.

Lastly, we evaluated the ability of our assembly to detect anti-sense transcription sites by aligning 388 million strand-specific Illumina reads to the 5b maize annotation transcripts. Transcripts with significant anti-sense transcription had $>$10% but $<$90% of the reads aligned to the anti-sense strand while transcripts with $>$90% of the reads aligned in the anti-sense orientation were considered to be extreme cases.

Data and software availability. Sequence data are available in the Sequence Read Archive (SRA053579). The final *de novo* assembly and isoform transcripts can be downloaded from the JGI ftp site: (ftp://ftp.jgi-psf.org/pub/JGI_data/genome_analysis/Maize_denovo_assem/). The Rnnotator software package is available under a BSD license from the JGI software page (http://www.jgi.doe.gov/software/).

1. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* **10**, 57–63 (2009).
2. Trapnell, C. *et al.* Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc* **7**, 562–78 (2012).
3. Trapnell, C. *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* **28**, 511–5 (2010).
4. Martin, J. A. & Wang, Z. Next-generation transcriptome assembly. *Nat Rev Genet* **12**, 671–82 (2011).
5. Caldana, C., Scheible, W. R., Mueller-Roeber, B. & Ruzicic, S. A quantitative RT-PCR platform for high-throughput expression profiling of 2500 rice transcription factors. *Plant Methods* **3**, 7 (2007).
6. Martin, J. *et al.* Rnnotator: an automated de novo transcriptome assembly pipeline from stranded RNA-Seq reads. *BMC Genomics* **11**, 663 (2010).
7. Rabinowicz, P. D. & Bennetzen, J. L. The maize genome as a model for efficient sequence analysis of large plant genomes. *Current opinion in plant biology* **9**, 149–56 (2006).
8. Schnable, P. S. *et al.* The B73 maize genome: complexity, diversity, and dynamics. *Science* **326**, 1112–5 (2009).
9. Treangen, T. J. & Salzberg, S. L. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nat Rev Genet* **13**, 36–46 (2012).
10. Schnable, J. C., Springer, N. M. & Freeling, M. Differentiation of the maize subgenomes by genome dominance and both ancient and ongoing gene loss. *Proc Natl Acad Sci U S A* **108**, 4069–74 (2011).
11. Swigonova, Z. *et al.* On the tetraploid origin of the maize genome. *Comp Funct Genomics* **5**, 281–4 (2004).
12. Hansey, C. N. *et al.* Maize (Zea mays L.) genome diversity as revealed by RNA-sequencing. *PLoS One* **7**, e33071 (2012).
13. Maere, S., Heymans, K. & Kuiper, M. BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics* **21**, 3448–9 (2005).
14. Schnable, J. C. & Freeling, M. Genes identified by visible mutant phenotypes show increased bias toward one of two subgenomes of maize. *PLoS One* **6**, e17855 (2011).
15. Eid, J. *et al.* Real-Time DNA Sequencing from Single Polymerase Molecules. *Science* **323**, 133–138 (2009).
16. Koren, S. *et al.* Hybrid error correction and de novo assembly of single-molecule sequencing reads. *Nature Biotechnology* **30**, 692–+ (2012).
17. Bolot, S. *et al.* The 'inner circle' of the cereal genomes. *Current Opinion in Plant Biology* **12**, 119–125 (2009).
18. Coram, T. E., Settles, M. L. & Chen, X. Large-scale analysis of antisense transcription in wheat using the Affymetrix GeneChip Wheat Genome Array. *BMC Genomics* **10**, 253 (2009).
19. Faghihi, M. A. & Wahlestedt, C. Regulatory roles of natural antisense transcripts. *Nat Rev Mol Cell Biol* **10**, 637–43 (2009).
20. Ozsolak, F. & Milos, P. M. RNA sequencing: advances, challenges and opportunities. *Nature Reviews Genetics* **12**, 87–98 (2011).
21. Levin, J. Z. *et al.* Targeted next-generation sequencing of a cancer transcriptome enhances detection of sequence variants and novel fusion transcripts. *Genome Biology* **10**, R115 (2009).
22. Mercer, T. R. *et al.* Targeted RNA sequencing reveals the deep complexity of the human transcriptome. *Nat Biotechnol* **30**, 99–104 (2012).
23. Christodoulou, D. C., Gorham, J. M., Herman, D. S. & Seidman, J. G. Construction of normalized RNA-seq libraries for next-generation sequencing using the crab duplex-specific nuclease. *Curr Protoc Mol Biol* **Chapter 4**, Unit4 12 (2011).
24. Haas, B. J. *et al.* Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Research* **31**, 5654–5666 (2003).
25. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society Series B-Methodological* **57**, 289–300 (1995).
26. Rodrigue, S. *et al.* Unlocking short read sequencing for metagenomics. *PLoS One* **5**, e11840 (2010).
27. Kent, W. J. BLAT--the BLAST-like alignment tool. *Genome research* **12**, 656–64 (2002).
28. Quevillon, E. *et al.* InterProScan: protein domains identifier. *Nucleic Acids Res* **33**, W116–20 (2005).
29. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic Local Alignment Search Tool. *Journal of Molecular Biology* **215**, 403–410 (1990).
30. Harris, R. S. Improved pairwise alignment of genomic DNA. *PhD thesis, Penn State Univ.* (2007).
31. Lyons, E., Pedersen, B., Kane, J. & Freeling, M. The value of nonmodel genomes and an example using SynMap within CoGe to dissect the hexaploidy that predates the rosids. *Tropical Plant Biology* **1**, 181–190 (2008).
32. Lyons, E. *et al.* Finding and comparing syntenic regions among Arabidopsis and the outgroups papaya, poplar, and grape: CoGe with rosids. *Plant Physiol* **148**, 1772–81 (2008).

Acknowledgments

The work was conducted by the U.S. Department of Energy Joint Genome Institute and supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231 and the U. S. Department of Energy Great Lakes Bioenergy Research Center (Department of Energy Office of Science BER DE-FC02-07ER64494).

Author contributions

Z.W. conceived the study. M.W., E.L., C.W. and F.C. carried out the RNA-seq experiments. J.M. and X.M. developed computational tools for the maize transcriptome assembly and validation study. J.M. annotated the transcriptome. N.J., Z.W. and S.G. wrote the manuscript and N.J., S.G., J.S. and S.K. contributed to biological interpretation of the results. D.C. performed the RT-PCR experiments. N.J. and Z.W. revised the manuscript and N.J. performed additional analyses in response to reviewer comments. All authors read and approved the manuscript for publication.

Additional information

Supplementary information accompanies this paper at http://www.nature.com/scientificreports

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Martin, J.A. *et al.* A near complete snapshot of the *Zea mays* seedling transcriptome revealed from ultra-deep sequencing. *Sci. Rep.* **4**, 4519; DOI:10.1038/srep04519 (2014).