This README file contains the descriptions of the bulk download files on the CLiP website.

Basic_mutant_data_table.txt: Full library data: colony positions, insertion sites, deconvolution and LEAP-Seq summary, gene annotation.

Each line is an insertion junction (corresponding to a flanking sequence derived from the 5' or 3' side of an inserted cassette). Note that each mutant may have one or multiple sequenced insertion junctions, which may be two sides of the same insertion, or multiple independent cassette insertions in one mutant. Two sides of the same insertion will usually be 5' and 3', but may sometimes both be 5' or both be 3', if the insertion consisted of two cassettes in reverse orientations. Sometimes there is a genomic DNA fragment on one side of the insertion - approximately 25% insertion junction positions are incorrect due to this and other factors.

The columns for each line are:

- mutant_ID – the ID assigned to the colony (note that LMJ.SG and LMJ.RY mutants are from two different libraries, with different cassettes and somewhat different transformation protocols).
- side – the side of the cassette that the flanking sequence in this line was derived from.
- chromosome – the chromosome to which the insertion was mapped.
- strand – '+' if the insertion cassette's resistance gene is transcribed in the + chromosome direction, '-' otherwise.
- min_position – the genomic base immediately before the insertion (either sequenced directly or inferred based on the flanking sequence from the other side), 1-based.
- full_position – the bases immediately before and after the insertion, with the sequenced side given a number and the other (inferred) side shown as '?'
- gene – the Phytozome ID of the gene containing the insertion. If the insertion junction is inside two overlapping genes, they're separated with &.
- orientation – 'sense' if the insertion cassette's resistance gene is transcribed in the same direction as the gene, 'antisense' otherwise. For overlapping genes, there are two &-separated values.
- feature – which feature of the gene the insertion is in. For genes with multiple splice variants, this wasn't determined. For overlapping genes, there are two &-separated values. For insertions precisely between two features, the two features are separated with /.
- intergenic_adjacent_genes – for intergenic insertions, IDs of the nearest genes before and after the insertion.
- intergenic_adj_orientations – whether the intergenic insertion is upstream or downstream of each of the adjacent genes, relative to the gene orientations.
- intergenic_adj_distances – distances of the intergenic insertion from the nearest edges of the adjacent genes.
- transcript_names – the common name of the gene containing the insertion (or adjacent to the insertion if the insertion is intergenic), from Phytozome. If the insertion junction is inside two overlapping genes, or is intergenic and bordering two genes, they're separated with &. Some genes have no common names.
- genome_version – the version of the *Chlamydomonas* genome the gene information was derived from – can be different for different libraries.
- flanking_seq – the genomic sequence immediately flanking the cassette, shown in the orientation from the cassette junction outwards (may match the genome sequence or be a reverse-complement; may have 1 bp mismatch compared to reference genome sequence.)
- IB – the internal barcode on the given side of this insertion cassette. Only present for LMJ.RY series mutants.
- mapping_confidence – what % of insertions in this category are mapped correctly, based on check PCRs of ~20 randomly chosen mutants
- if_both_insertion_sides – whether this flanking sequence has a matching flanking sequence from the other side of the same insertion (perfect insertion, or with a deletion or duplication, or with a corrected junk fragment on one side) – there will always be a matching pair of insertion junctions with the same value, corresponding to the two sides of the insertion.
- if_fixed_position – whether the insertion position was corrected based on the conclusion that the flanking sequence was part of a "junk" genomic DNA fragment co-inserted with the cassette. This conclusion was made if most of the LEAP-Seq distal reads were located in a different region than the proximal read. If the value here is "yes", several fields are different than usual:

  - the flanking_seq field is not the sequence immediately flanking the cassette, but the distal LEAP-Seq read closest to the presumed true insertion site.
  - all the insertion position information (chromosome, strand, min_position, full_position, gene, orientation, feature, adjacent gene data, all gene annotation) is based on the mapping position of the above LEAP-Seq read, not the flanking sequence.
  - if the new insertion position is mapped to the same locus as another insertion in the same mutant (with positions/orientations consistent with the two being the two sides of a single inserted cassette), the if_both_insertion_sides field for both lines is set to "with-junk"
  - LEAP-Seq_confirmed_distance will instead be the highest distance between two LEAP-Seq distal reads mapping to the new position locus
  - the remaining LEAP-Seq_* fields will also refer to confirming the new corrected position
- LEAP-Seq_confirmed_distance – the longest distance between the proximal and distal LEAP-Seq reads for this insertion junction that mapped to the same locus. The longer the distance, the higher the confidence that this insertion location is correct and not an artifact due to a genomic DNA fragment co-inserted with the cassette.
- LEAP-Seq_percent_confirming – percent of matching reads (see next two fields for description). The higher the percentage, the higher the confidence that this insertion location is correct.
- LEAP-Seq_N_confirming – number of LEAP-Seq read pairs in which the proximal and distal reads mapped to the same locus, .
- LEAP-Seq_N_other – number of LEAP-Seq read pairs in which the proximal and distal reads mapped to different genomic regions. This can be due to a genomic DNA fragment inserted with the cassette, or to LEAP-Seq artifacts or PCR/sequencing errors.
- plate_iteration, colony_iteration – which iteration of the deconvolution process this flanking sequence was deconvolved in (the LMJ.SG mutant series had three iterations; the LMJ.RY series only had one).
- plate_errors, colony_errors – the number of differences between the expected and observed super-pool signatures during plate and colony deconvolution.
- plate_avg_rpm, colony_avg_rpm – average ChlaMmeSeq read numbers during plate and colony deconvolution, normalized to reads per million.
- gene_synonyms, defline, description, PFAM, Panther, KOG, KEGG_ec, KEGG_Orthology, Gene_Ontology_terms, best_arabidopsis_TAIR10_hit_name, best_arabidopsis_TAIR10_hit_symbol, best_arabidopsis_TAIR10_hit_defline – annotation from Phytozome.