Basic RNA-seq analysis results structure

At the top-level, the main results folder should contain the following files and sub-folders:

- ***multiqc_report.html** - One or more web pages containing sequence-level and analysis pipeline quality control data.
- **differential_expression** - Contains differential expression results at the level of gene, transcript and (optionally) alternate splicing event.
- **graphs** - PCA and heatmap plots for each differential expression comparisons, plus overall PCA and heatmap plots for all samples.
- (plus **sessionInfo.txt** - reproducibility information, for our records).

Graphs

For each main differential expression comparison made there are two plots describing the samples included in the comparison. The PCA plot graphs samples along the two axes of gene expression by which samples differ by the greatest amount (thus if there are large differences in expression between conditions, you would expect samples from each condition to cluster together - however, a lack of such clustering does not necessarily mean that there will be no differentially expressed genes). The heatmaps show inter-sample "distance" - where each sample is considered as a point in an n-dimensional space (n=number of genes, sample location in the space determine by read count for each gene), and the distance is then just the Euclidean distance between points in that space; again, when there are large differences in expression between conditions, you would expect clustering of samples from each condition.

There are also PCA and heatmap plots showing all samples.

n.b. these plots are mainly used to spot gross QC problems and sample mix-ups, but frequently nothing conclusive can be drawn from them.

Differential gene expression

Differential gene expression results are in the sub-folder **differential_expression/de_gene**.

- ***deseq2_results_fpkm*.csv** - One or more CSV files (these can be opened directly in Excel) containing results of differential gene expression analysis.
- ***de_summary*.csv** - One or more CSV files containing summaries of the numbers of genes detected as differentially expressed (at a particular false discovery rate - unless otherwise requested, this will normally be 5%).

Each of the one or more main differential gene expression results files (there may be more than one if there are results for multiple species, or a single file was too unwieldy) contains the following columns:

- **gene**: Ensembl gene ID.
- **gene_name**: Ensemble gene name.
- **chromosome**: Chromosome on which the gene lies.
- **description**: Ensembl gene description.
- **entrez_id**: NCBI Gene ID (this is mainly for our internal use).
- **gene_type**: e.g. protein coding, pseudogene, LINC RNA etc. etc.
- **gene_length**: number of bases contained in the union of all transcripts for the gene.
- **max_transcript_length**: maximum length of any one isoform of the gene - this is used to calculate FPKMs.
- ***_fpkm**: Per-sample gene abundances measured in **f**ragments **p**er **k**ilobase per **m**illion mapped reads (n.b. if Sargasso has been used, there may be separate pre- and post-Sargasso FPKM columns for single-species samples).
- ***_fpkm_avg**: Per-condition FPKMs averaged over the samples in that condition.

and for each differential expression comparison:

- **<comparison_name>_l2fc**: log2 fold change in gene expression between conditions.
- **<comparison_name>_pval**: raw p-value for statistical significance of differential expression.
- **<comparison_name>_padj**: adjusted p-value (i.e. false discovery rate) after correcting for multiple testing.
- **<comparison_name>_raw_l2fc**: raw log2 fold change in gene expression between conditions, calculated directly from normalised read counts (this may be different to the fold change calculated by DESeq2, which does some further processing beyond the raw value).

n.b.:

- In the first instance, the `l2fc` and `padj` columns are the ones to look at.
- In a comparison named "B_vs_A", a positive log2 fold change means that gene expression is higher in the "comparison" condition B than in the "base" condition A, and vice-versa for a negative fold change. If there is any confusion about which are the "comparison" and "base" conditions (e.g. if we have named the comparison incorrectly) then consulting the appropriate entry in the ***de_summary*.csv** file will indicate exactly which conditions were used as comparison and base.
- p-values and fold changes may be missing for some genes. If the raw p-value is missing, this means that the expression of the gene across all samples is negligible, and the gene was filtered before differential expression was performed (in this case, the fold change and adjusted p-value columns will be empty too). If the raw p-value is present, but the adjusted p-value is missing, this means that the gene was filtered by DESeq2 prior to multiple testing correction (DESeq2 tries to reduce the burden of this correction by excluding genes from its final calculation that it believes, due to their expression being below a comparison-specific expression threshold, are unlikely to be discovered as significantly differentially expressed).
- The raw log2 fold change column may contain the values `Inf` or `-Inf` – these indicate that expression in one of the experimental conditions being compared was zero in all samples, and hence a finite raw fold change could not be calculated.

Gene Ontology enrichment analysis

In the sub-folder **differential_expression/go**, there are Gene Ontology (GO) enrichment analyses for each differential expression comparison (these may be in further species-specific sub-folders).

For each differential expression comparison we take (i) all genes called differentially expressed at false discovery rate < 0.05, (ii) just the up-regulated genes and (iii) just the down-regulated genes. Then for each of the GO categories of "biological process", "cellular compartment" and "molecular function", we use an algorithm to test whether the differentially expressed genes are enriched in genes annotated with particular GO terms, as compared to the background set of all genes expressed in this data. These results can start to give some clue as to the particular biological processes that are being affected, without having to just scan through huge lists of gene expression data.

n.b. some of these GO analysis files might not be present, if there were no, or only a few, differentially expressed genes for a particular differential expression comparison.

Each GO enrichment results CSV file contains the following columns:

- **GO.ID**: Gene Ontology term ID.
- **Term**: Gene Ontology term description.
- **Annotated**: Number of expressed genes annotated with this term.
- **Significant**: Number of differentially expressed genes annotated with this term.
- **Expected**: The number of genes you would expect to be annotated with this term in a random set of genes of the same sizes as the differentially expressed set.
- **weight_fisher**: A p-value for enrichment of this Gene Ontology term in the differentially expressed set.
- **Genes**: The differentially expressed genes annotated with this Gene Ontology term.

Note that the p-values here are *not* corrected for multiple testing (it's not straightforward to do this, since the statistical tests for different Gene Ontology terms are not independent). A common convention, however, is to consider GO terms with p-value < 0.01 as potentially interesting.

Gene set enrichment analysis

While GO analyses are useful, they can suffer from our having imposed an arbitrary significance cutoff for the sets of genes considered; whereas sub-threshold, yet coherent, shifts in the expression of groups of genes between conditions can also be biologically meaningful. These coherent shifts in expression can be investigated using "gene set enrichment analysis"; results are found in the sub-folder **differential_expression/gsa** (and may be in further species-specific sub-folders).

The analysis method used here is called "Camera" (Wu & Smyth, "Camera: a competitive gene set test accounting for inter-gene correlation", Nucleic Acids Research (2012)). This implements a "competitive gene set test" - that is, for a particular set of genes of interest, it takes the expression values of all genes and tests whether the fold change in expression between experimental conditions for the genes in the set is different (as a whole) to the genes not in the set. (It also takes into account that fold changes of different

genes are not necessarily independent - e.g. in cases where a bunch of genes in a pathway are all coherently differentially regulated, so the p-value obtained should be more robust than some other competitive gene set test methods that are out there.)

The gene sets we use are divided into three categories, and consist of gene sets that have been compiled by the Broad Institute:

- **CURATED**: gene sets curated from various sources such as online pathway databases, the biomedical literature, and knowledge of domain experts.
- **MOTIF**: gene sets representing potential targets of regulation by transcription factors or microRNAs.
- **GO**: Gene sets that contain genes annotated by the same GO term.

The results for each differential expression comparison are contained in a sub-folder under the `gsa` folder. Each sub-folder contains:

1) up to three CSV files, i.e. *CURATED_enriched_sets.csv*, *MOTIF_enriched_sets.csv* and *GO_enriched_sets.csv*. These detail, for the particular comparison of experimental conditions, those genes sets in each category which were found by Camera to be significantly differentially expressed, as a whole, when compared to all other genes (with a False Discovery Rate cut-off of **10%**). If there were no significant differentially expressed gene sets for a category, then the file for that category will be missing.

The columns in these CSV files are:

- **GeneSet**: Name of the set of interesting genes.
- **NGenes**: Number of genes in the set.
- **Direction**: Whether expression of the genes in this set is shifted up or down in this particular comparison of experimental conditions.
- **PValue**: Raw p-value for the significance of this shift.
- **FDR**: p-value corrected for multiple testing (i.e. a false discovery rate). Only gene sets with FDR < 0.1 are included (hence all entries in these spreadsheets can be considered significant).

2) Up to three matching CSV files, i.e. *CURATED_genes_in_sets.csv*, *MOTIF_genes_in_sets.csv* and *GO_genes_in_sets.csv*, corresponding to those gene categories for which we got significant results. These files can be used to study the behaviour of the actual genes in those significant gene sets.

The columns in these CSV files are:

- **gene**: Ensembl gene ID.
- **gene_name** : Ensemble gene name.
- **entrez_id**: NCBI Gene ID (this is mainly for our internal use).
- **<comparison>.l2fc** , **<comparison>.pval**, **<comparison>.padj**, **<comparison>.raw_l2fc**: Original differential gene expression results for the gene, for the appropriate differential expression comparison.
- **<SIGNIFICANT_GENE_SET_1>, <SIGNIFICANT_GENE_SET_2>, etc**: These columns contain a "T" if and only if the gene in this row is contained in the particular significant gene set.

Thus a good way to examine the behaviour of genes in a particular significant gene set is to order the spreadsheet first by the gene set column (e.g. <SIGNIFICANT_GENE_SET_1>) and then by the adjusted p-value for the differential expression comparison.

n.b. for more information about the provenance of a particular gene set, see the MSigDb database provided by the Broad Institute (free to access, but registration required I think).

Reactome

In the sub-folder **differential_expression/reactome**, there are enrichment analyses for pathways defined in the REACTOME pathway database.

These are very similar analyses to those performed for GO annotations. Again we take (i) all genes called differentially expressed at false discovery rate < 0.05, (ii) just the up-regulated genes and (iii) just the down-regulated genes. But now we test whether the differentially expressed genes are enriched in genes annotated as belonging to particular pathways defined in REACTOME.

n.b. some REACTOME analysis files might not be present, if there were no, or only a few, differentially expressed genes for a particular differential expression comparison.

Differential transcript expression

Differential transcript expression analysis is optional - let us know if you would like it to be performed.

Differential transcript expression results are in the sub-folder **differential_expression/de_tx**.

- **deseq2_results_tpm_*_tx_transcript.csv** - One or more CSV files containing results of differential transcript expression analysis.
- **de_summary_mouse_tx_transcript_salmon.csv** - A CSV file containing a summary of the numbers of transcripts detected as differentially expressed (at a false discovery rate of 5%).

Each of the one or more main differential transcript expression results files (there may be more than one if there are results for multiple species, or a single file was too unwieldy) contains the following columns:

- **transcript**: Ensembl transcript ID.
- **transcript_length**: Length of transcript in bases.
- **gene**: Ensembl gene ID.
- **number*of*transcript**: Number of isoforms defined for this gene in Ensembl.
- **gene_name**: Ensemble gene name.
- **chromosome**: Chromosome on which the gene lies.
- **description**: Ensembl gene description.
- **entrez_id**: NCBI Gene ID (this is mainly for our internal use).
- **gene_type**: e.g. protein coding, pseudogene, LINC RNA etc. etc.
- ***_tpm**: Per-sample transcript abundances measured in **t**ranscripts **p**er **m**illion (n.b. if Sargasso has

been used, there may be separate pre- and post-Sargasso TPM columns for single-species samples).
- ***_avg_tpm**: Per-condition TPMs averaged over the samples in that condition.

and for each differential expression comparison:

- **<comparison_name>_l2fc**: log2 fold change in transcript expression between conditions.
- **<comparison_name>_pval**: raw p-value for statistical significance of differential expression.
- **<comparison_name>_padj**: adjusted p-value (i.e. false discovery rate) after correcting for multiple testing.
- **<comparison_name>_raw_l2fc**: raw log2 fold change in transcript expression between conditions, calculated directly from normalised read counts (this may be different to the fold change calculated by DESeq2, which does some further processing beyond the raw value).

rMATS

Differential splicing analysis is optional - let us know if you would like it to be performed.

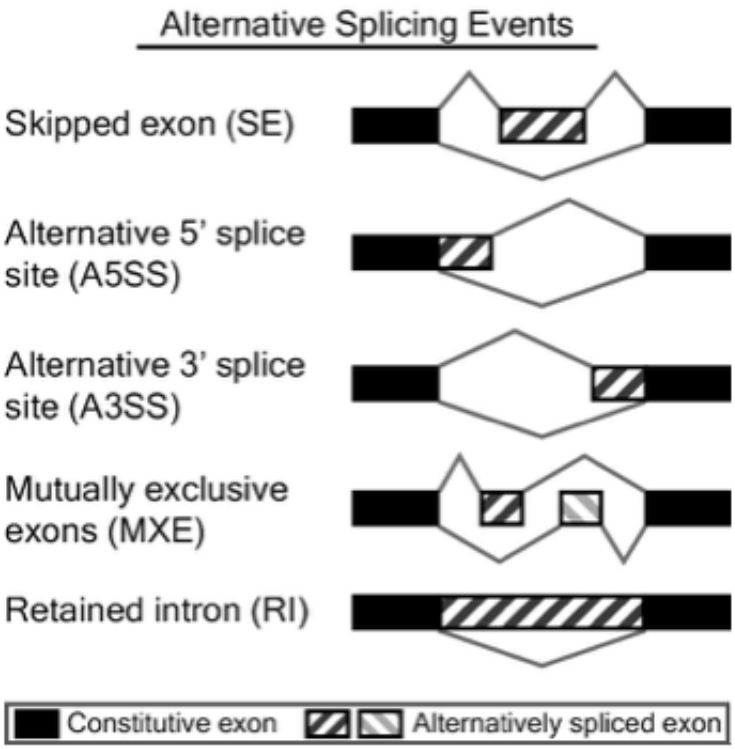
rmats (paper) is a tool to detect differential alternative splicing events from RNA-Seq data.

The statistical model of MATS calculates the p-value and false discovery rate that the difference in the "isoform ratio" of a gene between two experimental conditions exceeds a given user-defined threshold. We run rMATS for each comparison between conditions for which we produce differential gene expression results (e.g. mutant vs control). rMATS takes the samples from the two conditions in the comparison and works out the statistically significant (FDR < 0.05) alternative splicing events.

In the rMATS results folder (**differential_expression/de_rmats**), you will find one or more *AS_summary_{species}.csv* files, which contain brief summaries of all the comparisons made. Each comparison has 5 rows of summary data, corresponding to the 5 different types of splicing events supported by rMATS:

- skipped exons (SE)
- alternative 5′ splice sites (A5SS)
- alternative 3′ splice sites (A5SS)
- mutually exclusive exons (MXE)
- retained introns (RI)

Most of the columns are fairly self-explanatory. The **Up_regulated**, **Down_regulated** and **D.E.total** colums show the number of significant up- and down- regulated AS events, plus the total (at the false discovery rate specified by the column **p.adj.cutoff**). The **Up_regulated_gene**, **Down_regulated_gene** and **D.E.total_gene** columns show the corresponding number of genes in which the above events are found (i.e. there may be more than one alternative splicing event per-gene). Note that each AS "event" comprises constitutive exon sequence (i.e. isoform sequence which is present whether the splicing event has taken place or not) and alternatively spliced exon sequence (i.e. isoform sequence which is present when the splicing event has taken place):



In the rMATS result folder you will also find a sub-folder for each comparison (these may be in further species-specific sub-folders if data from multiple species are being sequenced), which in turn holds CSV files containing the details of the results for each of the five supported types of AS events. The columns in these files are:

- **gene**, **gene_name**, **chromosome**, **description**, **entrez_id**, **gene_type**, **gene_length**, **max_transcript_length**: similar columns to those that appear in the differential gene expression results files
- **avg_fpkm**: Average gene FPKM measured across all samples involved in this comparison (may be useful for filtering rMATS results)
- **ID**: rMATS defines a database of "known" alternative splicing events for which it calculates p-values - this ID identifies the splicing event in the rMATS database.
- **strand**: the strand of the splicing event (+ or -)
- **exonStart_0base**, **exonEnd** etc., or equivalent: there follow a number of columns specific to the particular type of alternative splicing event which define its genomic location; for example, for skipped exons, these define the locations of the upstream, downstream, and (potentially) skipped exon.
- **IJC_SAMPLE_1**: Numbers correspond to the samples listed for the "base" condition in the rMATS summary file; each is the number of reads in that sample that support the alternatively-spliced isoform sequence.
- **SJC_SAMPLE_1**: Numbers correspond to the samples listed for the "base" condition in the rMATS summary file; each is the number of reads in that sample that support the non-alternatively-spliced isoform sequence.
- **IJC_SAMPLE_2**: Numbers correspond to the samples listed for the "comparison" condition in the rMATS summary file; each is the number of reads in that sample that support the alternatively-spliced isoform sequence.

- **SJC_SAMPLE_2**: Numbers correspond to the samples listed for the "comparison" condition in the rMATS summary file; each is the number of reads in that sample that support the non-alternatively-spliced isoform sequence.
- **IncLevel1**: Numbers correspond to the samples listed for the "base" condition in the rMATS summary file; each is an "inclusion level" (this nomenclature only really makes sense for the skipped exon type of AS event, but is used for all types) - that is a measure of the frequency with which the alternative splicing event takes place in isoforms which contain the constitutive isoform sequence. These values are calculated from the read counts in fields **IJC_SAMPLE_1** and **SJC_SAMPLE_1**, but include a correction for the length of the respective constitutive and AS sequences. A value of 0 means the alternatively-spliced sequence is never used, while 1 means the AS sequence is always used. For a good explanation, see here.
- **IncLevel2**: Numbers correspond to the samples listed for the "comparison" condition in the rMATS summary file; these are inclusion levels for the comparison condition samples.
- **IncLevelDifference**: Equals average(**IncLevel1**) - average(**IncLevel2**). That is, a positive value means more use of the alternative-spliced sequence in the base condition, while a negative value means more use of the alternatively-spliced sequence in the comparison condition.
- **PValue**: A likelihood-ratio test is used to calculate a p-value that **IncLevelDifference** is greater than some user-specified value; we have selected the value 0.1, with the belief that inclusion level differences smaller than this are unlikely to be of great biological significance.
- **FDR**: A false discovery rate calculated from the p-value.

By default, rMATS outputs *a lot* of data, and we have observed that it has a tendency to assign very low p-values to AS events supported by very few reads. To attempt to remove some of this noise, we filter out all events supported by an average read count (across the columns **IJC_SAMPLE_1**, **SJC_SAMPLE_1**, **IJC_SAMPLE_2** and **SJC_SAMPLE_2**) of less than 5.