Extended methods for the annotation of *Triticum aestivum* CS42.

Luca Venturini*[1], Gemy Kaithakottil†[1], and David Swarbreck‡[1]

[1]Earlham Institute, Norwich, UK

October 13, 2016

*luca.venturini@earlham.ac.uk
†gemy.kaithakottil@earlham.ac.uk
‡david.swarbreck@earlham.ac.uk

1 Construction of the wheat gene set

The wheat gene set for wheat was generated using a custom pipeline integrating wheat-specific transcriptomic resources, including PacBio transcriptomic data, similarity to proteins of related species, and evidence-guided ab initio predictions generated with AUGUSTUS (Stanke et al., 2006).

The pipeline was divided in five different phases. In the first phase, RNA-Seq models were generated with 4 different assembly methods utilising data from multiple tissues and conditions, and integrated together with PacBio transcripts into a coherent and non-redundant set of models using Mikado (Venturini et al., 2016). In the second phase, PacBio reads were classified based on protein similarity and a subset of high quality (e.g. full length, canonical splicing, non-redundant) transcripts employed to train an AUGUSTUS wheat-specific gene prediction model. In the third phase, AUGUSTUS was used to generate a first draft of the genome annotation, using as input Mikado-filtered transcript models, reliable junctions identified with Portcullis (Mapleson et al., 2016), and peptide alignments of proteins from five different species closely related to wheat (*Brachypodium distachyon* 314 v. 3.1, *Zea mais* 284 v. 6a, *Oryza sativa* 204 v. 7.0, *Sorghum bicolor* 313 v. 3.1, and *Setaria italica* 312 v. 2.2, all downloaded from Phytozome (Goodstein et al., 2012)). In the fourth stage, this draft annotation was refined and polished by identifying and correcting probable gene fusions, missing loci and alternative splice variants. Finally, the polished annotation was functionally annotated and all loci were assigned a confidence rank based on their similarity to known proteins and their agreement with wheat transcriptomic data.

1.1 Reference guided transcriptome reconstruction

1.1.1 Alignment of Illumina RNA-seq data

Data preparation RNA-Seq data from three different datasets was utilised for the annotation: ERP004714 (used for the annotation provided in IWGSC (2014)), ERP004505 (used for the grain-development analyses in Pfeifer et al. (2014)) and an internally generated dataset of 250bp paired-end strand-specific reads from six different tissues (PRJEB15048; Table 1). In total, the three datasets comprised over 3.2 billion paired-end reads. For each dataset, read samples were collapsed by tissue and filtered using trim-galore v. 0.3.7 (BabrahamLab, 2014), with the command line options:
`-q 20 --phred33 --stringency 5 --fastqc --length 60`
Due to concerns of high concentration of ribosomal RNA in the internally produced samples, reads from that dataset were further filtered using SortMeRNA v. 2.0 (Kopylova et al., 2012), with the command line options:
`--num_alignments 1 --fastx --paired_in`
and using RFam (5S and 5.8S) and Silva (Archea 16S-23S, Bacteria 16S-23S, Eukariota 18S-28S) as databases.

Alignment with STAR Filtered reads were aligned to the wheat genome using a forked version of STAR-2.5.0-alpha (Dobin et al. (2013), commit `f82c5a0028`; see (`https://github.com/alexdobin/STAR/issues/85`)). The genome was indexed using the option
`--genomeChrBinNbits 14`
in accordance with STAR documentation, and the process had to be performed on a UV supercomputer due to the memory requirements (∼2TB of RAM). Reads were aligned with stringent parameter in a two pass approach to ensure alignment accuracy, a first pass using the custom command-line options
`--outFilterMismatchNmax 3 --alignEndsType EndToEnd`
`--alignIntronMin 20 --alignIntronMax 200000`
`--outSJfilterIntronMaxVsReadN 10000 10000 10000`
to increase the accuracy of the alignments and
`--outSAMattributes NH HI NM MD AS XS`
to ensure the compatibility of the output with downstream tools such as Cufflinks (Trapnell et al., 2010). All 1,519,861reliable junctions detected by STAR in at least one sample during this first pass were collapsed, and given as input for a second round of alignments, with the same command line parameters but also providing the merged junction file with the options:
`--limitSjdbInsertNsj 2000000 --sjdbOverhang 250`
Finally, the alignments from all samples were filtered with portcullis v. 0.10.1 (Mapleson et al., 2016) to exclude spliced reads with non-canonical junctions that were on manual review identified as predominantly due to misalignment.

Alignment with TopHat2 As the original IWGSC annotation had been created using the aligner TopHat2 (Kim et al., 2013), we also aligned reads from the ERP004714 dataset using this program. To retrieve splicing junctions related to the original annotation, IWGSC models were aligned against our reference using GMAP v. 2015-09-29 (Wu and Watanabe, 2005), with the command line options:
`--min-identity=0.99 --min-trimmed-coverage=0.90 -n 1`
and subsequently collapsed and filtered for models only with canonical junctions using gffread from Cufflinks v. 2.2.2beta (Trapnell et al., 2012; Roberts et al., 2011a,b). 281,562 unique splicing junctions from the aligned models were retrieved with a custom Python3 script from the surviving 85,242 models and provided to TopHat v.2.1.0 (patched to use Bowtie2.2.5 (Langmead and Salzberg, 2012) long indices; the patch was subsequently integrated into the later TopHat v.2.1.1). Reads from ERP004714 were then aligned in single pass using the CLI options
`-a 13 -i 20 -I 400000 -g 20 --no-discordant -N 1 --read-edit-dist 1 --read-realign-edit-dist 1 --read-gap-length 1 --library-type fr-unstranded`
and additionally providing the junction file from above.

Table 1: Sequencing reads used in this study. ERP004714: Grain, Leaf, Root, Spike and Stem, ERP004505: 10DPA, AL_20DPA, AL.SE_30DPA, REF_20DPA, SE_20DPA, SE_30DPA and TC_20DPA, PRJEB15048: seedling, root, leaf, stem, spike and seed.

	ERP004714	ERP004505	PRJEB15048
Number of samples	5	7	6
Number of reads	1,536,051,415	873,709,556	824,241,135
Number of filtered reads	1,412,029,174	873,550,049	731,931,657
Average no. filtered reads per sample	282,405,834.8	124,792,864.1	121,988,609.5
Aligned reads (STAR)	1,203,100,456	744,087,908	488,750,691
Aligned reads (STAR second pass)	1,267,816,403	759,278,032	579,642,183
Aligned reads (TopHat2)	1,299,830,440	NA	NA

Table 2: Number of PacBio reads, per sample and size-fraction.

Stage	Size Fraction	Leaf	Root	Seed	Seedling	Spike	Stem	Total
Reads of insert	0.7 - 2 kbps	345,566	482,417	410,969	227253	353,196	210,462	2,029,863
	2-3 kbps	267,379	410,186	364,988	330,525	375,062	376,717	2,124,857
	3-5 kbps	367,571	356,396	301,030	110,628	311,537	370,739	1,817,901
	Total	980,516	1,248,999	1,076,987	668,406	1,039,795	957,918	5,972,621
IsoSeq + Quiver	0.7 - 2 kbps	69,817	116,164	86,031	77,211	98,848	79,909	527,980
	2-3 kbps	55,789	125,622	77,619	97,894	90,340	104,293	551,557
	3-5 kbps	73,513	73,351	56,315	34,818	88,516	103,272	429,785
	Total	199,119	315,137	219,965	209,923	277,704	287,474	1,509,322
Aligned		187,583	297,970	205,990	197,535	259,329	265,816	1,414,223
% aligned		94.21%	94.55%	93.65%	94.10%	93.38%	92.47%	93.70%

1.1.2 Alignment of PacBio RNA-seq data

Data preparation PacBio sequencing data from six tissues was analysed initially using the SMRTanalysis package (v2.3.0.140936), stopping at the quiver step. The "CircularConsensus" step of the ConsensusTools utility was called with the command-line options
`--minFullPasses 0 --minPredictedAccuracy 75`
while during the classification step the option
`--min_seq_len 300`
was invoked. The pipeline provided a total of over 1.5 million PacBio transcriptomic reads for downstream analyses (Table 2).

Read alignment PacBio reads were aligned using the gmapl utility from GMAP v. 2015-11-20 (Wu and Watanabe, 2005), with the command line options
`-f 2 --no-chimera -n 1 --min-trimmed-coverage=0.90 --min-identity=0.95 --split-output`
We further discarded alignments deemed to be translocations by GMAP (those reported in the .transloc file).

1.1.3 Transcript assembly

The illumina RNA-Seq alignments (18 from STAR and 5 from TopHat2) were assembled by tissue/condition using three different tools: CLASS v. 2.12 (Song et al., 2016), Cufflinks v. 2.2.2 beta (commit 753c109e31; Trapnell et al. (2010); Roberts et al. (2011a,b)) and StringTie v.1 .10 (Pertea et al., 2015). CLASS was called using the option `-F 0.05`; Cufflinks was invoked asking to limit the intron size to 200,000 and using both the fragment-bias correction and the multi-read rescue method:
`-I [200000] -b -u`
Samples from the internal dataset were assembled using also the option:

Table 3: Illumina and PacBio transcript assembly statistics. For each tool, assembled transcripts have been clustered into loci using `cuffcompare` (v.2.2.1, command line options "`-C -G`"; Trapnell et al. (2010))

Method	Loci	Transcripts	Average number of exons	Average cDNA size	Number of monoexonic transcripts
CLASS	181,259	3,188,679	5.48	1,304.55	326,210
Cufflinks	270,456	3,281,661	4.37	1,595.44	1,078,721
StringTie	285,728	3,826,431	4.47	1,554.83	1,117,717
Trinity	244,384	646,244	2.96	1,301.02	333,428
PacBio (4 samples)	81,752	1,020,650	6.80	2,109.06	131,357
PacBio (all 6 samples)	88,609	1,330,372	6.79	2,100.97	173,661

Table 4: Mikado transcript assembly statistics.

	Genes	Transcripts	Average number of exons	Average cDNA size	Number of monoexonic transcripts
Mikado (4 PacBio)	81,848	120,886	6.36	2,098.83	18,554
Mikado (6 PacBio)	83,144	128,030	6.29	2,182.37	19,175
Mikado (Illumina and PacBio)	273,243	373,861	4.07	1,377.70	93,564

`--library-type fr-firststrand`
StringTie was invoked asking for assemblies longer than 200bp ("`-m 200`"). In addition the alignments of reads from the internal dataset (6 tissues) were merged using the MergeSamFiles utility from picard (Wysokar et al., 2016). The merged BAM file was used as input for Trinity v.2.1.1 (Haas et al., 2013) in genome-guided mode, using the command line options:
`--SS_lib_type RF --genome_guided_max_intron 200000`
The assembled transcripts were then aligned against the genome using gmap from GMAP v. 2015-11-20 (Wu and Watanabe, 2005), using the command line options:
`-f 2 --min-trimmed-coverage=0.80 --min-identity=0.90`
Uniquely and multiply mapping transcripts were further filtered using a custom python3 script to retain only those alignments in which the assembled transcript mapped against the same region from which its original read cluster originated from. The number and features of transcripts detected by each method is reported in Table 3.

We used Mikado (Venturini et al., 2016) to integrate the ∼11 million Illumina assemblies generated by multiple assembly tools (CLASS, Cufflinks, StringTie, Trinity) and ∼1.4 million aligned PacBio reads. Mikado leverages transcript assemblies generated by multiple methods to improve transcript reconstruction. Loci are first defined across all input assemblies with each assembled transcript scored based on metrics relating to ORF and cDNA size, relative position of the ORF within the transcript, UTR length and presence of multiple ORFs. The best scoring transcript assembly is then returned along with additional transcripts (splice variants) compatible with the representative transcript.

We generated three Mikado selected transcript sets for use in gene predictor training or annotation (Table 4):

1. Alignments from 4 PacBio samples (Root, Seedling, Spike, Stem) were analysed with Mikado 0.11.0, without BLAST data and disabling the "chimera_split" algorithm. The transcript set was used in gene predictor training.

2. Mikado (v. 0.19.2) run on the full set of 6 PacBio samples, with BLAST data, and enabling the chimera_split option in "PERMISSIVE" mode.

3. The 70 RNA-Seq assemblies (23 alignments * 3 assemblers + Trinity) and PacBio alignments (Root, Seedling, Spike, Stem) were analysed using Mikado v. 0.18.0 with the "chimera_split" option set to PERMISSIVE.

For Mikado runs incorporating BLAST data transcripts passing the "prepare" step were blasted against filtered and masked proteins of *B. distachyon*, *O. sativa*, *S. bicolor*, *S. italica* and *Z. mays* using BLAST+ v. 2.2.30 and limiting each result to the best 15 matches.

1.2 Gene predictor training

The primary PacBio alignments from 4 samples (Root, Seedling, Spike, Stem) analysed with Mikado 0.11.0 were filtered for full-length complete and coding transcripts using Full-lengtherNEXT (v0.0.8; Fernandez and Guerrero (2012)) with open reading frames (ORFs) predicted using TransDecoder v2.0.1 (Grabherr et al., 2011). A reliable set of transcripts were selected for training AUGUSTUS having single full length ORF, with 5' and 3' UTR present, consistent Full-lengtherNEXT and TransDecoder CDS coordinates, a minimum CDS to transcript ratio of 50% and a single transcript per gene. We excluded genes with a genomic overlap within 1000bp of a second gene and gene models that are homologous to each other with a coverage and identify of 80%. The filtered PacBio set contained 9952 transcripts selected for training AUGUSTUS. The trained AUGUSTUS model resulted in 0.941 sn, 0.844 sp nucleotide level, 0.798 sn, 0.756 sp exon level and 0.455 sn, 0.367 sp at the gene level.

1.3 Gene prediction using evidence guided AUGUSTUS

Protein coding genes were predicted using AUGUSTUS (Stanke et al., 2006) by means of a Generalized Hidden Markov Model (GHMM) that takes both intrinsic and extrinsic information into account.

1.3.1 Generation of external hints for gene prediction

Junctions RNA-Seq junctions (defining introns) were derived from RNA-Seq alignments (From TGAC: Leaf, Stem, Spike, Seed, Seedling and Root samples; From accession ERP004505: 10DPA, AL_20DPA, AL.SE_30DPA, REF_20DPA, SE_20DPA, SE_30DPA and TC_20DPA samples; From accession ERP004714: Grain, Leaf, Root, Spike and Stem samples), using portcullis v.0.12.0 (Mapleson et al., 2016) and the default set of filtering parameters. Junctions that pass and fail the portcullis filter were classified as Gold and Silver respectively.

Table 5: Description of reference protein datasets used with AUGUSTUS (Stanke et al., 2006). Proteins were filtered at 50% identity and 80% coverage and junctions checked against the Illumina junctions as an additional filtering criterion. Any intron over 50kb resulted in the protein alignment being removed.

	B. distachyon	***O. sativa***	***S. bicolor***	***S. italica***	***Z. mays***
Total Proteins	52,972	49,061	47,205	43,001	88,760
Proteins Aligned	30,354	23,929	23,231	23,107	38.653
Proteins Aligned (%)	57.30%	48.77%	49.21%	53.74%	43.55%
Protein Alignments	105,190	89,739	83,561	86,381	142,217

Proteins Protein sequences from 5 species (*Brachypodium distachyon*, *Oryza sativa*, *Sorghum bicolor*, *Setaria italica* and *Zea mays*) were soft masked for low complexity (segmasker from NCBI BLAST+ 2.3.0) and aligned to the soft masked genome (using PGSB repeats) with exonerate v2.2.0 (Slater and Birney, 2005) with parameters:

```
--model protein2genome --softmaskquery yes --softmasktarget yes --bestn 10 --minintron 20
```

To identify a high confidence set of alignments, exonerate results were filtered at 50% identity and 80% coverage. Furthermore, alignments whose introns were either longer than 50kbps or that were not present in the set of Illumina RNA-Seq junctions were removed from further analysis (see Table 5).

PacBio transcript classification To generate high confidence evidence hints for gene prediction, Mikado filtered PacBio transcripts (Root, Seedling, Spike, Stem) were classified into the following three categories:

Gold : PacBio reads having a full length hit (complete/putative complete) with Full-LengtherNEXT and having a maximum of 2 complete 5'UTR exons and 1 complete 3'UTR exon;

Silver : Remaining models meeting the maximum 5'UTR and 3'UTR restrictions with an additional constraint of having at least 900bp CDS length;

Bronze : any remaining Mikado PacBio transcripts were assigned to the bronze category.

In addition, polished (Quiver high and low quality filtered) PacBio reads were filtered for splice sites that are concordant with Illumina RNA-Seq alignments and were used along with other evidences for the gene prediction.

Classification of Mikado transcripts The Mikado models (combining Illumina and PacBio assemblies) were classified into the following three categories:

Gold : Mikado transcripts having a full length hit (complete/putative complete) with Full-LengtherNEXT and having having a maximum of 2 complete 5'UTR exons and 1 complete 3'UTR exon;

Silver : Remaining models meeting UTR restrictions with an additional constraint of having at least 300bp CDS length;

Bronze : Any remaining Mikado transcripts were assigned to bronze category if they had a maximum intron length of 50kbp.

RNA-seq coverage hints Individual RNA-Seq bam files from STAR were merged together and reads were extracted from merged bam using picardtools (SamToFastq.jar v1.84; Wysokar et al. (2016)). The extracted PE reads were then normalised using a Trinity utility (v2.0.2; Grabherr et al. (2011)):

```
insilico_read_normalization.pl --max\_cov 50 --pairs_together --KMER_SIZE 25
```

and were used to create the normalised bam with picardtools (FilterSamReads.jar v1.84; Wysokar et al. (2016)). The wig file was generated using RSeQC v2.3.7 (bam2wig.py; Wang et al. (2012)) and then converted to a hints file using a utility provided with AUGUSTUS (v2.7; (Stanke et al., 2006)):

```
wig2hints.pl --width=10 --margin=10 --minthresh=2 --minscore=4 --prune=0.1 --radius=4.5
```

1.3.2 Gene prediction

AUGUSTUS (v2.7) was used to predict gene models for the Wheat CS42 TGACv1 genome assembly by utilising the evidence hints generated from five sets of cross species protein alignments, PacBio models, Mikado PacBio models, PacBio plus Illumina Mikado models and RNA-Seq junctions (defining introns). Interspersed repeats were provided as "nonexonpart" hints and RNA-Seq read density was provided as "exonpart" hints. We assigned higher bonus scores and priority based on evidence type and classification (Gold, Silver, Bronze) to reflect the reliability of different evidence sets. Statistics of the generated models are presented in Table 6).

Table 6: AUGUSTUS gene prediction statistics.

Gene Count	224,994
Total transcripts	224,994
Transcripts per gene	1
Transcript mean size (incl. intron) (bp)	3547.89
Transcript mean size cDNA (bp)	1447.66
Transcript median size cDNA (bp)	1239
Min cDNA	8
Max cDNA	15,613
Total exons	833,929
Exons per transcript	3.71
Exon mean size (bp)	390.58
Total exons (distinct)	827,714
Exon mean size (distinct) (bp)	392.09
CDS mean size (bp)	302.18
CDS mean size (distinct) (bp)	302.22
Transcript mean size CDS (bp)	959.71
Transcript median size CDS (bp)	747
Min CDS	3
Max CDS	14,259
5UTR mean size (bp)	154.03
5UTR mean size (distinct) (bp)	153.96
3UTR mean size (bp)	249.69
3UTR mean size (distinct) (bp)	249.73

1.4 Gene model refinement

The primary gene models generated by AUGUSTUS were corrected to remove long terminal introns spanning over 10kbp, identified from manual review as likely artefacts. To identify incorrectly split genes, AUGUSTUS gene models were compared against the high quality Mikado PacBio Gold and Silver set of gene models to identify cases where more than one AUGUSTUS model was contained within a PacBio model with at least 80% nucleotide precision (specificity), in which case we retained only the AUGUSTUS gene model with the highest nucleotide F1.

To add reliable alternative splice variants we ran PASA (Haas, 2003) with a filtered set of transcripts, removing from Mikado transcripts and PacBio reads those which had introns greater than 10kb, and retaining PacBio splice junctions that were consistent with RNA-Seq Illumina alignments. Transcripts were integrated into the annotation via a PASA utility:

```
validate_alignments_in_db.dB --MIN_INTRON_LENGTH=20 --MAX_INTRON_LENGTH=50000
 --MIN_PERCENT_ALIGNED=70 --MIN_AVG_PER_ID=95 --NUM_BP_PERFECT_SPLICE_BOUNDARY=3
```

A second round of updates to the annotation was generated with PASA assemblies constructed from only PacBio reads. To identify and correct gene annotation artefacts, any incorrectly fused PASA models were replaced with a PacBio Gold gene model when the latter was found to overlap with a nucleotide recall of at least 30%. PASA transcripts associated with the incorrectly fused PASA gene but not found to overlap with the PacBio Gold gene model were clustered into new loci and retained. Transcript models with cDNAs shorter than 300bp were removed from further analysis.

1.5 Assignment of gene biotypes and confidence classification

Gene models were classified as coding, non-coding and repeat associated and assigned as high or low confidence based on support from cross species protein similarity and wheat transcripts.

We decided to assign a confidence ranking to each transcript, in three levels:

Protein ranking : this rank is based on similarity - or lack thereof - of the transcript against publicly available protein datasets. The rankings go from 1 (best) to 5 (worst).

Transcript ranking: this rank is based on support for the model - or lack thereof - from our multiple sources of transcriptomic evidence. The rankings go from 1 (best) to 5 (worst).

Confidence: we assigned a general binary confidence tag ("High" vs "Low") for each transcript. To qualify to be considered a high-confidence *coding* transcript, a model has to fall in one of the following categories:

- Protein ranking P1 and transcript ranking T4 or better
- Protein ranking P2 and transcript ranking T4 or better
- Protein ranking P3 and transcript ranking T1

1.5.1 Cross species protein similarity ranking

Each gene model was assigned a protein rank (P1–P5) reflecting the level of coverage of the best identified homolog in a plant protein database. Protein ranks were assigned as:

Protein Rank 1 (P1) : proteins identified as full length in Full-LengtherNEXT with the UniProt database or at least 80% coverage in a supplementary BLAST database consisting of *A.thaliana*, *B. distachyon*, *O. Sativa*, *S. bicolor*, *S. italica* and *Z. mays* proteins

Protein Rank 2 (P2) : proteins with at least 60–80% coverage in the supplementary BLAST database;

Protein Rank 3 (P3) : proteins with at least 30–50% coverage in the supplementary BLAST database;

Protein Rank 4 (P4) : proteins with a low coverage hit (between 0–30%) in the supplementary BLAST database;

Protein Rank 5 (P5) : proteins with no hit in the supplementary BLAST database.

1.5.2 Wheat transcript support ranking

A transcript rank (T1–T5) was assigned based on the extent of support for the predicted gene model from either wheat PacBio reads or assembled wheat RNA-Seq data (all 10,943,015 transcripts assembled from all four transcript assembly methods).

We calculated a variant of annotation edit distance (*AED*) and used this to determine a transcript level ranking. First we define accuracy *AC* as:

$$AC = (SN + SP)/2$$

where *SN* is sensitivity and *SP* specificity, and then derived the AED:

$$AED = 1 - AC.$$

Rather than taking the union of all transcript evidence, we calculate *AED* at base, exon and splice junction level against all individual wheat transcripts used in our gene build (Illumina assemblies, cDNAs and PacBio reads), we then take the mean of base, exon and junction *AED* based on the transcript that best supported the gene model. *AED* statistics were calculated using the compare utility from Mikado (Venturini et al., 2016).

Transcript ranking was assigned based on:

Transcript Rank 1 (T1) : Full length support from cDNA or Pacbio read;

Transcript Rank 2 (T2) : full length support from Illumina assemblies;

Transcript Rank 3 (T3) : Best average *AED* less than 0.5;

Transcript Rank 4 (T4) : Best average *AED* between 0.5 and 1;

Transcript Rank 4 (T5) : No transcriptomic support (best average $AED = 1$).

1.5.3 Assignment of a locus biotype

Following the assignment of protein and transcript rankings, we assigned a locus biotype to each gene.

Repeat associated biotypes Genes were classified as repeat associated if all their transcripts aligned with at least 20% similarity and 30% coverage to the TransposonPSI library (v08222010; Haas (2010)) and had at least 40% coverage by PGSB interspersed repeats. In addition, genes with transcripts that had at least 20% similarity and 50% coverage to the TransposonPSI library or had at least 60% coverage by the PGSB interspersed repeats were also classified as repeat associated. In order to reduce the number of false positive calls, the combined set of putative repetitive transcripts identified above were further checked using a BLAST dataset (comprising protein sequences from *A. thaliana* TAIR10.31, *B. distachyon* v3.1, *H. vulgare* v1.31, *O. sativa* v7.0, *S. bicolor* v3.1, *S. italica* v2.2 and *Z. mays* v6a, all from Phytozome) filtered specifically for repeats, by excluding any sequence corresponding to one of the following parameters:

- Protein with a match for "retrotransposon", "transposon" or both in their description
- At least 30% similarity and 60% coverage to a hit in TransposonPSI

Any assignment of repeat-associated status was judged a false positive call if the protein had a hit with at least 30% coverage against the filtered protein dataset above.

Non-coding RNAs Genes where all the transcript had a protein rank of P4 or P5 were checked to verify whether they could constitute putative non-coding RNAs. Transcript sequences were analysed with CPC v. 0.9.2 (Kong et al., 2007) in conjunction with Uniref90 from Uniprot (retrieved on 11th March 2016). Transcripts were called as putative non-coding RNAs if they met the following conditions:

- PR4 and CPC score lower or equal than -1
- PR5 and CPC score lower than 0

Table 7: Rankings and confidence of coding transcripts.

Protein Rank	Transcript Rank	Confidence	Transcript Count
P1	T1	High	66404
P1	T2	High	43423
P1	T3	High	20937
P1	T4	High	10013
P1	T5	Low	21469
P2	T1	High	3461
P2	T2	High	3545
P2	T3	High	3392
P2	T4	High	2084
P2	T5	Low	6213
P3	T1	High	1813
P3	T2	Low	4521
P3	T3	Low	3995
P3	T4	Low	3406
P3	T5	Low	12210
P4	T1	Low	781
P4	T2	Low	3116
P4	T3	Low	2846
P4	T4	Low	2494
P4	T5	Low	7484
P5	T1	Low	2079
P5	T2	Low	4638
P5	T3	Low	3944
P5	T4	Low	2915
P5	T5	Low	12364

Protein-coding genes Genes not assigned as non-coding were classified as protein coding; all the transcripts associated with them were assigned the same biotype.

1.5.4 Removal of spurious genes

After assigning a biotype to each gene, we performed a final polish of the annotation by marking for removal loci where all the transcripts met the following criteria:

- Putative non-coding transcripts lacking transcript support (TR5)
- Putative coding transcript lacking transcript and protein similarity support (TR5,PR5)
- Protein coding transcripts harbouring an in-frame stop-codon

Before discarding these transcripts, we performed an expression estimation against all of our samples using Kallisto v 0.42.5 (Bray et al., 2016); in parallel, we aligned all high-confidence protein coding transcripts from the previous annotation (IWGSC, 2014) using GMAPL v. 2015-11-20 (Wu and Watanabe, 2005) and asking for the best match with coverage over 90% and identity over 95% (excluding chimeric alignments). Genes were retained if one of their transcripts met at least one of the following conditions:

- Expression level over 0.5 TPM in at least one of our samples, as measured by `Kallisto`
- BLAST hit from the Full-LengtherNEXT analysis with the UniProt database.
- Match against the IWGSC set, with *AED* lower than 1, as measured by Mikado compare

Any gene whose transcripts were all marked for removal, even after these last checks, was excluded from the final annotation. Table 7 reports the final number of coding transcripts per each rank.

1.5.5 Assignment of high and low confidence tags

Based on the above ranking, gene models were classified as high and low confidence as follows:

A **High confidence (biotype Protein_coding)** - any protein coding gene where any of its associated gene models meet the following criteria:

Table 8: TGACv1 annotation biotype and gene confidence assignment.

Confidence Level	Biotype	Gene Count
High	protein_coding	104091
High	ncRNA	10156
Low	Protein_coding_repeat associated	8556
Low	protein_coding	83217
Low	ncRNA_repeat_associated	1954
Low	ncRNA	9933

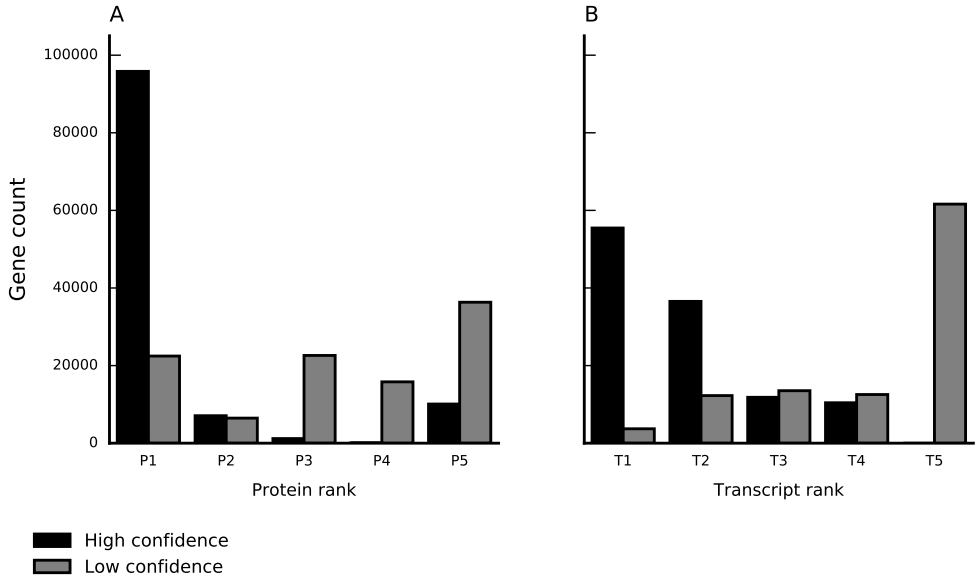

Figure 1: Assessment of confidence rankings for the protein coding portion of the wheat gene set. Protein **(A)** and transcript **(B)** classification for high and low confidence genes (gene level) based on classification of the representative gene model.

- PR1 and TR1 to TR4
- PR2 and TR1 to TR4
- PR3 and TR1

B **Low confidence (biotype Protein_coding)**: any protein coding gene where all of its associated transcript models do not meet the criteria to be considered as high confidence protein coding transcripts.

C **High confidence (biotype ncRNA)**: any ncRNA gene where any of its associated gene models meet the following criteria:

- TR1
- TR2

D **Low confidence (biotype ncRNA)**: any ncRNA gene where all of its associated transcript models do not meet the criteria to be considered as high confidence non-coding transcripts.

E **Low confidence (biotype Protein_coding_Repeat_associated, ncRNA_Repeat_associated)** all repeat associated genes are classed as low confidence.

This classification defines four locus biotypes (protein_coding, ncRNA, protein_coding_repeat_associated and ncRNA_repeat associated) and two locus level confidence classifications: “high” or “low”. Transcript classifications were harmonised within each gene so that each of them only harbours transcripts of one classification, following the order of rankings in the list above.

The number of genes within each category can be found in Table 8, and a graphical summary of the genes associated with each protein and transcript ranking can be found in Figure 1.

1.5.6 Assignment of a representative gene model

We assigned a representative model for a gene by selecting a model with the highest confidence ranking (as described in Table 7, where a rank 1 is greater than a rank 5 model, i.e., PR1 is better than PR5, TR1 is better than TR5) and lowest *AED* by keeping the order:

Table 9: Characteristics of predicted high (HC) and low (LC) confidence wheat genes including coding (mRNA) and long non-coding (ncRNA) RNA.

	All TGAC Models	mRNA HC	mRNA LC	ncRNA HC	ncRNA LC	Repeat-associated
Genes	217,907	104,091	83,217	10,156	9,933	10,510
Transcripts	273,739	154,798	85,778	11,591	10,438	11,134
Transcripts per gene	1.26	1.49	1.03	1.14	1.05	1.06
Transcript mean cDNA size (bp)	1,766.12	2,119.52	1,304.53	1,368.24	1,083.98	1,462.71
Exons per transcript	4.48	5.83	2.8	2.58	2.76	2.27
Exon mean size (bp)	394.15	363.73	465.27	530.25	392.24	644.09
Transcript mean CDS size (bp)	1,165.52	1,361.82	839.97	-	-	891.05
Mono-exonic transcripts	60,322 22.04%	19,034 12.30%	30,479 35.53%	3,061 26.41%	3,044 29.16%	4,704 42.25%
Genes with alternative splicing	32,616 14.97%	28,608 27.48%	2,033 2.44%	1,037 10.21%	460 4.63%	478 4.55%

1. highest protein rank
2. highest transcript rank
3. lowest *AED*.

For ncRNA genes, we assigned the representative model by considering the order:

1. highest transcript rank
2. lowest *AED*.

We compiled a summary of the annotation statistics in Table 9.

1.5.7 Assessment of the TGACv1 annotation

Comparison with *B. dystachion* models. We assessed the coherence in gene length between a selected set of TGACv1 Triticum aestivum and Brachypodium distachyon genes. We have downloaded 2707 Brachypodium distachyon proteins identified as single copy in *Brachypodium distachyon*, *Oryza sativa*, *Sorghum bicolor*, *Setaria italica* and *Zea mays* from Phytozome 11 (BioMart URL link: urlhttps://goo.gl/5Ujnkj). The B.distachyon proteins were blasted (ncbi-blast-2.3.0+, maximum evalue 1×10^{-5}) against TGACv1 *T. aestivum* proteins and the reciprocal best hit was selected using a custom perl script. A high coherence in gene length was found between *B. distachyon* proteins and TGACv1 *T. aestivum* proteins (Figure 2).

Comparison with IWGSC gene models We compared the previous annotation with ours (IWGSC, 2014; Choulet et al., 2014) by aligning the gene models onto our assembly with GMAPL (version `2015-11-20`; Wu and Watanabe (2005)) with the following command line options:

```
gmapl --no-chimeras -n 1 -f 2 --min-trimmed-coverage=0.90 --min-identity=0.95
```

The alignment has been effectuated separately for the high confidence genes and the low confidence set. The alignments were compared against our annotation with Mikado compare (v. 0.22.0; Venturini et al. (2016)), and binned into four different classes:

1. TGAC model missed (class code in the refmap file: NA, X, x, P, p, i, I, ri, rI, u).
2. Structural difference between the TGAC model and the IWGSC model (class codes in the refmap file: f, j, J, n, h, O, C, mo, m, o, e).
3. IWGSC contained within the TGAC model (class codes in the refmap file: c).
4. Concordance between the two annotations (class codes in the refmap file: =, _)

Results are reported in Figure 3 of the manuscript.

1.5.8 Evaluation of non-coding RNAs

Comparison with coding models in *T. aestivum* We extracted the GFF3 of the 10,156 high-confidence ncRNA genes of the TGACv1 annotation using the `grep` utility from Mikado v0.24.0; only representative transcripts for each gene were retained. Likewise, we extracted the GFF3 of all coding genes (both high and low confidence). Mikado `compare` was then used to find the best match for each entry in the former GFF in the latter one. For the purposes of this evaluation, class codes in the TMAP file of `u`,`p` and `P` were considered as intergenic, `X` and `x` as matches on the opposite strand, and finally `i` and `I` as intronic.

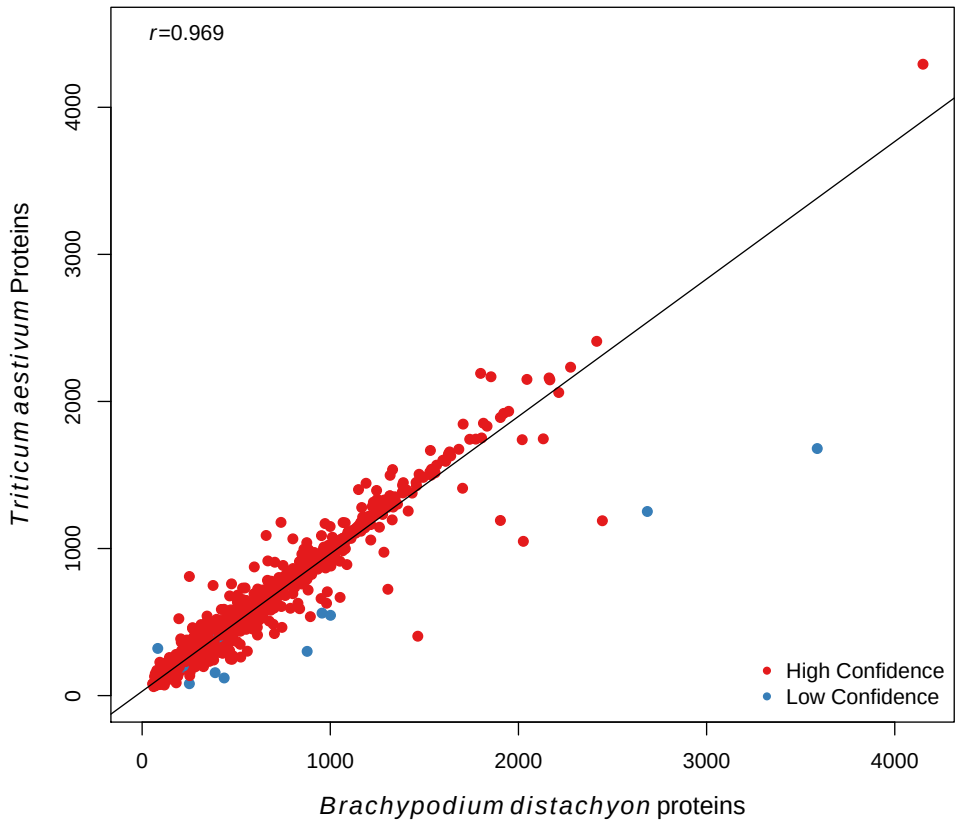

Figure 2: Coherence in gene length between *Triticum aestivum* and *Brachypodium distachyon* proteins. Blast analysis (1×10^{-5}) identified 2686 proteins that had reciprocal best hits to 2707 *Brachypodium distachyon* proteins identified as single copy in *B. distachyon*, *O. sativa*, *S. bicolor*, *S. italica*, *Z. mays* (Phytozome). A high coherence in gene length was found between *Triticum aestivum* and *Brachypodium distachyon*, with a correlation coefficient *r* equal to 0.969.

Alignment against the genomes of progenitors We downloaded the genomes of two progenitors of *Triticum aestivum*, *Triticum urartu* and *Aegilops tauschii*, from EnsEMBL plants release 32. The representative transcripts of the 10,156 high-confidence ncRNA genes of the TGACv1 annotation were aligned against each of these genomes using GMAP v2015-11-20 (Wu and Watanabe, 2005), with the command line options:

```
gmap --no-chimeras -n 5 -f 2 --cross-species
```

The matches were then extracted from the GFF files, filtered for hits with identity and coverage greater than 90%, and merged into a unique list.

1.6 Alternative splicing analysis

RNA-Seq reads generated via the Illumina platform are often too short to cover a full transcript and unambiguously link alternative 5' and 3' splicing events. Furthermore, mapping of relatively short (100–300bp) reads can lead to misalignment and the identification of a substantial number of false positive splice junctions (Sturgill et al., 2013). With different assembly methods showing considerable variation in the number and structure of transcripts assembled we chose to take a conservative approach to annotating alternative splicing in the TGACv1 gene set, giving greater emphasis to long PacBio reads and excluding transcripts with severely truncated coding sequences. To provide a more comprehensive representation of alternative splicing we subsequently integrated transcripts assemblies generated from six strand specific Illumina libraries (Table 1, BioProject accession number PRJEB15048). RNA-Seq transcript assemblies were generated from the six samples using cufflinks (v2.2.1) and subsequently merged via cuffmerge (Roberts et al., 2011b), the TGACv1 gene models were provided as reference annotation. The merged transcripts assemblies were filtered to contain transcripts that are novel isoforms to the TGACv1 annotation, i.e. share at least one splice junction with the reference transcript. Splice variants identified from this additional analysis are provided as a separate track in the Ensembl wheat browser `http://plants.ensembl.org/Triticum_aestivum`, and can be retrieved from the Earlham Institute server (see Section 1.8) In order to analyse different alternative splicing events and to identify transcripts that are susceptible to nonsense mediated decay (NMD), a bioconductor package, spliceR (Vitting-Seerup et al., 2014), was used with the output generated from running cuffdiff (Trapnell et al., 2012).

1.7 Functional annotation of protein coding transcripts

All the proteins of our annotation were annotated using AHRD v.3.1 (Hallab et al., 2014). Sequences were blasted against TAIR10 *A. thaliana* protein sequences (Lamesch et al., 2012) and the plant sequences of UniProt v. 2016_05, both SwissProt and TREMBL datasets (The UniProt Consortium, 2014). Proteins were BLASTed using BLASTP+ v. 2.2.31 asking for a maximum e-value of 1. We adapted the standard example configuration file pathtest/resources/ahrd_example_input.yml, distributed with the AHRD tool, changing the following apart from the location of input and output files:

1. we included the GOA mapping from uniprot,
2. The regular expression used to analyse the TAIR header was amended to correct a parsing error to:

```
^>(?<accession>[aA][tT][0-9mMcC][gG]\\d+(\\.\\d+)?)\\s+\\|  Symbols
    :[^\\|]+\\|\\s+(?<description>([^\\|]+|))(\\s*\\|.*)?\$
```

Concurrently, we analysed the same set of sequences using InterProScan 5.18.57 (Jones et al., 2014). A custom Perl script was used to integrate the ranking, biotype, and functional classification from both tools into a unified file available at: `http://opendata.earlham.ac.uk/Triticum_aestivum/TGAC/v1/annotation/Triticum_aestivum_CS42_TGACv1_scaffold.annotation.gff3.functional_annotation.tsv.gz`.

1.8 Data Access

Sequencing reads generated for this study have been submitted to the European Nucleotide Archive under the accession code PRJEB15048. The annotation is available in Ensembl Plants genomic repository (release 32) at `http://plants.ensembl.org/Triticum_aestivum` and from the Earlham Institute server at `http://opendata.earlham.ac.uk/Triticum_aestivum/TGAC/v1/annotation`. The latter repository contains the following files:

- TGACv1 annotation, in GFF3 format:
  - `Triticum_aestivum_CS42_TGACv1_scaffold.annotation.gff3.gz`
- Sequences for the transcript models of TGACv1 cDNAs, CDS and proteins:
  - `Triticum_aestivum_CS42_TGACv1_scaffold.annotation.gff3.cdna.fa.gz`
  - `Triticum_aestivum_CS42_TGACv1_scaffold.annotation.gff3.cds.fa.gz`
  - `Triticum_aestivum_CS42_TGACv1_scaffold.annotation.gff3.pep.fa.gz`
- Functional annotation of TGACv1 models:
  - `Triticum_aestivum_CS42_TGACv1_scaffold.annotation.gff3.functional_annotation.tsv.gz`
- Annotation of alternative splicing events (see Section 1.6), in both GFF3 and GTF format:

- Triticum_aestivum_CS42_TGACv1_scaffold.annotation.AS.gff3.gz
- Triticum_aestivum_CS42_TGACv1_scaffold.annotation.AS.gtf.gz

2 References

BabrahamLab, 2014. Trim Galore.

Bray, N. L., Pimentel, H., Melsted, P., and Pachter, L., 2016. Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology*, **34**(5):525–527.

Choulet, F., Alberti, A., Theil, S., Glover, N., Barbe, V., Daron, J., Pingault, L., Sourdille, P., Couloux, A., Paux, E., *et al.*, 2014. Structural and functional partitioning of bread wheat chromosome 3B. *Science*, **345**(6194):1249721–1249721.

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T. R., 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, **29**(1):15–21.

Fernandez, N. and Guerrero, D., 2012. Full Lengther Next.

Goodstein, D. M., Shu, S., Howson, R., Neupane, R., Hayes, R. D., Fazo, J., Mitros, T., Dirks, W., Hellsten, U., Putnam, N., *et al.*, 2012. Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Research*, **40**(D1):D1178–D1186.

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., *et al.*, 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology*, **29**(7):644–652.

Haas, B. J., 2003. Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Research*, **31**(19):5654–5666.

Haas, B. J., 2010. TransposonPSI.

Haas, B. J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P. D., Bowden, J., Couger, M. B., Eccles, D., Li, B., Lieber, M., *et al.*, 2013. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature Protocols*, **8**(8):1494–1512.

Hallab, A., Klee, K., Boecker, F., Girish, S., and Schoof, H., 2014. Automated assignment of Humand Readable Descriptions (AHRD).

IWGSC, 2014. A chromosome-based draft sequence of the hexaploid bread wheat (Triticum aestivum) genome. *Science*, **345**(6194):1251788–1251788.

Jones, P., Binns, D., Chang, H.-Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., Nuka, G., *et al.*, 2014. InterProScan 5: genome-scale protein function classification. *Bioinformatics*, **30**(9):1236–1240.

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., and Salzberg, S. L., 2013. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biology*, **14**(4):R36.

Kong, L., Zhang, Y., Ye, Z.-Q., Liu, X.-Q., Zhao, S.-Q., Wei, L., and Gao, G., 2007. CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. *Nucleic Acids Research*, **35**(Web Server issue):W345–9.

Kopylova, E., Noe, L., and Touzet, H., 2012. SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. *Bioinformatics*, **28**(24):3211–3217.

Lamesch, P., Berardini, T. Z., Li, D., Swarbreck, D., Wilks, C., Sasidharan, R., Muller, R., Dreher, K., Alexander, D. L., Garcia-Hernandez, M., *et al.*, 2012. The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. *Nucleic Acids Research*, **40**(D1):D1202–D1210.

Langmead, B. and Salzberg, S. L., 2012. Fast gapped-read alignment with Bowtie 2. *Nature Methods*, **9**(4):357–359.

Mapleson, D. L., Venturini, L., and Swarbreck, D., 2016. Portcullis. `https://github.com/maplesond/portcullis`.

Pertea, M., Pertea, G. M., Antonescu, C. M., Chang, T.-C., Mendell, J. T., and Salzberg, S. L., 2015. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature Biotechnology*, **33**(3):290–295.

Pfeifer, M., Kugler, K. G., Sandve, S. R., Zhan, B., Rudi, H., Hvidsten, T. R., International Wheat Genome Sequencing Consortium, Mayer, K. F. X., and Olsen, O.-A. O.-A., 2014. Genome interplay in the grain transcriptome of hexaploid bread wheat. *Science*, **345**(6194):1250091.

Roberts, A., Pimentel, H., Trapnell, C., and Pachter, L., 2011a. Identification of novel transcripts in annotated genomes using RNA-Seq. *Bioinformatics*, **27**(17):2325–2329.

Roberts, A., Trapnell, C., Donaghey, J., Rinn, J. L., and Pachter, L., 2011b. Improving RNA-Seq expression estimates by correcting for fragment bias. *Genome Biology*, **12**(3):R22.

Slater, G. S. C. and Birney, E., 2005. Automated generation of heuristics for biological sequence comparison. *BMC bioinformatics*, **6**(1):31.

Song, L., Sabunciyan, S., and Florea, L., 2016. CLASS2: accurate and efficient splice variant annotation from RNA-seq reads. *Nucleic Acids Research*, **44**(10):e98–e98.

Stanke, M., Tzvetkova, A., and Morgenstern, B., 2006. AUGUSTUS at EGASP: using EST, protein and genomic alignments for improved gene prediction in the human genome. *Genome biology*, **7 Suppl 1**(May 2005):S11.1–8.

Sturgill, D., Malone, J. H., Sun, X., Smith, H. E., Rabinow, L., Samson, M.-L., and Oliver, B., 2013. Design of RNA splicing analysis null models for post hoc filtering of Drosophila head RNA-Seq data with the splicing analysis kit (Spanki). *BMC Bioinformatics*, **14**(1):320.

The UniProt Consortium, 2014. Activities at the Universal Protein Resource (UniProt). *Nucleic Acids Research*, **42**(D1):D191–D198.

Trapnell, C., Hendrickson, D. G., Sauvageau, M., Goff, L., Rinn, J. L., and Pachter, L., 2012. Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nature Biotechnology*, **31**(1):46–53.

Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., Salzberg, S. L., Wold, B. J., and Pachter, L., 2010. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology*, **28**(5):511–515.

Venturini, L., Caim, S., Mapleson, D. L., Kaithakottil, G. G., and Swarbreck, D., 2016. Mikado. `https://github.com/lucventurini/mikado`.

Vitting-Seerup, K., Porse, B., Sandelin, A., and Waage, J., 2014. spliceR: an R package for classification of alternative splicing and prediction of coding potential from RNA-seq data. *BMC Bioinformatics*, **15**(1):81.

Wang, L., Wang, S., and Li, W., 2012. RSeQC: quality control of RNA-seq experiments. *Bioinformatics*, **28**(16):2184–2185.

Wu, T. D. and Watanabe, C. K., 2005. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics*, **21**(9):1859–1875.

Wysokar, A., Tibbetts, K., McCown, M., Homer, N., and Fennell, T., 2016. Picard: A set of Java command line tools for manipulating high-throughput sequencing data (HTS) data and formats.