Supplementary Methods

InterProScan

InterProScan was run as below to identify and classify protein functional domains.

```
cat $PROTEIN \
    | paste - - \
    | grep -v "\*" \
    | sed 's/\t/\n/g' \
    | parallel -j $NSLOTS --pipe --block 100k --recstart '>' \
    "nice interproscan.sh -T /run/shm/ -i - -d $OUTDIR -dp -t p -appl
TIGRFAM-13.0,ProDom-2006.1,SMART-6.2,SignalP-EUK-4.0,PrositePatterns-20.97,PRIN
TS-42.0,SuperFamily-1.75,Gene3d-3.5.0,PfamA-27.0,PrositeProfiles-20.97,Phobius-
1.01,TMHMM-2.0c,Coils-2.2 -f TSV"
cat -- $OUTDIR/* > $PROTEIN.interproscan
gzip $PROTEIN.interproscan
```

Result files were imported into the Lepbase Ensembl instance for visualisation on protein sequences and are available for download at http://download.lepbase.org/v2/interproscan/.

blastp

Protein fasta files at http://download.lepbase.org/v2/sequence/, were compared against the UniProt/SwissProt database using blastp with the command below to identify matches with e-values stronger than $1e^{-10}$. The -outfmt includes the btop (BLAST traceback operations) field, which can be converted to a CIGAR string as a summary of the alignment.

```
cat $PROTEIN \
    | parallel -j $NSLOTS --pipe --block 10k --recstart '>' \
    "nice blastp -query - -db /exports/blast_db/uniprot_sprot.fasta -evalue
1e-10 -outfmt '6 std qlen slen stitle btop'" \
    | gzip >$OUTDIR/$PROTEIN.blastp.uniprot_sprot.1e-10.gz
```

Result files were imported into the Lepbase Ensembl instance to provide additional gene descriptions and are available for download at http://download.lepbase.org/v2/blastp/.

Repeat masking

Repetitive sequences were modeled and masked using RepeatModeler/RepeatMasker. For each assembly scaffold fasta file at http://download.lepbase.org/v2/sequence/, a taxon-specific repeat library was generated using RepeatModeler. This library was filtered using the corresponding protein fasta file to remove any hits to annotated proteins that were not annotated as repeats in RepBase. For two species, proteomes were not available for the same assembly so *Bicyclus anynana* v1.2 repeats were filtered using the nBa.0.1 assembly proteome and *Spodoptera frugiperda* v2 repeats were filtered using the *Bombyx mori* ASM15162v1 proteome. Resulting repeat libraries were combined with annotated

Lepidoptera repeats from RepBase and then used to mask the assembly fasta file. Commands used were:

```
# Run RepeatModeler
BuildDatabase -engine ncbi -name $SCAFFOLDS $SCAFFOLDS.fa
RepeatModeler -database $SCAFFOLDS -engine ncbi -pa $NSLOTS
REPMODLIB=$SCAFFOLDS.fa.consensi.fa.classified

# Blast proteome against RepeatMasker TE database
blastp -query $PROTEINS -dbRepeatPeps.lib \
    -outfmt '6 qseqid staxids bitscore std sscinames sskingdoms stitle' \
    -max_target_seqs 25 -culling_limit 2 -num_threads $NSLOTS -evalue 1e-5 \
    -out $PROTEINS.vs.RepeatPeps.25cul2.1e5.blastp.out

# Remove TEs from proteome
fastaqual_select.pl -f $TRANSCRIPTS \
    -e <(awk '{print $1}' $PROTEINS.vs.RepeatPeps.25cul2.1e5.blastp.out | sort
| uniq) \
    > $TRANSCRIPTS.no_tes.fa

# Blast proteome against RepeatModeler library
makeblastdb -in $TRANSCRIPTS.no_tes.fa -dbtype nucl
blastn -task megablast \
       -query $DATADIR/$REPMODLIB \
       -db $TRANSCRIPTS.no_tes.fa \
       -outfmt '6 qseqid staxids bitscore std sscinames sskingdoms stitle' \
       -max_target_seqs 25 \
       -culling_limit 2 \
       -num_threads $NSLOTS \
       -evalue 1e-10 \
       -out $REPMODLIB.vs.$TRANSCRIPTS.25cul2.1e10.megablast.out

# Remove hits from RepeatModeler library
fastaqual_select.pl -f $REPMODLIB \
    -e <(awk '{print $1}' $REPMODLIB.vs.$TRANSCRIPTS.25cul2.1e10.megablast.out
| sort | uniq) \
    > $REPMODLIB.filtered_for_CDS_repeats.fa

# Add RepBase Lepidoptera repeats to repeat library
queryRepeatDatabase.pl -species "lepidoptera" > repeatmasker.lepidoptera.fa
tail -n+2 repeatmasker.lepidoptera.fa > repeatmasker.lepidoptera.noheader.fa
cat repeatmasker.lepidoptera.noheader.fa \
    $REPMODLIB.filtered_for_CDS_repeats.fa \
    > $SCAFFOLDS.repeatlib.fa

# Run RepeatMasker
RepeatMasker -pa $NSLOTS -lib $DATADIR/$REPLIB -dir . -xsmall \
    $DATADIR/$SEQFILE

# summarise content of .tbl files
```

```
perl -0777 -ne '$ARGV=~s/_-.+//;m/total.+?(\d+).+bases.+?(\d+)/s;print
"$ARGV\t$1\t",($2/$1*100),"\n"' $SCAFFOLDS.fa.tbl \
    > repeat_content_summary.txt
```

Resulting `.out` files were imported into the Lepbase Ensembl instance to provide annotation tracks. These and summarised `.tbl` files have been made available for download at http://download.lepbase.org/v2/repeats/.