This document contains script-specific documentation for scripts in the https://github.com/averydavisbell/wormstrainrnaiexpr git repo.
Written by Avery Davis Bell with the intention of providing useful context for any interested in these analyses, especially anyone interested in running the scripts and understanding input and output files.

getstrainspectranscriptome.nf

- Software that must be on path:
  - Gffread
  - Bioawk
  - seqkit
- Parameters [subset from earlier workflow]

Category	Flag for script (in script as params.<this)	Default value (if highlighted, need to provide)	Description
General input	--strainlist	""	Path to one-column file with strains to process (one per line). These must be column headers in VCFs and will be used for output labelling as well.
General output	--outputdir	""	Parent output directory. Will be created if doesn't exist.
General input	--snpvcf	WI.20210121.hard-filter.isotype.snpsonly.vcf.gz	VCF containing all SNPs (only) for all strains of interest vs. reference genome of interest
General input	--indelvcf	WI.20210121.hard-filter.isotype.indelsonly.vcf.gz	VCF containing all INDELs (only) for all strains of interest vs. reference genome of interest
General input	--reffasta	c_elegans.PRJNA13758.WS276.genomic.fa	FASTA reference genome - for strain-specific genome creation; what variants were called against. Should be **unzipped**
General input	--refgtf	c_elegans.PRJNA13758.WS276.canonical_geneset_nounderscoretranscriptids.gtf	GTF file for reference genome. **Transcripts must NOT have underscores in their name if salmon is to be run through EMASE. Modify transcript names to exclude them if necessary.**
General input	--idx	salmon	Which index to build? Possible values: 'bowtie2', 'salmon', 'all' (builds both), 'none' (builds none)
General input	--salmdecoy	no	'yes' or 'no' - add REFERENCE *genome* sequence as decoy sequence for salmon

			index building? Default 'yes'. *(true/false here not smooth - I think gets interpreted as logical, which I didn't want to figure out how to deal with)*
organizational	--g2gtoolsconda	/.conda/envs/g2gtools	path to python 2 conda environment where g2gtools properly set up *Note: using ~ here doesn't work when running on node*
organizational	--procgtfscriptdir	.	Directory containing Python GTF worker scripts gtfnonchroverlaps.py and gtftoemasegenemapping.py
organizational	--salmonenv	/.conda/envs/salmon	path to conda environment where salmon is installed

- Processes & outputs

Name	**Description**	**Any saved outputs in [subdirectories of outdir]**
g2gvcf2chain	Chain indels onto reference	N/A [only want final genomes/transcriptomes used for bowtie & emase]
g2gpatch	Patch SNPs onto reference genome	N/A [only want final genomes/transcriptomes used for bowtie & emase]
g2gtransform	chain indels onto patched genome	N/A [only want final genomes/transcriptomes used for bowtie & emase]. *Would keep this if wanted the fasta long term*
g2gconvert	update GTF file based on new genome *After this process, strain-specific genome generation is complete.*	N/A [only want final genomes/transcriptomes used for bowtie & emase]
getexclseqs	For alt strain, splits GTF into sequences that should be excluded vs. included from making transcriptome (those that end after chromosome ends are excluded) Runs gtfnonchroverlaps.py	N/A [all intermediates]
straintranscriptome	Generate strain-specific transcriptome, to later be combined into diploid transcriptome	strain specific dir: $params.outputdir/$mystrain ${mystrain}_transcriptome_namesorted.fa - strain-specific transcriptome generated from reference & input VCFs

	A couple processes together: gffread to extract, then some formatting (bioawk, seqkit)	
straintrnslengths	Get lengths of each transcript for one strain, adding in any that were excluded	strain specific dir: $params.outputdir/$mystrain <strain>_transcriptlengths.txt - transcript lengths; key is it has 0s for any that weren't included in strain-specific transcriptome (but are in reference genome)
bowtie2idx	Builds bowtie2 single end index for diploid transcriptome. Bowtie2/2.3.5.1 is PACE version. **Only run if --idx is 'bowtie2' or 'all'**	strain specific dir: $params.outputdir/$mystrain *.bt* files - 6 total
salmondecoyprep	Creates fasta of diploid transcriptome + reference genome (as decoy), + chromosome names for use as salmon index's decoy **Only run if --idx is 'salmon' or 'all' AND --salmdecoy is 'yes'**	N/A
salmonidxdecoy	builds salmon index (using ref genome as decoy) **Only run if --idx is 'salmon' or 'all' AND --salmdecoy is 'yes'**	strain specific dir: $params.outputdir/$mystrain Directory *salmon_idx* in here has all salmon index info.
salmonidxnodecoy	builds salmon index without decoy **Only run if --idx is 'salmon' or 'all' AND --salmdecoy is 'no'**	strain specific dir: $params.outputdir/$mystrain Directory *salmon_idx* in here has all salmon index info.

strainspecsalmon.nf

- Parameters

Category	**Flag for script (in script as params.<this)**	**Default value (if highlighted, need to provide)**	**Description**
General input	--sampleinfo	""	Path to tab-delimited file containing sample information. Column names (descriptions): SampleID (sample ID as in input filenames, to be used in output filenames); RefDescrip (description of reference genome(s) to use in output file names); SalmonIndexDir (path to

			directory generated by salmon index); fldMean (mean library fragment length per sample, passed to --fldMean in salmon quant); fldSD (standard deviation library fragment length per sample, passed to --fldSD in salmon quant)
General output	--outputdir	""	Parent output directory. Will be created if doesn't exist.
General input	--fastqdir	""	Directory containing all fastq.gz files to process. One or more per sample.
Trimmomatic	--trimmodir	trimmomatic-0.39	Path to trimmomatic v0.39 directory containing jar file and adapters directory (which itself contains TruSeq3-SE.fa).
Trimmomatic	--trimmoseedmism	1	Input to trimmomatic ILLUMINACLIP. How many of 16 bp can mismatch and still be counted as match.
Trimmomatic	--trimmoadapclipthresh	12	Input to trimmomatic ILLUMINACLIP. How accurate match between adapter sequence and read must be. Each correct base adds 0.6. They recommend 7-15 (12 bases needed for 7, 25 for 15).
salmon	--slibtype	SR	salmon --libtype option matching the library being aligned here
organizational	--salmonenv	'/.conda/envs/salmon'	path to conda environment where salmon is installed (Anaconda3)

- Processes & outputs

Name	Description	Any saved outputs in [subdirectories of outdir]
mergeLaneFastqs	Merge files across lanes so that there's one fastq per sample	No
trimmoIlluminaAdapters	Use trimmomatic to trim Illumina adapters from merged fastqs	Keeping trim logs. Don't need, but if hadn't done before might be interesting. In /triminfo

salmonquant	quantify RNA-seq data with salmon	Keeping *all* salmon outputs - quite a lot: might be useful longterm In /salmonout/<sample>_<dip reference descrip> - one dir per

diffexp_lrt_straintreat_salmon_deseq2.R

- Inputs [run script with --help to reproduce]

```
-s, --sampinfo        Path to sample information file. Must include
                      column SampleID and any columns that are used
                      in modeldesign.
-b, --baseoutname     Base name for all output files [default: out]
-o, --outdir          Outer output directory. Sub-directories will
                      be created internally.
-e, --exampquantsf    example filepath to salmon quant.sf (or
                      quant.sf.gz) RNA quantifiaction file for one
                      sample. Transcripts in name-sorted order.
                      Where each Sample ID goes, needs to have
                      _sampid_ (e.g.
                      path/to/file/_sampid__genecounts.txt.gz).  For
                      any other differences in filepath, include *
                      for interpolation.
-t, --tx2genef        Path to file mapping transcripts to genes. Two
                      columns (transcript ID, gene ID), no header.
-r, --refcategoryinfo  Path to matrix describing the reference level
                      for each factor in the model. Columns
                      'colname', 'reflevel'. 'colname' must have one
                      entry for every column of sampinfo used in the
                      modeldesign.
-m, --modeldesign     Quote-wrapped model formula for DESeq2, e.g. ~
                      Batch + Strain + Treatment + Strain:Treatment.
                      Must include an interaction term for the
                      analyses performed in this script.  Formula
                      MUST include spaces between all terms
                      including ~. The interaction term is used to
                      generate progressively reduced models for
                      likelihood-ratio tests (interaction dropped
                      first, then the terms making up the
                      interaction separately) and to create a
```

secondary model to get condition-specific results.

-g, --genegff Path to *genes only* gff3 file containing info on all gene_ids present in input counts file

-a, --alpha Alpha p-value threshold for FDR-like filtering. [default: 0.1]

-l, --lfcthresh Log2 fold change threshold for summarizing among-group/pairwise comparison results. Not used for LRT tests or for filtering/multiple hypothesis testing correction, just for categorizing results passing alpha threshold. Default (0.5849625) corresponds to 1.5x fold change. [default: 0.5849625]

- Outputs
  - *Notes*
    - DOES re-generate some that differentialexpr_straintreat_salmon_deseq2.R does - want this to possibly be able to be independent
    - This *should* work with any model that has main1 + main2 + main1:main2 (but untested!)
  - DESeq full datasets [probably more than is necessary especially as data saved each time!]
    - *_dds_group.Rdata. DESeq1 object containing results of DESeq2 on "group" model of data, i.e. where the variables in the interaction term are combined to get within-variable results (for example, can contrast any Strain-Treatment group with any other Strain-Treatment group). With genes with <10 reads across all samples excluded. This can be queried for DE results.
    - *_dds_LRT_interaction.RData - results from likelihood ratio test of full model vs. dropping interaction term only. E.g.: ~ Strain + Treatment + Strain:Treatment *vs.* ~ Strain + Treatment
    - *_dds_LRT_<main effect this is LRT result for e.g. Strain>.Rdata - results from likelihood ratio test of no-interaction model vs. dropping first main effect (so, first term's effects). E.g. the total effect of Strain (Strain in name of file) is for LRT ~Strain + Treatment *vs.* ~Treatment
      - ONE EACH for first and second main effects!
    - *_dds_LRT_<name of main effect>_EffectIn_<Reference level of other main effect term>Only.Rdata, e.g.: *_dds_LRT_StrainEffectInCTROnly.Rdata. Two of these, one for each main effect. These are for SUBSETS of data: only the samples in the reference level for one term, testing the other term - e.g. only CTR treatment samples testing Strain term (or only N2 samples testing Treatment term).
  - Overview/QC plots
    - Generated from top 500 most varible genes after variance stabilizing transform
    - *PCA plots (all in /pcaplots directory)*
      - One PDF per factor in --reflevels (so, should be one per element included in model). First page is PC 1 vs. 2, second is PC 2 vs 3, third is 3 vs. 4.

        - *_pcaplot_<name of column/element>.pdf
    - *_pcaplot_<term 1 from interaction>_and_<term2 from interaction>.pdf - points are colored by the first part of the interaction term (e.g. strain) and shaped by the second part of this term (e.g. treatment)
- *Euclidean distance heat map plot*
    - /heatplots/*_eucdistvstheatmap.pdf - heatmap showing Euclidean distances between samples' variance stabilizing transformed gene expression (all expressed genes included)

◦ DE summaries

- In /diffexpgenes/ directory
- *_degenes_LRTs_numsummary.txt - summary of number of DE genes from the likelihood ratio tests of interaction and main effects. Columns:
    - description, interaction or main term that these results are for (i.e., they're from DROPPING this term and comparing to either full model, for interaction, or model with both terms and no interaction for the main effects)
    - nSig, number of genes with genome-wide adjusted p-value under input --alpha threshold
    - nTotal, total # genes in input (same for any test - number not filtered before running DESeq2)
    - nNonNA, total # genes with adjusted p-values calculated (those with too few reads for this test and outliers are excluded) (can vary across tests)
    - pSig_ofAll, proportion of all genes in input that are significant
    - pSig_ofNonNA, proportion of genes with adjusted p-values calculated that are significant
- *_degenes_pairwise_numsummary.txt - summary of number of DE genes from *all* pairwise comparisons between main1s [eg strain] within a main2 [eg treatment], between main2s [eg treatment] within main1 [eg strain]. Columns:
  description, description of comparison described here (e.g.: <Strain1> vs <Strain2> in <Treatment>)
  nSig_p, # genes significant at padj<myalpha
  nSig_pLogFC, # genes significant at padj<myalpha AND abs(log 2 fold change) exceeds --lfcthresh
  nUp_p, # upregulated genes (p value threshold)
  nDown_p, # downregulated genes (p value threshold)
  nUp_pLogFC, # upregulated genes, p-value and log2FC threshold
  nDown_pLogFC, # downregulated genes, p-value and log2FC threshold
  nTotal,total # genes in input (same for any test)
  nNonNA, total # genes with adjusted p-values calculated (those with too few reads for this test, outliers are excluded) (can vary across tests)
  pSig_p_ofAll, proportion of all genes significant with just p-value threshold
  pSig_p_ofNonNA, proportion of non-NA genes significant with just p-value threshold
  pSig_pLogFC_ofAll, proportion of all genes significant & passing log2FC threshold
  pSig_pLogFC_ofNonNA, proportion of non-NA genes significant & passing log2FC threshold
  pUp_p_ofAll, proportion of all genes that are upregulated (p value threshold)
  pDown_p_ofAll, proportion of all genes that are downregulated (p value threshold)

pUp_pLogFC_ofAll, proportion of all genes that are upregulated (p value & log2FC thresholds)
pDown_pLogFC_ofAll, proportion of all genes that are downregulated (p value & log2FC thresholds)

- *Overlaps of DE genes across different categories*
  - All in directory /diffexpgenes/overlaps
  - Subdirectory for main effect 1 (e.g. strain) and main effect 2 (e.g. treatment) - all below generated for both, but only showing once. Using 'Strain' and 'Treatment' as examples, stereotyped names are generated
  - Venn diagrams: one for every among-main effect (here, Treatment) comparison that's done. For 3 conditions, 3 comparisons; for 5, there are 10.
    - *_<treatment 1>_vs_<treatment 2>_Deoverlapmultiple<Strain>.pdf
  - /*Treatment*/*_numoverlapsacross*Treatment*_comparisons.txt - numerical summary of how hits within one comparison overlap with others. One row per main effect category (e.g. Strain), comparison (e.g. across treatments), then direction (any DE, upregulated, downregulated) combination. Columns:
    - test, what is comparison (numerator on top) e.g. <treatment 1> vs. <treatment 2>
    - direction, differentially expressed, upregulated, or downregulated - what DE gene numbers included here
    - categ, category this row describes (e.g. the specific strain)
    - n, number of hits/elements in this category - e.g., number of DE genes between treatments 1 and 2 in a given strain
    - n.unique, # hits/elements ONLY in this category (i.e. not DE in any other strain)
    - n.shared, # hits/elements in this category and at least one more
    - n.shared.1other, # hits/elements in this category and only one more
    - n.shared.multiple, # hits/elements in this category and more than one more
    - p.unique, proportion of hits that are unique (n is denominator for all p-columns) (i.e. proportion of DE genes in this strain that aren't called DE in any other strain)
    - p.shared, proportion of hits that are shared
    - p.shared.1other, proportion of hits shared by only one other category
    - p.shared.multiple, proportion of hits shared my more than one other category

DE gene lists (saved so they can be queried without loading data; uploaded for GO analysis or the like; etc)

- LRT tests
  - All in directory /diffexpgenes/lrt_sigde/
  - All have genome-wide adjusted p < p$alpha
  - File naming - here, Strain would be whatever first main term in model is; Treatment whatever second main term in model is
    - *_interactionStrainTreatment_sigLRTDEgenes.txt.gz - results for comparing full model to model without interaction term. [*From DESeq2 analysis saved as *_dds_LRT_interaction.RData]*
    - *_Strain_allsamplesfullmodel_sigLRTDEgenes.txt.gz - results for comparing input model without interaction term to model with just main term 2 (effects are of main term 1); for all samples [*From DESeq2 analysis saved as *_dds_LRT_<main effect this is LRT result for e.g. Strain>.Rdata]*
    - *_Treatment_allsamplesfullmodel_sigLRTDEgenes.txt.gz - results for comparing input model without interaction term to model with just main

term 1 (effects are of main term 2); for all samples [*From DESeq2 analysis saved as *_dds_LRT_<main effect this is LRT result for e.g. Strain>.Rdata]*
- *_TreatmentEffectIn<Ref level of Strain>Only_sigLRTDEGenes.txt.gz - results for main term 2 for only reference-level of main term 1(RESTRICTED) samples. E.g. p-value for comparing ~Treatment to ~1 in only samples from reference strain. [*From DESeq2 analysis saved as_dds_LRT_<name of main effect>_EffectIn_<Reference level of other main effect term>Only.Rdata]*
- *_StrainEffectIn<Ref level of Treatment>Only_sigLRTDEGenes.txt.gz - results for main term 1 for only reference-level of main term 2 (RESTRICTED) samples. E.g. p-value for comparing ~Strain to ~1 in only control-treated samples [*From DESeq2 analysis saved as_dds_LRT_<name of main effect>_EffectIn_<Reference level of other main effect term>Only.Rdata]*

  - Columns:
    - gene_id, gene ID
    - gene_name, gene locus name
    - biotype, gene biotype
    - baseMean, base mean expression of gene (DESeq2-calculated)
    - stat, likelihood ratio test test statistic (DESeq2-calculated)
    - pvalue, unadjusted p-value (DESeq2-calculated)
    - padj, genome-wide adjusted p-value (NA for genes with too few reads, outliers) (DESeq2-calculated)

- Pairwise comparisons among groups
  - All in directory /diffexpgenes/lrt_sigde/
  - All have genome-wide adjusted p <input --alpha **and** absolute value of log2 FC > --lfcthresh
  - All extracted from the **_dds_group.Rdata* DESeq2 analysis via contrasts
  - Saved for all pairwise comparisons, which can be a lot! Presumably most (all?) of these won't be looked at. Filenaming:
    - <Category compared e.g. Strain>_<member 1 of that category in numerator eg first strain>_<member 2 of that category in numerator eg 2nd strain>
  - Columns:
    - gene_id, gene ID (from rownames of dds_lrt)
    - gene_name, gene locus name
    - biotype, gene biotype
    - baseMean, base mean expression of gene (DESeq2-calculated)
    - log2FoldChange, **ashr-shrunken** log2FoldChange (DESeq2-calculated)
    - lfcSE, standard error on log2FoldChange (DESeq2-calculated)
    - pvalue, unadjusted p-value (DESeq2-calculated)
    - padj, genome-wide adjusted p value (DESeq2-calculated)

mosdepthmergedexons.nf

needs mosdepth, gtftools on path

- Parameters

Category	Flag for script (in script as params.<this >)	Default value (if highlighted, need to provide)	Description
General input	--gtf	""	path to GTF containing genes for which to determine coverage from all BAMs. Columns as ws276 GTF from Wormbase.
General input	--sampleinfo	""	Path to sample information file. Columns SampleID (name of sample for output), bam (path to BAM file to process for this sample), bai (path to BAM .bai index file for this sample)
General input	--outdir	"out"	Path to output directory
General input	--outname	"out"	Prefix for output files that contain all samples' mosdepth information
GTFtools	--refname	"ref"	Prefix for output file containing merged exons - i.e. reference genome name
GTFtools	--chrs	I,II,III,IV,V,X,MtDNA	Chromosomes to process exons/genes for - need to match the GTF. Default is for *C. elegans* ws276
GTFtools	--gtftoolsdir	GTFtools_0.8.5	Directory containing GTF tools python script gtftools.py
mosdepth	--flag	1796	--flag (SAM flag bits to exclude) argument for mosdepth **Default is mosdepth default; may very well want to change!**
mosdepth	--mapq	0	-Q, mapq threshold argument for mosdepth, threshold below which read will be excluded **Default is mosdepth default; may very well want to change!**
Summary R script	--rscriptdir	../..	Directory containing exploregenecoverage_fromexons.R
Summary R script	--gff	""	Path to *genes only* gff3 file containing info on all genes from the GTF
Summary R script (subset)	--gsubset	''	OPTIONAL Path to no-header list of genes to run summary R script for - SUBSET of all genes. *It will also be run for all genes.*
Summary R script (subset)	--gsubsetname	""	OPTIONAL name of gene subset for output filenaming. Provide if provide --gsubset

- Processes & outputs

Name	Description	Any saved outputs
gtf2mergedexonbed	Get merged exons bed file from GTF using GTFtools	<refname>.mergedexons.bed.gz - bed file of merged exons made from GTF *[probably DON'T need to save this given same information is present in the coverage output beds, but keeping for now to be extra safe]*
mosdepth	Run mosdepth for genes. Also unzips bed output for downstream ease.	NA - combining together before saving
combinedpbeds	Combines *.regions.bed.gz mosdepth outputs into one file with all samples	<outname>.mergedexons.bed.gz - key file! One row per input gene; columns with gene info followed by one column per sample/strain containing mean coverage in that region (over that gene) for that strain
combinedpsumms	Combines *.mosdepth.summary.txt mosdepth outputs into one file with all samples	<outname>.mosdepth.summary.txt. 2 rows per chromosome and total per sample. Columns: SampleID - which sample. Repeated for all rows with this sample's data Chrom - chromosome ID, or chromosome ID _region: data for entire chromosome or for all the provided regions (genes) on a given chromosome Length - length of chromosome/sum of length of regions Bases - read bases total aligned here at thresholds above Mean - mean coverage (# reads covering) across specified region Min - min coverage Max - max coverage
comboexonsexplore	Runs exploregenecoverage_fromexons.R for all genes *Added second*	<outname>_genecoveragefromexons_raw.txt.gz and <outname>_genecoveragefromexons_mednorm.txt.gz - per GENE coverages computed from merged exon bed, raw and normalized to across-gene median In /plots subdirectory, lots of plots of coverage - see documentation for this script for full breakdown

exploregenecoverage_fromexons.R

- Inputs

-e, --exoncov Path to multi-sample/strain coverage-per-merged exon file (e.g. output of combinedpbeds process, mosdepthmergedexons.nf workflow). Columns chr, start, end, name, <1 per sample containing mean coverage over exon for that sample>. **One row per MERGED EXON, so multiple rows per name (name = gene ID)

-c, --covsumm Path to multi-sample/strain mosdepth summary file (e.g. output of combinedpsumms process, mosdepthmergedexons.nf workflow). Columns SampleID, chrom (chrom_region is for just merged exons on that chromosome; total is genome-wide), length, bases, mean, min, max

-g, --genegff Path to *genes only* gff3 file containing info on all genes in name column of --genecov input

-o, --outstem Output filestem. Include preceding path if don't want outputs in current directory. [default: out]

--genelist OPTIONAL path to (no-header) list of gene IDs (as in name column of --exoncov) to restrict analyses to (e.g. expressed genes, protein-coding genes, etc)

- Outputs
  - Coverage data
    - *_genecoveragefromexons_raw.txt.gz - coverage per gene as computed from merged exons - (coverage per exon * length of exon) / (total length of exons in gene). Columns:
      - gene_id, locus, sequence_name, biotype, chr, start, end, strand - information from GFF

      nMergedExons, # of merged exons in this gene (used to determine strain coverage)
      length.exons, merged exonic length of the gene (just for interest)
      <sample/strain IDs> - one column per input sample column. Has the gene coverage for this sample (computed as described)
- *_genecoveragefromexons_mednorm.txt.gz - Raw coverages above normalized to **MEDIAN** coverage across genes (just divided by medians). *Chose median because there are some serious outliers*. For all genes included here - subset to gene list if provided, not if not. *Might want to re-compute medians/median corrections in downstream analyses from raw!* Columns as above, but now values in sample coverage are median normalized
  - Plots (copying documentation from previous)
    - *Per-strain coverage plotted*
      - All of these made for RAW coverage (just for interest) and for coverage normalized to mean coverage across included genes. Filestems:

        - <outstem>_**raw**coverage
        - <outstem>_medngenenormcoverage
    - *_genecovviolinplots.pdf. Violin plots. First has all data, raw axes. Next is log10 axes - **careful, excludes 0s.** Then, axes cut to mean coverage (across all strains)*5, then to just mean coverage. *These latter only make sense to look at for mean-normalized data since all strains will have mean 1 there.*
    - *_genecovhists.pdf. Histograms, faceted by strain. Axes and caveats as with violin plots. Axis expands to all genes even when x axis value cut off.
    - *_genecovrankedlines.pdf. Line plots showing cumulative coverage, sort of: x axis is gene index, y axis is coverage. To compare shape. Same set of axes as above but no log10.
- *strainVstrainplots.pdf - strain vs. strain matrix plots; each gene is a point (unless I break R). First page has all included genes; second restricts axes to mean coverage (across all strains)*5, third to mean coverage (x and y axis restricted). **This definitely won't scale well.**
    - One each for _**raw**coverage [axis restriction doesn't make as much sense - will be diff # genes per plot], _medgenenormcoverage
- Number & proportion of genes at different coverage thresholds
    - Each coverage threshold is *that proportion of mean.* So, basically, the number <= 0.05 is the number where <raw expr/mean across genes raw expr> <= 0.05. Not the same as quantiles.
    - _ngenesunderpropofmeancoverage.txt - NUMBER genes in each strain
    - _propgenesunderpropofmeancoverage.txt - PROPORTION of genes in each strain
    - Format/columns:
        - Strain (obv)
        - nUnderPropMeanCov<num> or propUnderPropMeanCov<num>
            - <nums> are 0-0.5 by 0.05 intervals
            - Number is genes with **less than or equal to** this proportion of the mean coverage. So 0 is true 0-coverage genes.

de_dnacov_overlap.R

Inputs [run script with --help to reproduce]

```
-g, --genelist    Path to no-header list of all genes that were
                  considered to derive hits (e.g. included in
                  differential expression analysis thanks to having
                  enough coverage).  All of these should have also
                  been included in coverage analysis (not checked!).
                  Used as background/total set.
--hitgenes        Path to file containing genes that are hits to
                  overlap with no coverage (e.g. differentially
                  expressed genes).  Must have column gene_id (format
                  as in other inputs); may have other columns that
                  will be included in output but not otherwise used
```

here.

-n, --nocovgenes Path to list of genes flagged as not having coverage in strains of interest - output of genelistsfromcovprop.R. Presence in ANY strain will be enough for gene to be included in 'nocov' set - want to match 'hits' input with this carefully. Columns gene_id (as in other inputs), SampleID (strains), meannormval (value from filtering). One row for each strain-low coverage gene pair.

-o, --out Base out name - absolute path (include directory if not running from within desired output directory).

- Outputs
  - Summaries of no-coverage genes overlapping with input genes - but no hit information (added 1/19/22). *Descriptive, maybe duplicative but not clear this is found elsewhere.*
    - *_nocovgenewhichstrainsummary.txt - per-strain summaries of how that strain contributes no-coverage genes to total set. Columns:
      strain, strain ID
      n, total # no-coverage genes that were no-coverage in this strain
      n.unique, # no-coverage genes that were no-coverage *only* in this strain
      n.withothers, # no-coverage genes that were no-coverage in this strain and at least one other strain
      p.nocov, .nocov denotes proportion of total no-coverage gene set. Other parts of following names (and this one) as above.
      p.uniq.nocov,
      p.withothers.nocov,
      p.total, .total detotes proportion of total gene set (number in --genelist). Other parts of following names (and this one) as above.
      p.uniq.total,
      p.withothers.total
    - *_nocovgenewhichstraincombosummary.txt - per strain combination summary of how many genes no coverage in each unique strain combination. Columns:
      whichstrains, combination of strains;
      N - # nocov genes in that combination
      p.nocov - proportion of nocov genes (in nocovs) with this combination
      p.total - proportion of ntot genes with this combination
  - *_inputhitsnocovannotated.txt.gz - DE or similar hits exactly as in input, but now annotated with no-DNA-coverage information. Columns:
    - gene_id
    - nocov_nstrains, # strains this gene was called as no coverage DNA in (from input, no new calling)
    - nocov_whichstrains, which strains this gene was called as no coverage DNA in, or NA if not no coverage in any strain
    - <any others included in input>

- *_hitnocovgenenstrainsummary.txt - little table showing, *for genes that are hits & not covered in 1 or more strains given here,* how many were not covered in each strain combination observed. Sorted with strain or strain combination contributing most non-covered genes first.
- *_hitnocovgenewhichstrainsummary.txt- little table showing, *for genes that are hits ,* how many were not covered in 0-maximum number of strains
- *_hitsnocov_overlapsummary.txt- Numerical summary of overlap between DE/input hit genes and no DNA coverage genes. Lots of numbers and proportions - columns:
    - nHits, # input hit (e.g. differentially expressed) genes
    - nNoCov, # genes with no coverage in one of these strains that are in overall gene set (p$genelist)
    - nHitsNoCov, # genes that are both hits and no coverage
    - nHitsNoCov1Strain, # genes that are hits and are not covered in 1 strain
    - nHitsNoCovMultStrains, # genes that are hits and are not covered in multiple strains
    - pAllGenesNoCov, proportion of total # genes that are not covered in at least one strain
    - pHitGenesNoCov, proportion of hit (DE) genes that are not covered in at least one strain
    - pAllGenesHits, proportion of total # genes that are hits
    - pNoCovGenesHits, proportion of no-coverage genes that are hits
    - pvalueHypGeom, Hypergeometric test p-value for overlap of 'hit' and 'no coverage' enrichment
- *Summary plots*
    - *_hitsnocov_overlapunscaledvenn.pdf - Venn diagram showing overlap between all genes, hit genes, no-coverage genes. All genes necessarily contain the other two, but Venn doesn't scale its area. Does have numbers in each region of plot.
    - *_hitsnocov_overlapeulerplot.pdf - Euler diagram showing overlap between all genes, hit genes, no-coverage gene. With areas more scaled to numbers and showing that all genes subsume other sets.
    - *_propnocovgenesplot.pdf - bar plot showing proportion of all genes, hits (i.e. DE genes), non-hits (i.e. non-DE genes) that are no-coverage genes. Error bars are 95% binomial confidence intervals on proportion.

exploregenecoverage_fromexons_lowend.R

Inputs

```
-b, --baseoutname  Base name for all output files [default: out]
-o, --outdir       Outer output directory. Sub-directories will be
                   created internally.
-s, --strains      Strains to process - in other inputs. Either
                   comma-separated (no spaces) list or path to
                   no-header file with one line per strain.  Must
                   match how strains are named in input files.
                   Strains will be plotted/leveled in this order.
                   [default: N2,JU1088,EG4348,CB4856,QX1211]
-i, --isos         Isotypes of the strains provided in --strains *in
```

same order*. Either comma-separated (no spaces) list or path to no-header file with one line per strain. [default: N2,JU1088,EG4349,CB4856,QX1211]

-r, --rawcov Path to file containing at least gene_id, raw coverage for each gene of interest in each strain of interest (i.e. _genecoveragefromexons_raw.txt.gz output of exploregenecoverage_fromexons.R)

--offgenes Path to *_zeroexp_rnai_nocov_genes.txt file

-n, --nucdiv per-gene divergence metrics as output by nucdivcendr_geneswindows_allandasestrains.R

- Outputs
  - *Just DNA coverage data*
    - *_lowdnacovhists.pdf - Contains zoom in on lower tail of DNA coverage from input. Titles describe specifics.
    - *_number0coveragesummary.txt - per strain summary of how many genes have 0 raw coverage called
    - *_barplot0covlowcov.pdf - bar plot with # low coverage genes per strain, colored/split based on whether they're 0 raw coverage or > 0 raw coverage
  - *DNA coverage x off genes*
    - *_offgenesX0coveragesummary.txt - per strain summary of how many off genes have different coverage categories, including 0 raw coverage
    - *_barplotOffGenes0covlowcov.pdf - bar plot with # off genes per strain, colored/split based on whether they're zero, low, normal DNA coverage in that strain
      - in legend, "zero" coverage means raw 0x; "low" means above 0x but less than 25% median coverage; "normal" means >25% median coverage

offgenes_straintreatDE_deseq2_dnacov.R

- Inputs

-d, --dds DESeq2 strain/treatment interaction analysis object (from differentialexpr_straintreat_salmon_deseq2.R) path. Must be named dds (internally). Must provide BOTH this AND --ddsgrp.

--ddsgrp DESeq2 dds object for GROUP model object (from differentialexpr_straintreat_salmon_deseq2.R) path. Must be named dds_grp (internally). Must provide BOTH this AND --dds.

-o, --outdir Output directory. Created if it doesn't exist. [default: out]

--outstem Output filestem. [default: out]

-a, --alpha Alpha p-value threshold for FDR-like filtering

from LRT test of strain. [default: 0.1]

-l, --lfcthresh Log2 fold change threshold for summarizing RNAi/treatment results. Not used for filtering/multiple hypothesis testing correction, just for categorizing results passing alpha threshold only for RNAi. Default (0.5849625) corresponds to 1.5x fold change. [default: 0.5849625]

-n, --nocovgenes Path to list of genes flagged as not having coverage in these strains - output of genelistsfromcovprop.R. Columns gene_id (as in input dds's), SampleID (strains), meannormval (value from filtering). One row for each strain-low coverage gene pair.

- Outputs
  - zeroexponeormorestrains_Nsummary.txt, oneexponeormorestrains_Nsummary.txt, and fiveexponeormorestrains_Nsummary.txt - numerical summaries showing # genes that 1) had p < alpha in LRT test of strain as a whole and 2) have average expression in title (0, 1, 5) across ALL samples. Shows number for different strain combinations. Columns:
    -

whichunder	Comma-separated string: which strains under threshold
nstrains	# of strains this is
ngenes	# genes under threshold in this unique strain set

  - ***_zeroexp_rnai_nocov_genes.txt - **key output.** One row per 'off' gene ('off' = significant in strain LRT, 0 average normalized counts across all samples in at least one strain). Annotated with RNAi DE information, whether gene was flagged in low coverage in input file. Columns:
    -

Column Name	Description (if not obvious)
gene_id	
gene_name	
biotype	
nStrainsZeroExp	# strains with MEAN expression overall at 0
whichStrainsZeroExp	Comma-separated string specifying which strains have 0 expression
nnocov	# of strains with no coverage at this gene (as defined however input gene no-coverage list was created)

anyZeroExpIsNoCov	T, F, or NA: are ANY of the strains called zero expression also no coverage at this gene? NA if no no-coverage strains at this gene.
allZeroExpIsNoCov	T, F, or NA: are ALL of the strains called zero expression also no coverage at this gene? NA if no no-coverage strains at this gene.
<strain>.nocov	One column per strain, T or F - was this gene flagged as no coverage (in input no coverage list) for this strain?
nRNAiDE	# significant RNAi DE comparisons (one = POS1 in a strain, e.g.)
whichRNAiDE	Comma-separated string specifying which RNAi comparisons have DE
<strain>.CTR_vs_<strain>.CTR	One column per pair of strains. Value is log2 fold change between first and second *if p<alpha* in LRT test AND p<alpha in this specific pairwise test. (logFC threshold not used here!)
<strain>.<treatment>	One column per strain/treatment pair, CTR first. Value is MEAN EXPRESSION across the samples of this strain and treatment
<strain>.all	One column per strain, value is MEAN EXPRESSION across all samples of this strain (all treatments pooled)
<Strain>.<treatment>_vs_<same strain>.CTR	One column per treatment comparison (to CTR) *within strain.* Value is log2 fold change of treatment *if p<alpha & lfc > threshold* (NA'ed out if no significant DE)

- *zeroexp_rnai_nocov_genes_perstrain_nsummary.txt - summary of # off genes with different characteristics (coverage, RNAi DE, etc) on a per-strain basis. Currently one column per strain; categories/number summaries are rows *[if did for lots of strains at once, would want these to be column-wise instead]:*
    -

category	description
n	# genes 'off' in this strain
nStrainOnly	'off' only this strain
nStrainPlus	'off' this strain + others

nCovInStrain	'off' and not called no coverage in this strain
nNoCovInStrain	'off' and called no coverage in this strain
nStrainOnly_CovInStrain	off ONLY this strain, not no-cov
nStrainOnly_NoCovInStrain	off ONLY this strain, called no-cov
nStrainPlus_CovInStrain	off this and more strains, not no-cov
nStrainPlus_NoCovInStrain	off this and more strains, called no-cov
nRNAi	off in this strain AND has RNAi effect in some strain
nRNAi_CovInStrain	off + RNAi effect + not no-cov in this strain
nRNAi_NoCovInStrain	off + RNAi effect + called no-cov in this strain
nRNAi_nStrainOnly	off ONLY here + RNAi effect
nRNAi_nStrainPlus	off this and more strains + RNAi effect