**Table S1**

General architecture and processes of the TriAnnot pipeline V3.5 (full analysis) with a description of each module.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Panels** | **Modules or Steps** | **Software**  (1) | **DataBanks**  (2) | **Tasks** | **Module or Step dependency** | **Remarks** |
| I | 01 | 16,3,4 | 1,2,4,6,7,  8,9,10,11 | Repeats annotation |  | Transposable Elements (TEs) masking + masking of univec hits and *E. coli* hits |
| I | 02 | 2 | 3 | Similarity searches (BLASTX) |  | TEs proteins |
| I | 03 | 15 | 5 | TEs annotation |  |  |
| I | 04 | 21 | 73 | GTtallymer |  |  |
|  | | | | | | |
| II | 05 | 7,8,9,10,  11 | **-** | *ab initio* gene prediction | 01 | Augustus uses a maize or wheat matrix; FGeneSH uses a monocot matrix; GeneMarkHMM uses either wheat or barley or maize or rice matrix; GeneID uses wheat matrix. |
| II | 06 | 2,20 | 12,13,14, 15,16,17, 18,19,20, 21,22,23, 24,25,26, 27,50,53, 56,59,62, 65,70 | Transcripts spliced alignments | 01 | BLASTN/exonerate: spliced alignments for nucleic biological evidences. |
| II | 07 | 2,20 | 41,42,43, 44,45,46, 47,51,54, 57,60,63, 66,67,68, 69,72 | Protein spliced alignments | 01 | BLASTX/exonerate: spliced alignments for protein biological evidences. |
| II | 08 | 1  with 2 & 5 | 47,70,71 | SIMsearch  gene models  similarity only | 05 | Annotation of gene model of Categories 00, 01, 02 and 03. Need a soft-masked genome sequence. |
| II | 08 | 12 | 17,47 | EuGene  gene models  *ab initio* & similarities | 01,05,06,07 | EuGene is used here only as a combiner based on *ab initio* gene prediction (augustus wheat), similarity data (BLASTN/exonerate) against *Triticum aestivum* ESTs, similarity data (BLASTN/exonerate against SIMnuc) and similarity data (BLASTX/exonerate against SIMprot). |
| II | 08 | - | - | Merge between  SIMsearch, EuGene and augustus | 05,08 | For each locus the best category is kept with the following priority:  Cat 0 > Cat 1 > Cat 2 > Cat 3 (SIMsearch) >  Cat 4 (EuGene) > Cat5 (augustus) |
| II | 09 | 2,13 | 29,41,42, 48,54 | **Functional annotation**  (3) | 08 | known function  BLASTP: UniProtKB/Swiss-Prot  >80% identity & >80% coverage  putative function  BLASTP: UniProtKB/Swiss-Prot + UniProtKB/TrEMBL  >45% identity & >50% coverage without hypothetical protein  domain containing protein  PfamA and Gene Ontology (GO)  expressed sequences  TBLASTN: Plant ESTs databank  > 45% identity & >50% coverage  conserved unknown function  BLASTP: UniProtKB/Swiss-Prot + UniProtKB/TrEMBL  45% identity & 50% coverage  hypothetical protein only  hypothetical protein (only *ab initio*) |
| II | 10 | 2,20 | 46,51,54, 57,60,63, 66,67,68, 69,72 | Best hits | 08 | BLASTP/exonerate: spliced alignments for best hit protein.  The pipeline verifies also amino acids gaps leading to check the absence or addition of exon within the gene model |
| II | 11 | 14 | **-** | Protein domains, gene ontology | 08 | Uses only PfamA, Prosit, SMART and Gene Ontology (GO terms) |
|  | | | | | | |
| III | 12 | 2 | 37,38,39, 40,49,52, 55,58,61, 64 | BLASTN against  organelles genomes and  pseudo molecules |  | For Conserved Non Coding Sequences – CNSs  Genome model pseudo molecules: *Arabidopsis thaliana* (TAIR); Rice (IRGSP); Rice (MSU); Maize (MSO); Sorghum (DOE); *Brachypodium distachyon* (JGI). |
| III | 13 | 2 | 51,54,57, 60,63,66 | BLASTX against protein databanks |  | Proteins derived from CDS annotation from Genome models such as: *Arabidopsis thaliana* (TAIR); Rice (IRGSP); Rice (MSU); Maize (MSO); Sorghum (DOE); *Brachypodium distachyon* (JGI) |
| III | 14 | 18 | **-** | tRNA annotation |  |  |
|  | | | | | | |
| IV | 15 | 17 | **-** | SSRs annotation |  | The parameters of TRF are tuned to find mainly microsatellites: minimum score=30; maximum period=5 |

(1) Software number (see table below)

(2) Databanks number (see Supplementary Table 2)

(3) The functional annotation follows a modification of the IWGSC annotation guide line.

List of software used within the TriAnnot pipeline V3.5

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Number** | **Softwares** | **Version** | **URL** | **References** |
| 1 | SIMsearch | AUG 2011 | Developed by T. Itoh’s bioinformatics Group at NIAS, Japan | - |
| 2 | BLAST | 2.2.21 | ftp://ftp.ncbi.nlm.nih.gov/blast/executables/LATEST | ([Altschul et al., 1997](#_ENREF_1)) |
| 3 | wu-BLAST | 2.2.6 | http://www.advbiocomp.com/blast.html | Advanced Biocomputing LLC |
| 4 | cross-match | 0.990329 | http://www.incogen.com/public\_documents/vibe/details/crossmatch.html | Green P. *et al*. |
| 5 | est2genome | EMBOSS\_6.1.0 | ftp://ftp.sanger.ac.uk/pub/pmr/ | ([Mott, 1997](#_ENREF_13)) |
| 6 | Gmap | 2010-03-09 | http://research-pub.gene.com/gmap/ | ([Wu and Watanabe, 2005](#_ENREF_29)) |
| 7 | GeneID | 1.3 | http://www1.imim.es/software/geneid/index.html | ([Guigo et al., 1992](#_ENREF_7)) |
| 8 | GeneMarkHMM gmhmm2 | 2.2a | http://opal.biology.gatech.edu/GeneMark/ | ([Lukashin and Borodovsky, 1998](#_ENREF_12)) ([Lomsadze et al., 2005](#_ENREF_10)) |
| 9 | GeneMarkHMM gmhmm3 | 3.9d | http://opal.biology.gatech.edu/GeneMark/ | ([Lukashin and Borodovsky, 1998](#_ENREF_12)) ([Lomsadze et al., 2005](#_ENREF_10)) |
| 10 | FGeneSH | 3.1.1 | http://linux1.softberry.com/berry.phtml | SoftBerry |
| 11 | augustus | 2.4 | http://augustus.gobics.de/ | ([Stanke and Waack, 2003](#_ENREF_23)) |
| 12 | EuGene | 4.0 | http://eugene.toulouse.inra.fr/ | ([Schiex et al., 2001](#_ENREF_19)) |
| 13 | HMMER | 3.0 | http://hmmer.janelia.org/software | ([Durbin R, 1998](#_ENREF_4)) |
| 14 | InterProsScan | 4.6 | http://www.ebi.ac.uk/interpro/ | ([Zdobnov and Apweiler, 2001](#_ENREF_30)) |
| 15 | TEannot | 1.4 | http://urgi.versailles.inra.fr/index.php/urgi/Tools/REPET | ([Quesneville et al., 2005](#_ENREF_17)) |
| 16 | RepeatMasker | 3.2.6 | http://www.repeatmasker.org/ | ([Smit, 1993](#_ENREF_22)) |
| 17 | TRF | 4 | http://tandem.bu.edu/ | ([Benson, 1999](#_ENREF_2)) |
| 18 | tRNAscan-SE | 1.23 | http://lowelab.ucsc.edu/tRNAscan-SE/ | ([Lowe and Eddy, 1997](#_ENREF_11)) |
| 19 | GBrowse | 1.70 | http://gmod.org/wiki/Main\_Page | Generic Model Organism Database (GMOD) |
| 20 | exonerate | 2.2.0 | http://www.ebi.ac.uk/~guy/exonerate/ | ([Slater and Birney, 2005](#_ENREF_21)) |
| 21 | GTtallymer | 1.3.4 | http://seqanswers.com/wiki/Tallymer | ([Kurtz et al., 2008](#_ENREF_9)) |

**Table S2**

List of databanks used within the TriAnnot pipeline V3.5

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Number** | **Databanks (1)** | **Release** | **http and/or ftp URLs for downloading** | **References** |
|  | **Repeat Databanks** |  |  |  |
| 1 | TREP\_total | 10 | http://wheat.pw.usda.gov/ITMI/Repeats/index.shtml | ([Wicker et al., 2002](#_ENREF_28)) |
| 2 | TREP\_nr |
| 3 | TREP\_prot |
| 4 | TREP\_plus | FCREP-all-1.2 | TREPplus = TREPtotal + F. Choulet annotated TEs  ***Databank not publicly available*** | ([Choulet et al., 2010](#_ENREF_3)) |
| 5 | TREP\_cons | JAN2010 | \* 58 are consensus made from TE copies already in TREP release 10 (for which there is no consensus) and from TE copies found by Choulet *et al.* The number of copies used to make the consensus is in the sequence headers (“msa” stands for “multiple sequence alignment”)  \* 188 are TE copies (already in TREP release 10) belonging to known families for which there are less than 3 copies, thus no consensus are made for these families  \* 63 are TE copies (already in TREP release 10) that are “unnamed”, thus no consensus are made for these families  \* 140 are TE copies (Choulet *et al.*) that are “unnamed”, thus no consensus are made for these families  \* 91 are consensus already in TREP release 10  ***Databank not publicly available*** | ([Flutre et al., 2011](#_ENREF_6)) |
| 6 | TIGR\_Fam\_Repeats | V2-V3.3 | *Gramineae, Fabaceae, Solanaceae, Brassicaceae* ftp://ftp.tigr.org/pub/data/TIGR\_Plant\_Repeats/ | TIGR, US |
| 7 | TIGR\_GSS\_Repeats | JUL2004 | *Arabidopsis, Brassica, Glycine, Lotus, Lycopersicon, Medicago, Oryza, Sorghum, Triticum, Zea* ftp://ftp.tigr.org/pub/data/TIGR\_Plant\_Repeats/TIGR\_GSS\_Derived\_Repeats/ |
| 8 | MIPS\_repeat | 8.3 | ***Databank not publicly available*** | K. Mayer (MIPS) |
| 9 | ALL\_Repbase | 16.03 | http://www.girinst.org/server/RepBase/index.php | ([Jurka et al., 2005](#_ENREF_8)) |
| 10 | univec | 6.0 | http://www.ncbi.nlm.nih.gov/VecScreen/UniVec.html  ftp://ftp.ncbi.nih.gov/pub/UniVec/README.uv | NCBI-GenBank  Kitts *et al.* unpublished |
| 11 | Ecoli | 107 | http://srs6.ebi.ac.uk/srsbin/cgi-bin/wgetz?-page+srsq2+-noSession embl-Molecule: genomic dna embl-Class: std embl-Description: complete&genome  embl-Organism: Escherichia&coli | EBI-EMBL, UK |
|  | **Nucleic Databanks** |  |  |  |
| 12 | PlnCpltCDS (3) | 107 | http://srs6.ebi.ac.uk/srsbin/cgi-bin/wgetz?-page+srsq2+-noSession embl-Division: pln embl-Molecule: mrna embl-Class: std embl-Description: complete&cds | EBI-EMBL, UK |
| 13 | PlnMrnaSTD (3) | http://srs6.ebi.ac.uk/srsbin/cgi-bin/wgetz?-page+srsq2+-noSession embl-Division: pln embl-Molecule: mrna embl-Class: std | EBI-EMBL, UK |
| 14 | PoaMrnaSTD (3) | http://srs6.ebi.ac.uk/srsbin/cgi-bin/wgetz?-page+srsq2+-noSession embl-Division: pln embl-Molecule: mrna embl-Class: std embl-Taxon: Poaceae | EBI-EMBL, UK |
| 15 | PlnFL (3) | http://srs6.ebi.ac.uk/srsbin/cgi-bin/wgetz?-page+srsq2+-noSession embl-Division: pln embl-Molecule: mrna embl-Class: std embl-Description: full&insert or FLI\_CDNA | EBI-EMBL, UK |
| 16 | TaFL (3) | http://srs6.ebi.ac.uk/srsbin/cgi-bin/wgetz?-page+srsq2+-noSession embl-Division: pln embl-Molecule: mrna embl-Class: std embl-Description: FLI\_CDNA  embl-Taxon: Triticum|Aegilops  http://trifldb.psc.riken.jp/download.pl for 6146 FLcDNA http://trifldb.psc.riken.jp/download/ver.2.0/TaRFL5740.fas.gz for 5740 FLcDNA  http://trifldb.psc.riken.jp/download/ver.3.0/TaRFL4905.fas.gz for 4905 FLcDNA | EBI-EMBL, UK     Riken, Japan |
| 17 | TaEST (3) | http://srs6.ebi.ac.uk/srsbin/cgi-bin/wgetz?-page+srsq2+-noSession  embl-Division: pln embl-Organism: Triticum&aestivum embl-Class: est | EBI-EMBL, UK |
| 18 | HvFL (3) | http://srs6.ebi.ac.uk/srsbin/cgi-bin/wgetz?-page+srsq2+-noSession embl-Division: pln embl-Molecule: mrna embl-Class: std embl-Description: FLI\_CDNA  embl-Taxon: Hordeum |
| 19 | HvEST (3) | http://srs6.ebi.ac.uk/srsbin/cgi-bin/wgetz?-page+srsq2+-noSession  embl-Division: pln embl-Organism: Hordeum&vulgare embl-Class: est |
| 20 | AT\_unigene (3) | 70 | http://www.ncbi.nlm.nih.gov/unigene  ftp://ftp.ncbi.nih.gov/repository/UniGene/Arabidopsis\_thaliana/ | ([Wheeler et al., 2003](#_ENREF_27)) |
| 21 | HV\_unigene (3) | 56 | http://www.ncbi.nlm.nih.gov/unigene  ftp://ftp.ncbi.nih.gov/repository/UniGene/Hordeum\_vulgare/ |
| 22 | OS\_unigene (3) | 80 | http://www.ncbi.nlm.nih.gov/unigene  ftp://ftp.ncbi.nih.gov/repository/UniGene/Oryza\_sativa/ |
| 23 | TA\_unigene (3) | 56 | http://www.ncbi.nlm.nih.gov/unigene  ftp://ftp.ncbi.nih.gov/repository/UniGene/Triticum\_aestivum/ |
| 24 | ZM\_unigene (3) | 77 | http://www.ncbi.nlm.nih.gov/unigene  ftp://ftp.ncbi.nih.gov/repository/UniGene/Zea\_mays/ |
| 25 | SB\_unigene (3) | 29 | http://www.ncbi.nlm.nih.gov/unigene  ftp://ftp.ncbi.nih.gov/repository/UniGene/Sorghum\_bicolor/ |
| 26 | SO\_unigene (3) | 14 | http://www.ncbi.nlm.nih.gov/unigene  ftp://ftp.ncbi.nih.gov/repository/UniGene/Saccharum\_officinarum/ |
| 27 | HS\_unigene (3) | 224 | http://www.ncbi.nlm.nih.gov/unigene  ftp://ftp.ncbi.nih.gov/repository/UniGene/Homo\_sapiens/ |
| 28 | embl\_pl | 107 | ftp://ftp.ebi.ac.uk/pub/databases/embl/release/ | EBI-EMBL class std, UK |
| 29 | embl\_est\_pln | EBI-EMBL class est, UK |
| 30 | embl\_pat\_pln | EBI-EMBL class pat, UK |
| 31 | embl\_sts\_pln | EBI-EMBL class sts, UK |
| 32 | embl\_gss\_pln | EBI-EMBL class gss, UK |
| 33 | embl\_htg\_pln | EBI-EMBL class htgs, UK |
| 34 | embl\_htc\_pln | EBI-EMBL class htc, UK |
| 35 | embl\_tsa\_pln | EBI-EMBL class tsa, UK |
| 36 | embl\_wgs\_pln | EBI-EMBL class wgs, UK |
| 37 | WH\_embl\_mito | AP008982 | http://www.ncbi.nlm.nih.gov/nuccore/78675232? | ([Ogihara et al., 2005](#_ENREF_15)) |
| 38 | WH\_embl\_chloro | AB042240 | http://www.ncbi.nlm.nih.gov/nuccore/13928184? | ([Ogihara et al., 2002](#_ENREF_14)) |
| 39 | refSeq\_chloro | APR2011 | ftp://ftp.ncbi.nlm.nih.gov/refseq/release/README  ftp://ftp.ncbi.nlm.nih.gov/refseq/release/plastid/ | NCBI-GenBank |
| 40 | refSeq\_mito | ftp://ftp.ncbi.nlm.nih.gov/refseq/release/README  ftp://ftp.ncbi.nlm.nih.gov/refseq/release/mitochondrion/ |
|  | **Protein Databanks** |  |  |  |
| 41 | uniprot\_sprot | 2011\_04 | http://us.expasy.org/sprot/  ftp://ftp.uniprot.org/pub/databases/uniprot/README | ExPASy Proteomics Server |
| 42 | uniprot\_trembl |
| 43 | uniprot\_uniref100 | http://us.expasy.org/sprot/  ftp://ftp.uniprot.org/pub/databases/uniprot/uniref/uniref100/ |
| 44 | uniprot\_uniref90 | http://us.expasy.org/sprot/  ftp://ftp.uniprot.org/pub/databases/uniprot/uniref/uniref90/ |
| 45 | uniprot\_uniref50 | http://us.expasy.org/sprot/  ftp://ftp.uniprot.org/pub/databases/uniprot/uniref/uniref50/ |
| 46 | nr | APR2011 | ftp://ftp.ncbi.nlm.nih.gov/blast/db/FASTA/  ftp://ftp.ncbi.nlm.nih.gov/blast/db/README | NCBI |
| 47 | refSeqPlnProt | ftp://ftp.ncbi.nlm.nih.gov/refseq/release/plant/  ftp://ftp.ncbi.nlm.nih.gov/refseq/release/README |
|  | **Protein Domains** |  |  |  |
| 48 | PfamA | 25 | http://pfam.sanger.ac.uk/ | ([Sammut et al., 2008](#_ENREF_18))  ([Finn et al., 2010](#_ENREF_5)) |
|  | **Model Genomes & Proteomes** |  |  |  |
|  | *Arabidopsis thaliana* |  |  | **TAIR**  http://www.arabidopsis.org/  ([The Arabidopsis Genome Initiative, 2000](#_ENREF_25)) |
| 49 | genoAtTAIR | 9 | ftp://ftp.arabidopsis.org/home/tair/Genes/TAIR9\_genome\_release/TAIR9\_chr\_all.fas | - |
| 50 | cdsAtTAIR (3) | 10 | ftp://ftp.arabidopsis.org/home/tair/Genes/TAIR9\_genome\_release/ TAIR10\_sequences/TAIR10\_cds\_20101214 | - |
| 51 | protAtTAIR | ftp://ftp.arabidopsis.org/home/tair/Genes/TAIR9\_genome\_release/ TAIR10\_sequences/TAIR10\_pep\_20101214 | - |
|  | *Oryza sativa*  **MSU** |  | http://rice.plantbiology.msu.edu/annotation\_pseudo\_current.shtml | **MSU**  http://rice.plantbiology.msu.edu/ |
| 52 | genoOsMSU | 6.1 | ftp://ftp.plantbiology.msu.edu/pub/data/Eukaryotic\_Projects/o\_sativa/annotation\_dbs/pseudomolecules/version\_6.1/all.dir/all.con | - |
| 53 | cdsOsMSU (3) | ftp://ftp.plantbiology.msu.edu/pub/data/Eukaryotic\_Projects/o\_sativa/annotation\_dbs/pseudomolecules/version\_6.1/all.dir/all.cds | - |
| 54 | protOsMSU | ftp://ftp.plantbiology.msu.edu/pub/data/Eukaryotic\_Projects/o\_sativa/annotation\_dbs/pseudomolecules/version\_6.1/all.dir/all.pep | - |
|  | *Oryza sativa*  **IRGSP** |  |  | **IRGSP**  http://rapdb.dna.affrc.go.jp/  ([Tanaka et al., 2008](#_ENREF_24)) |
| 55 | genoOsIRGSP | 5 | http://rapdb.dna.affrc.go.jp/download/latest/   🡺 file IRGSPb5.fa.masked.gz | - |
| 56 | cdsOsIRGSP (3) | http://rapdb.dna.affrc.go.jp/download/latest/rap\_representative.tgz  🡺 file representative\_orf\_nuc.fa  http://rapdb.dna.affrc.go.jp/download/latest/rap\_predicted.tar.gz   🡺 file predicted\_orf\_nuc.fa | - |
| 57 | protOsIRGSP | http://rapdb.dna.affrc.go.jp/download/latest/rap\_representative.tgz   🡺 file representative\_orf.fa  http://rapdb.dna.affrc.go.jp/download/latest/rap\_predicted.tar.gz   🡺 file predicted\_orf.fa | - |
|  | *Zea mays* |  |  | **MaizeSequence.org**  http://maizesequence.org/index.html  ([Schnable et al., 2009](#_ENREF_20)) |
| 58 | genoZmMSO | 5b | http://ftp.maizesequence.org/current/assembly/ ZmB73\_RefGen\_v2.masked.tar.gz | - |
| 59 | cdsZmMSO (3) | http://ftp.maizesequence.org/current/filtered-set/ ZmB73\_5b\_FGS\_cds.fasta.gz | - |
| 60 | protZmMSO | http://ftp.maizesequence.org/current/filtered-set/ ZmB73\_5b\_FGS\_translations.fasta.gz | - |
|  | *Sorghum bicolor* |  | ftp://ftp.jgi-psf.org/pub/JGI\_data/phytozome/v6.0/Sbicolor/README | **Phytozome**  http://www.phytozome.net/sorghum.php  ([Paterson et al., 2009](#_ENREF_16)) |
| 61 | genoSbDOE | 6 | ftp://ftp.jgi-psf.org/pub/JGI\_data/phytozome/ v6.0/Sbicolor/assembly/sbi1\_n\_repeatMasker.fasta.gz | - |
| 62 | cdsSbDOE (3) | ftp://ftp.jgi-psf.org/pub/JGI\_data/phytozome/ pub/JGI\_data/phytozome/v6.0/Sbicolor/annotation/Sbicolor\_79\_cds.fa.gz | - |
| 63 | protSbDOE | ftp://ftp.jgi-psf.org/pub/JGI\_data/phytozome/ pub/JGI\_data/phytozome/v6.0/Sbicolor/annotation/Sbicolor\_79\_peptide.fa.gz | - |
|  | *Brachypodium distachyon* |  |  | **Orego State U.**  http://www.brachypodium.org/  ([The International Brachypodium Initiative, 2010](#_ENREF_26)) |
| 64 | genoBdJGI | 8x | http://ftp.brachypodium.org/ files/8X\_ASSEMBLY/Brachypodium\_distachyon\_Bd21.main\_genome.scaffolds.fasta.tar.gz | - |
| 65 | cdsBdJGI (3) | 1.2 | http://ftp.brachypodium.org/ files/Annotation/brachypodium\_1.2\_CDS.fa | - |
| 66 | protBdJGI | 1.2 | http://ftp.brachypodium.org/ files/Annotation/brachypodium\_1.2\_Protein.fa | - |
|  | *Triticum* |  |  |  |
| 67 | ProtTRI | 107 | Using a homemade perl program for SRS request (4) | EBI-EMBL uniprot |
|  | *Hordeum* |  |  |  |
| 68 | ProtHOR | 107 | Using a homemade perl program for SRS request (4) | EBI-EMBL uniprot |
|  | *Saccharum* |  |  |  |
| 69 | ProtSAC | 107 | Using a homemade perl program for SRS request (4) | EBI-EMBL uniprot |
|  | **Special Databanks** |  |  |  |
| 70 | SIMnuc | JUN2011 | Expertise\_CDS (CAT00) + TaFL(CAT01) + PoaFL(CAT02) + CDSplus (CAT03) (3) | Local databank |
| 71 | SIMprot | RefSeqPlnProt + protBdJGI + protOsIRGSP + ProtTRI + ProtHOR |
| 72 | protPlant | Using a homemade perl program for SRS request (4)  with Taxon = *Magnoliophyta* |
| 73 | CS3B2X\_OCC1 | 2011 | Wheat cv. Chinese Spring Chromosome 3B 2X 454 sequences | Local databank  ([Choulet et al., 2010](#_ENREF_3)) |

(1) Databank names used within the TriAnnot pipeline

(3) These databank are post processed: polyA tails are deleted; sequences are strong-masked with RepeatMasker against the univec databank (N on 5’ and 3’ ends are deleted), and soft-masked with RepeatMasker against TREPplus. CAT00 is related to mRNA/CDS which have been already manually expertised.

(4) (([uniprot-Organism:$species\*]) & ((([uniprot-ProteinExistence:1: evidence at protein level] | [uniprot-ProteinExistence:2: evidence at transcript level]) | [uniprot-ProteinExistence:3: inferred from homology]) | [uniprot-ProteinExistence:4: predicted])). $species being the genus of interest (for example: *Arabidopsis, Triticum, Hordeum, Oryza, Zea, Sorghum, Saccharum, Brachypodium*).

**Table S3**

Parameters for TriAnnot V3.5 default analysis (step.xml).

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Panel** | **Modules or steps** | **Softwares**  (1) | **DataBanks**  (2) | **Topic** | **Sequence status and threshold of main parameters** |
| I | 01 | 16 | 4 | TEs annotation and masking | Initial sequence  RepeatMasker cutoff 225  ‘qq’ parameter (rush job) about 10% less sensitive, 4 to 10 times faster than default repeat options  masked with lower-case and Ns |
| I | 02 | 2 | 3 | BLASTX | Initial sequence  Evalue = 10-5 |
|  | | | | | |
| II | 05 | 11 | - | *ab initio* gene prediction | Masked TEs (N-masked) sequence  Wheat matrix |
| II | 06 | 2, 20 | 16,17,18,53, 56,59,62,65,70 | Spliced transcript alignments. BLASTN followed by exonerate. | Masked TEs (soft-masked) sequence  For TaEST databank:  BLASTN: Evalue = 10-5  Identity = 95%  Coverage = 95%  exonerate coverage threshold = 95%  For all other databanks:  BLASTN: Evalue = 10-5  Identity = 80%  Coverage = 70%  exonerate coverage threshold = 80% |
| II | 07 | 2 & 20 | 41,47,54,57,60,63,66,67,68,71 | Spliced transcript alignments. BLASTX followed by exonerate | Masked TEs (soft-masked) sequence  BLASTX: Evalue = 10-5  Identity = 70%  Coverage = 60%  exonerate coverage threshold = 70% |
| II | 08 | 1,12 | 10,13,48,72 | SIMsearch  gene models  similarity only | Masked TEs (soft-masked) sequence  Priority = CAT,CV\_NA,ORF\_LEN,ID\_AA,ID\_NA,RANGE,EXON  3’/5’ extension on |
| II | 08 | 12 | 17,70,71 | EuGene  gene models  *ab initio* & similarities | Masked TEs (specific IG file for EuGene)  Combines: augustus\_wheat and BLASTN/exonerate of TaEST, SIMnuc and BLASTX/exonerate of SIMprot |
| II | 09 | 2,13 | 29,41,42, 48,54 | Functional annotation | On proteins derived from the best gene models obtained at step08 |
| II | 10 | 2 | 46,51,54,57,60,63,66,67,68,69,72 | Best hits | On proteins derived from the best gene models obtained at step08  BLASTp: Evalue = 10-5  Positive = 50%  Coverage = 60%  Exonerate coverage threshold = 50% |
| II | 11 | 14 | - | Protein domains and gene ontology annotations | On proteins derived from the best gene models obtained at step08  InterProScan default parameters (Pfam,Prosite,SMART & GO) |
|  | | | | | |
| III | 12 | 2 | 39,40,49,52, 55,58,61,64 | CNSs  BLASTN against  organelles genomes and  pseudo molecules | Masked TEs and Genes (N-masked) sequence  BLASTN: Evalue = 10-5 |
| III | 13 | 2 | 51,54,57, 60,63,66 | BLASTX against protein databanks | Masked TEs and Genes (N-masked) sequence  BLASTN: Evalue = 10-5 |
| III | 16 | 18 | - | Annotation of tRNA | Initial sequence  tRNAscan default parameters |
|  |  |  |  |  |  |
| IV | 19 | 17 | - | Microsatellite (SSR) annotations | Initial sequence  match = 2; mismatch = 7; delta = 7; pm = 80; pi = 10; minscore = 30;  maxPeriod = 5 |

(1) Software number (**see Supplementary table S1**)

(2) Databanks number (**see Supplementary table S2**)

References

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25**,** 3389-3402.

Benson, G. (1999). Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. 27**,** 573-580.

Choulet, F., Wicker, T., Rustenholz, C., Paux, E., Salse, J., Leroy, P., Schlub, S., Le Paslier, M.C., Magdelenat, G., Gonthier, C., Couloux, A., Budak, H., Breen, J., Pumphrey, M., Liu, S., Kong, X., Jia, J., Gut, M., Brunel, D., Anderson, J.A., Gill, B.S., Appels, R., Keller, B., and Feuillet, C. (2010). Megabase level sequencing reveals contrasted organization and evolution patterns of the wheat gene and transposable element spaces. Plant Cell 22**,** 1686-1701.

Durbin R, E.S., Krogh a, Mitchison G (ed.). (1998). *Biological sequence analysis: probabilistic models of proteins and nucleic acids.*: Cambridge University Press.

Finn, R.D., Mistry, J., Tate, J., Coggill, P., Heger, A., Pollington, J.E., Gavin, O.L., Gunasekaran, P., Ceric, G., Forslund, K., Holm, L., Sonnhammer, E.L., Eddy, S.R., and Bateman, A. (2010). The Pfam protein families database. Nucleic Acids Res. 38**,** D211-222.

Flutre, T., Duprat, E., Feuillet, C., and Quesneville, H. (2011). Considering transposable element diversification in de novo annotation approaches. PLoS ONE 6**,** e16526.

Guigo, R., Knudsen, S., Drake, N., and Smith, T. (1992). Prediction of gene structure. J. Mol. Biol. 226**,** 141-157.

Jurka, J., Kapitonov, V.V., Pavlicek, A., Klonowski, P., Kohany, O., and Walichiewicz, J. (2005). Repbase Update, a database of eukaryotic repetitive elements. Cytogenet Genome Res. 110**,** 462-467.

Kurtz, S., Narechania, A., Stein, J.C., and Ware, D. (2008). A new method to compute K-mer frequencies and its application to annotate large repetitive plant genomes. BMC Genomics 9**,** 517.

Lomsadze, A., Ter-Hovhannisyan, V., Chernoff, Y.O., and Borodovsky, M. (2005). Gene identification in novel eukaryotic genomes by self-training algorithm. Nucleic Acids Res. 33**,** 6494-6506.

Lowe, T.M., and Eddy, S.R. (1997). tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res. 25**,** 955-964.

Lukashin, A.V., and Borodovsky, M. (1998). GeneMark.hmm: new solutions for gene finding. Nucleic Acids Res. 26**,** 1107-1115.

Mott, R. (1997). EST\_GENOME: a program to align spliced DNA sequences to unspliced genomic DNA. Comput. Appl. Biosci. 13**,** 477-478.

Ogihara, Y., Isono, K., Kojima, T., Endo, A., Hanaoka, M., Shiina, T., Terachi, T., Utsugi, S., Murata, M., Mori, N., Takumi, S., Ikeo, K., Gojobori, T., Murai, R., Murai, K., Matsuoka, Y., Ohnishi, Y., Tajiri, H., and Tsunewaki, K. (2002). Structural features of a wheat plastome as revealed by complete sequencing of chloroplast DNA. Mol. Genet. Genomics 266**,** 740-746.

Ogihara, Y., Yamazaki, Y., Murai, K., Kanno, A., Terachi, T., Shiina, T., Miyashita, N., Nasuda, S., Nakamura, C., Mori, N., Takumi, S., Murata, M., Futo, S., and Tsunewaki, K. (2005). Structural dynamics of cereal mitochondrial genomes as revealed by complete nucleotide sequencing of the wheat mitochondrial genome. Nucleic Acids Res. 33**,** 6235-6250.

Paterson, A.H., Bowers, J.E., Bruggmann, R., Dubchak, I., Grimwood, J., Gundlach, H., Haberer, G., Hellsten, U., Mitros, T., Poliakov, A., Schmutz, J., Spannagl, M., Tang, H., Wang, X., Wicker, T., Bharti, A.K., Chapman, J., Feltus, F.A., Gowik, U., Grigoriev, I.V., Lyons, E., Maher, C.A., Martis, M., Narechania, A., Otillar, R.P., Penning, B.W., Salamov, A.A., Wang, Y., Zhang, L., Carpita, N.C., Freeling, M., Gingle, A.R., Hash, C.T., Keller, B., Klein, P., Kresovich, S., Mccann, M.C., Ming, R., Peterson, D.G., Mehboob Ur, R., Ware, D., Westhoff, P., Mayer, K.F.X., Messing, J., and Rokhsar, D.S. (2009). The Sorghum bicolor genome and the diversification of grasses. Nature 457**,** 551-556.

Quesneville, H., Bergman, C.M., Andrieu, O., Autard, D., Nouaud, D., Ashburner, M., and Anxolabehere, D. (2005). Combined evidence annotation of transposable elements in genome sequences. PLoS Comput. Biol. 1**,** 166-175.

Sammut, S.J., Finn, R.D., and Bateman, A. (2008). Pfam 10 years on: 10,000 families and still growing. Brief Bioinform. 9**,** 210-219.

Schiex, T., Moisan, A., and Rouzé, P. (2001). "EuGene: An Eucaryotic Gene Finder that combines several sources of evidence.," in *Computational Biology*. Gascuel, O., Sagot, M-F.), 111-125.

Schnable, P.S., Ware, D., Fulton, R.S., Stein, J.C., Wei, F., Pasternak, S., Liang, C., Zhang, J., Fulton, L., Graves, T.A., Minx, P., Reily, A.D., Courtney, L., Kruchowski, S.S., Tomlinson, C., Strong, C., Delehaunty, K., Fronick, C., Courtney, B., Rock, S.M., Belter, E., Du, F., Kim, K., Abbott, R.M., Cotton, M., Levy, A., Marchetto, P., Ochoa, K., Jackson, S.M., Gillam, B., Chen, W., Yan, L., Higginbotham, J., Cardenas, M., Waligorski, J., Applebaum, E., Phelps, L., Falcone, J., Kanchi, K., Thane, T., Scimone, A., Thane, N., Henke, J., Wang, T., Ruppert, J., Shah, N., Rotter, K., Hodges, J., Ingenthron, E., Cordes, M., Kohlberg, S., Sgro, J., Delgado, B., Mead, K., Chinwalla, A., Leonard, S., Crouse, K., Collura, K., Kudrna, D., Currie, J., He, R., Angelova, A., Rajasekar, S., Mueller, T., Lomeli, R., Scara, G., Ko, A., Delaney, K., Wissotski, M., Lopez, G., Campos, D., Braidotti, M., Ashley, E., Golser, W., Kim, H., Lee, S., Lin, J., Dujmic, Z., Kim, W., Talag, J., Zuccolo, A., Fan, C., Sebastian, A., Kramer, M., Spiegel, L., Nascimento, L., Zutavern, T., Miller, B., Ambroise, C., Muller, S., Spooner, W., Narechania, A., Ren, L., Wei, S., Kumari, S., Faga, B., Levy, M.J., Mcmahan, L., Van Buren, P., Vaughn, M.W., et al. (2009). The B73 maize genome: complexity, diversity, and dynamics. Science 326**,** 1112-1115.

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

Smit, A.F. (1993). Identification of a new, abundant superfamily of mammalian LTR-transposons. Nucleic Acids Res. 21**,** 1863-1872.

Stanke, M., and Waack, S. (2003). Gene prediction with a hidden Markov model and a new intron submodel. Bioinformatics 19 Suppl 2**,** ii215-225.

Tanaka, T., Antonio, B.A., Kikuchi, S., Matsumoto, T., Nagamura, Y., Numa, H., Sakai, H., Wu, J., Itoh, T., Sasaki, T., Aono, R., Fujii, Y., Habara, T., Harada, E., Kanno, M., Kawahara, Y., Kawashima, H., Kubooka, H., Matsuya, A., Nakaoka, H., Saichi, N., Sanbonmatsu, R., Sato, Y., Shinso, Y., Suzuki, M., Takeda, J., Tanino, M., Todokoro, F., Yamaguchi, K., Yamamoto, N., Yamasaki, C., Imanishi, T., Okido, T., Tada, M., Ikeo, K., Tateno, Y., Gojobori, T., Lin, Y.C., Wei, F.J., Hsing, Y.I., Zhao, Q., Han, B., Kramer, M.R., Mccombie, R.W., Lonsdale, D., O'donovan, C.C., Whitfield, E.J., Apweiler, R., Koyanagi, K.O., Khurana, J.P., Raghuvanshi, S., Singh, N.K., Tyagi, A.K., Haberer, G., Fujisawa, M., Hosokawa, S., Ito, Y., Ikawa, H., Shibata, M., Yamamoto, M., Bruskiewich, R.M., Hoen, D.R., Bureau, T.E., Namiki, N., Ohyanagi, H., Sakai, Y., Nobushima, S., Sakata, K., Barrero, R.A., Sato, Y., Souvorov, A., Smith-White, B., Tatusova, T., An, S., An, G., S, O.O., Fuks, G., Fuks, G., Messing, J., Christie, K.R., Lieberherr, D., Kim, H., Zuccolo, A., Wing, R.A., Nobuta, K., Green, P.J., Lu, C., Meyers, B.C., Chaparro, C., Piegu, B., Panaud, O., and Echeverria, M. (2008). The Rice Annotation Project Database (RAP-DB): 2008 update. Nucleic Acids Res. 36**,** D1028-1033.

The Arabidopsis Genome Initiative (2000). Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature 408**,** 796-814.

The International Brachypodium Initiative (2010). Genome sequencing and analysis of the model grass Brachypodium distachyon. Nature 463**,** 763-768.

Wheeler, D.L., Church, D.M., Federhen, S., Lash, A.E., Madden, T.L., Pontius, J.U., Schuler, G.D., Schriml, L.M., Sequeira, E., Tatusova, T.A., and Wagner, L. (2003). Database resources of the National Center for Biotechnology. Nucleic Acids Res. 31**,** 28-33.

Wicker, T., Matthews, D.E., and Keller, B. (2002). TREP: a database for Triticeae repetitive elements. Trends Plant Sci. 7**,** 561-562.

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

Zdobnov, E.M., and Apweiler, R. (2001). InterProScan--an integration platform for the signature-recognition methods in InterPro. Bioinformatics 17**,** 847-848.