Caenorhabditis tRNA allelic variation analysis (github.com/paabylab/trnavcfalleles) Code details

Avery Davis Bell
September 2025

This document describes inputs and outputs (or other relevant aspects) of all scripts in this repository that are run on their own – *i.e.,* if a script is only called via an included workflow, that script may not be fully documented here (but you can see workflow doc to figure out its use case). Scripts are grouped into their general categories, but be aware outputs of some scripts are used for multiple types of analyses.

All script paths here are internal to https://github.com/paabylab/trnavcfalleles (relative paths)

Table of Contents

Related to generating initial information about tRNA genes, alleles

generate_alleles/trnavcfvariants.nf

Workflow that generates all unique tRNA allele sequences; runs tRNAScan-SE to characterize them; collects summaries for this.

- Parameters/inputs

Category	Flag for script (in script as params.<this)	Description

Data input	--runinfo	/ tab-delimited file, one row per sample/species to run. Columns:
		// id, sample/species ID (used for output file naming)
		// trnasout, highest-confidence tRNAscan-SE .out filepath OR full - just keep track of whether you're including pseudogenes or not!
		// trnass, highest-confidence tRNAscan-SE secondary structure (usually .SS) out filepath
		// trnabed, tRNA bed file path
		// vcf, path to VCF for this species - used to find all the tRNA variants
		// vcftbi, path to VCF .tbi index
		// refstrain, ref strain to strip out before data summarization

Data output	--out	
software	--vcfconda	path to conda environment set up to run scripts that import pyvcf
software	--pyscriptdir	Directory containing python scripts run here - get_strain_variants.py etc
software	--rscriptdir	Directory containing R scripts run here (e.g. getstrainxtrnacalls.R)

- Processes

Name	Description	Any saved outputs? (all in output dir / ID of this species/sample) And relevant notes
subsetvcf	Slices vcf to only be tRNA regions (otherwise collectvarstrnas is quite slow) // AND Edits VCF so python vcf reader can work with it (unzips; removes "malformed filter lines", etc)	no (*but definitely resume this, it's big)*
collectvarstrnas	Runs get_strain_variants.py (Collects variants in tRNAs	in /data/ subdir _strain_variants.txt: variants in tRNAs

	(important preprocessing step for next scripts)) VCFs have to be unzipped here [there is built in for .gz but encoding issues seem to happen]	get_strain_variants.log: log file, *but also has record of who has missing genotypes. might be useful.*
buildaltseqs	Runs build_alt_sequences.py (Generates strain-specific tRNA sequences)	in /data/ subdir _strain_trnas_gen.txt, _strain_trnas.fa, _strain_trnas_info.txt
straintrnascanse	Run tRNAscan-SE on strain-specific FASTA	in /data/trnascan subdir: .out: main output file .SS: secondary structure .stats: about tRNAscan-SE run .isospecific.out: tRNA isotype info
missingness	Run catalogmissingness_trnavcfscripts.R to get record of where missing VCF genotypes are	in /data/ subdir: _trna_variant_info_wmissingness_pervariant.txt _trna_variant_info_wmissingness_pertrna.txt _trna_variant_info_wmissingness_perstrain.txt In /summaries/ subdir: _trna_variant_info_wmissingness_pertrna.pdf _trna_variant_info_wmissingness_perstrain.pdf
strainxtrna	Run getstrainxtrnacalls.R to get final data and summaries	in /summaries/ subdir - all generated files: _referencestrain<refstrain>_pertRNAalleles.txt (if ref strain provided, included) _strainsxtrnas_alleles_wmissing.txt.gz (***this is data though too)*** _alleleinfo_counts_wmissing.txt _strain_alleletype_counts_wmissing.txt _strain_alleletype_counts_wmissing_variablegenesonly.txt _bygeneup_tRNA_counts_wmissing.txt _bygeneup_codon_counts_wmissing.txt _bygeneup_AA_counts_wmissing.txt

generate_alleles/catalogmissingness_trnavcfscripts.R

Gets which strains have missing calls per variant, per tRNA

- Inputs (also get by running script with --help)

```
-v, --varinfo      Path to *_strain_variants.txt output of
                   get_strain_variants.py. **strains are inferred
                   from this
-t, --trnainfo     Path to *_strain_trnas_info.txt output of
                   build_alt_sequences.py
-b, --baseoutname  Base name for all output files [default: out]
-o, --outdir       Outer output directory. Sub-directories will be
                   created internally. **NB: if you provide getwd()
                   here (quote wrapped), current directory will be
                   used
```

- Outputs
  - *_trna_variant_info_wmissingness_pervariant.txt: variant information annotated with which tRNA it is in. Columns:
    -

Chr	chromosome variant & tRNA are on
Pos	variant position
Ref	variant ref allele
Alt	variant alt allele
tRNA	tRNA ID (e.g. chrI.trna1)
Start	tRNA start
End	tRNA end
Strand	tRNA strand
AA	tRNA AA
Codon	tRNA codon
nhomRef	# strains with ref allele
nhomAlt	# strains with alt allele
nMissingOrHet	# strains with missing (or het called) genotypes
missingOrHet	comma-separated list of strains with missing (or het called) genotypes. If none for this variant, entry is 'None'

  - *_trna_variant_info_wmissingness_pertrna.txt : tRNA-level summary: number and identity of strains with missing-ness variants overlapping tRNA (or not where there aren't variants/missing variants). Here, missing/het is for *any or all* of the overlapped variants

- *Use this to annotate next level summaries, probably*
- Columns:
-

tRNA	tRNA ID (e.g. chrI.trna1)
Chr	chromosome
Start	tRNA start
End	tRNA end
Strand	tRNA strand
AA	tRNA AA
Codon	tRNA codon
nVCFVars	# variants from VCF were in this tRNA
nMissingOrHet	# strains with missing (or het called) genotypes
missingOrHet	comma-separated list of strains with missing (or het called) genotypes. If none for this variant, entry is 'None'

- *_trna_variant_info_wmissingness_perstrain.txt: quick strain level summary of number of tRNAs in that strain with missing variant calls. Columns:

  # strain, strain ID

  # nvars_missing, number of variants in tRNAs that had missing (or het) calls in this strain

  # ntrnas_missingvars, number of distinct tRNAs with one or more variants with missing (or het) calls in this strain

  # ntrnas_nomissingvars, number of tRNAs with no missing/het calls - just total number input minus previous number

- *quick summary plots*
  - *_trna_variant_info_wmissingness_pertrna.pdf - histogram showing number strains with missing calls (x axis) vs number tRNAs (y axis). Each bin is a single count.
  - *_trna_variant_info_wmissingness_perstrain.pdf - scatter plot showing how *many reference* tRNAs each strain has with any missing called variants (and total # missing-called variants on y axis)

generate_alleles/getstrainxtrnacalls.R

Getting strains x tRNA genes matrix with allele ID, including if there was any missingness
NB: 'Lost' is the given classification if : as pseudo gene (low inf but tRNAscan-SE does this) or if tRNAscan-SE doesn't find the generated fasta seq

- *How allele classifications are done:*

Classification	Criteria
Lost	not found by tRNAscan-SE, or found and called a pseudo gene (or secondary or undetermined isotype)
Altered	AA and alleleCM don't match (isotype switch) [or IPD in note - these are the same]
Invariant	only one allele in population AND not altered/Lost. **Deprecated - now we have a separate column for VariableInPop T/F, can use this to interpret other classifications**
Best	isotype matches what is expected and has the best infernal score for this tRNA should best be highest or lowest? highest based on what's classified as pseudo. [only confusion was in an old plot interpretation]
Functional	isotype matches what is expected, has a functional infernal score for this tRNA that is less than the Best score (Infernal > 20 but really these should all be flagged in note if fail this)

- Inputs (also get by running script with --help)

```
-t, --trnainfo     Path to *_strain_trnas_info.txt output of
                   build_alt_sequences.py
-m, --missinfo     Path to
                   *_trna_variant_info_wmissingness_pertrna.txt
                   output of catalogmissingness_trnavcfscripts.R
--trnascanout      Path to strain-specific fasta tRNAscan-SE output
                   (.out file)
--trnafasta        Path to strain_trnas.fa from
                   build_alt_sequences.py (used to check allele
                   names)
-r, --refstrain    ID of reference strain(s) to strip out of final data; comma-separated if more than one
-b, --baseoutname  Base name for all output files [default: out]
-o, --outdir       Outer output directory. Sub-directories will be
                   created internally. **NB: if you provide getwd()
                   here (quote wrapped), current directory will be
                   used
```

- Outputs
  - *Main data*
    - If reference strain is known and provided, it is removed from final data and saved separately:
      - _referencestrain<refstrain>_pertRNAalleles.txt
      - Columns: tRNA (tRNA gene ID), <refstrain>_allele (allele ref strain has there - generally Reference one, ha!)
    - *_strainsxtrnas_alleles_wmissing.txt.gz: *gzipped* Strains x tRNAs matrix with allele identified for each strain at each tRNA; NA if strain had any missing variant calls at that tRNA
      - columns are strain then each tRNA
  - *Data summaries (excludes reference strain duplication where reference strain ID provided)*
    - _alleleinfo_counts_wmissing.txt: **allele info & counts of strains with each allele.** Columns:

tRNA	tRNA ID
allelename	name of allele with date in this row
AA	amino acid the following DNA codon codes for (from tRNAscan-SE)
Codon	specific codon (from tRNAscan-SE) [DNA: reverse comp of the RNA code]
Infernal	infernal score (from tRNAscan-SE)
AlleleCM	expected amino acid based on backbone [*best guess, haven't found clear def of this field]* (from tRNAscan-SE) yep as far as I can tell - "The sequences were then grouped according to isotype, this time primarily based on which isotype-specific covariance model yielded the highest score for each sequence (regardless of the anticodon sequence). " (2021 tRNAscan-SE paper)
IsotypeScore	(from tRNAscan-SE)
Note	(from tRNAscan-SE)
VariableInPop	T or F, does this allele vary in the population
Classification	Lost, Altered, Functional, Best - see classification table above
n.allele	# strains with this allele

n.called	# strains with any allele called here (same for all alleles within gene)
n.missing	# strains with missing genotype calls for any variants in this tRNA (same for all alleles within gene)

- Updated 5/12/25 to make sure ones with no codon called by tRNAScan-SE are classified as pseudo/lost (some with 'Undet' amino acid/not-found codon were coming through as 'Altered')

- *_strain_alleletype_counts_wmissing.txt: strainwise count summary (*less useful for initial mutational variation piece, more for possible fitness differences/mutational burdens).*
  long format data with one row per strain, amino acid combination (and all together). Columns:
  - # strain,
  - # AA, amino acid alleles/tRNAs are combined for - if 'all', that's top-level strain summary
  - # <one for each classification - Altered, lost, best, functional> number of alleles called <this classification> for this amino acid in this strain
  - # nNA, number of alleles not called due to genotype missingness

    NB: [per AA is the ones actually coded for -- not the alleleCM. could be delivering wrong ones or...?]

- *_strain_alleletype_counts_wmissing_variablegenesonly.txt: same as above [columns *may* end up with different orders though], but only genes that have more than one observed allele in population used for summary - so can look at counts of just variable ones
- Summaries of number of alleles, genes, etc for each tRNA, codon, AA. *Note for all, AA is the one actually coded for from codon - so one mutated away from its original codon is counted in its new group*
  - *_bygeneup_tRNA_counts_wmissing.txt. one row per tRNA gene. Columns:

tRNA, gene ID
AA, AA codon codes for
Codon, actual codon in tRNA
VariableInPop, is gene variable in population
nAlleles, # alleles observed in population
anyMissingCalls, T or F, any alleles have any missing genotype calls in sequence
Best, number alleles classified as best
Functional, number alleles classified as functional
Altered, number alleles classified as altered
Lost, number alleles classified as altered

  - *_bygeneup_codon_counts_wmissing.txt. one row per triplet codon - the one that's actually CALLED, not the backbone! Columns:

AA, AA codon codes for
Codon, actual codon in tRNA
nGenes, # genes that have this codon
VariableInPop, is ANY gene for this codon variable in population
nAlleles, # alleles observed in population across any gene
nGenesMissingCalls, how many genes have alleles with any missing genotype calls (./. in VCF)
Best, number alleles classified as best across all genes
Functional, number alleles classified as functional across all genes
Altered, number alleles classified as altered across all genes
Lost, number alleles classified as altered across all genes

- *_bygeneup_AA_counts_wmissing.txt. one row per amino acid - the one that's actually CALLED, not the backbone! (from the codon)

AA, AA summarized here
nCodons, # codons for this AA in this genome
nGenes, # genes that have this AA
VariableInPop, is ANY gene for this AA variable in population
nAlleles, # alleles observed in population across any gene for this AA
nGenesMissingCalls, how many genes have alleles with any missing genotype calls (./. in VCF)
Best, number alleles classified as best across all genes for this AA
Functional, number alleles classified as functional across all genes for this AA
Altered, number alleles classified as altered across all genes for this AA
Lost, number alleles classified as altered across all genes for this AA

Related to secondary structure of tRNA (variant positions relative to secondary structure, secondary structure-aware alignments, etc)

secstruct_related/get_strain_variants_relpos.py

Gets variants internal to tRNAs (like get_strain_variants.py), but also includes their relative position (e.g. base number in acceptor stem).

Notes/caveats/method

- This WILL NOT assign a variant to multiple tRNAs - if tRNAs overlap, have a problem
    - *(laziness founded in biology*)
- If tRNA secondary structure isn't well parsed, info about relative position will be NA'ed out
- If tRNA has an intron, relative position is for *non-intronic length*
- Indels sometimes cause things to break
- *** see also notes for secstruct_related/secstruct2seqpieces.py
- Logic used to assign bases to structures (combined with secondary structure paired/unpaired calls):
- What about using KNOWN lengths of tRNA structures to guide? [using this in coding]
    - Require GOOD EVIDENCE to vary from this (except in variable loop)
    - For good looking tRNA:
    -

acceptor stem	1 non paired (base 73, often) 7 paired nts [but this can vary or have mismatches]
possible linker	0-…2?
d arm (paired)	usually looks like 4 paired
d loop	8-10? *varies?*
possible linker	0? 1+?
anticodon arm (paired)	5 paired nts [usually]
anticodon loop	easy to ID - has anticodon & other unpaired nts surrounding
variable loop	3+
t arm (paired)	4?
t loop (unpaired)	however many

    - SO, I can take this as baseline, and vary for 'nice' tRNAs that clearly have longer paired regions (*but not for pairing failures!)*
    - Possible structures [using this in coding]

Number (in tRNA gene)	What	optional or required?
1	acceptor stem (left)	required
2	acc-d linker	optional
3	d arm (left)	required

4	d loop	required
5	d arm (right)	reauired
6	d-ant linker	optional
7	anticodon arm (left)	required
8	anticodon loop (left of actual anticodon)	required
9	anticodon itself	required
10	anticodon loop (right of actual anticodon)	required
11	anticodon arm (right)	required
12	variable loop	required
13	t arm (left)	required
14	t loop	required
15	t arm (right)	required
16	t-acc linker	optional
17	acceptor stem (right)	required
18	acceptor stem overhang	required

Inputs (get by running with --help)

-trnass trnas.SS Path to tRNAscan-SE .SS file for tRNAs to process here (reference tRNAs). Should have TRUE genomic coordinates to match those in VCF

-vcf variants.vcf Path to VCF file for species whos tRNAs are provided

-out outfile.txt.gz Path to output file. If .gz, will be gzipped

-forcepseud [True, False]
During secondary structure processing: If gene is pseudo with codon listed as NNN as 0-0, try to force processing by trying codon at 34-36. (Default : True)

Outputs

- log file
- flags of when there were issues
- Main output file: one line per variant in VCF that was in one of the tRNA locations provided.
  Columns:
  chrom: chromosome
  pos: position of **VCF variant**
  ref: ref allele of variant - as in VCF
  alt: alt allele of variant - as in VCF
  nHomRef: # samples in VCF with 0/0 gts

nHomAlt: # samples in VCF with 1/1 gts
nNotMissingHet: # samples in VCF with 0/0 or 1/1 gts (not 0/1, missing)
tRNA: tRNA gene name
tRNA_strand: strand of tRNA
tRNA_pos: position of variant **relative to tRNA non intronic bases
structure: structure of tRNA this base is in
substructure: substructure of tRNA this base is in
nSubStructure: count substructure is from L to R (of all substructures)
substructure_pos: position of variant **within its substructure
nMissingHet: # samples in VCF with 0/1 or missing gts
homRef: comma-separated samples with 0/0 gts
homAlt: comma-separated samples with 1/1 gts
missingOrHettRNA: comma-separated samples with 1/1 or ./. gts

secstruct_related/trna2salignments.nf

Wrote python script that will break each tRNA allele into its secondary structure components (where possible); now want to run this and then align each structure component taking into account structure using 4sale.

- Parameters

Category	Flag for script (in script as params.<this)	Default value (if highlighted, need to provide)	Description

data	--runinfo	""	tab-delimited file, one row per sample/species to run. Columns:
			// id, sample/species ID (used for output file naming)
			// trnass, tRNAscan-SE .ss filepath for ALL alleles
			// specprefix, species prefix (no spaces) for FASTA seq naming for combining across species

data	--out	""	Outer/parent output directory (multiple directories created internally)
data	--alignmentout	"all_species_alleles_foursale_aligned.xfasta.gz"	File name for final alignment file
software param	--gapopen	0	-gapopen passed to foursale/clustal22
software param	--gapextend	0	-gapextend passed to foursale/clustal22
software param	--combfull	True	-full for fastax concat'ing: True or False: keep all sequences, like full/outer join
software param	--combfullfill	"N"	-fill for fastax concat'ing: "fill with N bases/residues for IDs missing in some files when using -full"

software param	--replaceU	True	-replaceU for fastax concat'ing: True or False: replace any Us in seq with Ts
software	--pyscriptdir		directory containing script secstructrelated/secstruct2seqpieces.py; utilityscripts/*concatxfastas.py*
software	--foursalerunjar		tring to put after java to get foursale to run. Should include all cp dependencies.
software	--clustalw2		/ fully articulated path to clustalw directory that runs with foursale (passed to foursale)
software	--params.vcfconda		path to conda environment set up to run scripts that import pyvcf (don't actually need pyvcf, but this one has other dependencies I use)

- Processes

Name	**Description**	**Any saved outputs? Where?**
secstruct2seqpieces	runs secstruct2seqpieces.py	ALL ouptuts of this script saved for future reference In directory: allelesecstruct (subdirectories for each species)
combinexfastas	Combines X fastas across species (within secondary structure pieces)	NA
foursale	Runs foursale for each secondary structure alignment piece	*no; these are intermediates*
combinealignments	Combines aligned outputs across secondary structures *With custom script: concatxfastas.py* *(seqkit works for fasta but not xfasta)*	Alignment in directory: fullalignments

secstruct_related/secstruct2seqpieces.py

Script that identifies sequence elements for each tRNA so, for example, broken up segments can be separately aligned (within secondary structure)

- NOTES on classifications
  - If stems have a mismatch, they are assigned the typical known length (in some cases, this could mean a linker or a bit of a different stem is included with them). *Am working to flag when things like this might be the case*
  - There ARE cases where borders are off by one - especially T arm vs. acceptor stem, for example [tricky cases where first acceptor stem base is unpaired, gets classified as having 2 bp overhang instead, etc]
  - See also notes for secstruct_related/get_strain_variants_relpos.py
- Inputs (can also get by running script with --help)

  -trnass trnas.SS Path to tRNAscan-SE .SS file for tRNAs to process here

  -outdir outdir Output directory

  -out outfilestem Prefix for output files

  -forcepseud [True, False]

  If gene is pseudo with codon listed as NNN as 0-0, try to force processing by trying codon at 34-36. (Default : True)

  -fasta [True, False] Write a fasta for each sensible secondary structure chunk of tRNAs. [default : True]

  -xfasta [True, False]

  Write a X fasta (where line after sequence is secondary structure in (). format) for each sensible secondary structure chunk of tRNAs. [default : True]

  -fanameprefix species_prefix

  Optional prefix for all sequences in output fasta/xfasta files - e.g., species identifier if files will be compined across multiple species.

  -maskanticodon [True, False]

  N out anticodons in output fasta/X fasta? (they are saved in their own FASTA) [default : True]

- Outputs
  - *_tRNA2struct_info.txt.gz: Information on tRNAs, one row per tRNA. NB doesn't include global stuff like position, strand as this doesn't make sense for strain-specific sequences that might be provided [has to come from elsewhere]. [doesn't include ones where couldn't split the seq well; DOES include ones with 'fake' codon positions if they were pseudogenes with 000 and NNN as codon in .SS file and script could make it work calling codon at 34-36]

    Columns (DESCRIPTIONS of them can be found in summary file):
    tRNA: gene name
    PossiblePseudogene: T or F
    Intron: T or F
    NoFlags: T or F. Not flagged at all - followed all rules perfectly as far as we can tell
    TooShort: T or F
    CodonIssue: T or F
    ReqStructuresMissing: number of required tRNA cloverleaf structures missing in this one (should be 0)

RemainingBases: number of bases in tRNA not assigned to a structure (should be 0)
AnticodonArmIssue: T or F. Anticodon arm stem not as expected (could be just one unpaired base, or worse)
AccStemIssue: T or F. Acceptor stem not as expected (could be just one unpaired base, or worse)
DArmIssue: T or F. D arm not as expected (could be just one unpaired base, or worse)
DLoopIssue: T or F. Unable to identify D loop, VERY WEIRD & UNTESTED [*can happen when fake out pseudogenes to try to process them]*
TArmIssue: T or F. T arm not as expected (could be just one unpaired base, or worse)
TLoopIssue: T or F. Unable to identify T loop, VERY WEIRD & UNTESTED

- *_tRNA2struct_summary.txt: summary of number of tRNAs with various characteristics. Columns:
  Category: flag category (intron, codon issue, etc)
  Description: description of category (as defined in this script)
  N: number of tRNAs flagged for this category
- *_tRNA2struct.txt.gz: For each base in tRNA, what structure does it map to. For all tRNAs where this was doable (NOT necessarily all tRNAs), one line per base in non-intronic length of that tRNA [1 indexed].
  - Columns:
    - tRNA: gene name
    - pos.tRNA: position in the tRNA (1-length of tRNA) this row describes
    - Structure: overall structure this position is assigned to (e.g. 'D' for D arm)
    - SubStructure: substructure this position is assigned to (e.g. 'arm_L' for L/first part of paired arm before D loop)
    - num.structure: number this substructure is in the tRNA (with acceptor stem L starting at 1)
    - pos.substructure: position this base is within its substructure [for comparing across tRNA]
    - nucleotide: nt at this structure (from sequence in input SS file)
    - secstruct: sec structure in . >< format at this structure (from sec structure in input SS file)
- FASTA related (if fasta or x fasta or both specified by inputs): ***excludes any it couldn't predict for****
  - *_<secondary structure descrip>.<x>fasta: fasta/x fasta for each piece of each tRNA. Name is tRNA name optionally prefixed (e.g. with species name) if input dictates. Files written (<secondary structure descrip>):

accstemL2d	acceptor stem to d loop [half], not including any d
darm2ant	d arm to anticodon arm, not including any anticodon
antarm	anticodon arm *with actual anticodon masked to N if specified*
anticodon	Actual anticodon (for checking; for use later if N'ed it out)
varloop	variable loop (everything between anticodon arm and t arm)

tarm2acc	d arm to acceptor stem, not including any acceptor stem
accstemR2end	second half acceptor stem plus any overhang

Related to variation in tRNA flanking regions

flank_specific/trnaflankvcfvars.nf

Workflow that gets the variants in tRNA gene flanking regions, notes which strains have which variants.

- Parameters/inputs

Category	Flag for script (in script as params.<this)	Default value (if highlighted, need to provide)	Description
data	--runinfo	""	tab-delimited file, one row per sample/species to run. Columns: // id, sample/species ID (used for output file naming) // trnasout, tRNAscan-SE .out filepath // trnabed, tRNA bed file path // vcf, path to VCF for this species - used to find all the tRNA variants // vcftbi, path to VCF .tbi index
data	--out		Outer/parent output directory (one per sample/species ID created internally)
data	--totalflank	40	flank length to get for tRNA flanking bed. **Just for output bed file, this does NOT change how Corinne's script works
data	--innerflank	20	flank length for inner tRNA flank for output bed file (outer is this to total). **Just for output bed file, this does NOT change how Corinne's script works
software	--vcfconda		path to conda environment set up to run scripts that import pyvcf
software	--wormtrnarepo		Directory containing gitrepo python scripts run here - updatedinitial/get_strain_variants-flank.py and utilityscripts/bed2flanks.py

- Processes

<table>
<tr><th>Name</th><th>Description</th><th>Any saved outputs? (all in output dir / ID of this species/sample) And relevant notes</th></tr>
<tr><td>bed4vcf</td><td>Generate bed file that includes tRNAs and generous flanking regions</td><td>no</td></tr>
<tr><td>subsetvcf</td><td>slice VCF to be only tRNA + flanking regions; Edits VCF so python vcf reader can work with it (unzips; removes "malformed filter lines", etc)</td><td>no</td></tr>
<tr><td>strainflankvars</td><td>Runs Corinne's get_strain_variants-flank.py</td><td>*variants-flank.txt, info on variants in flanks (in Corinne's format)
*variants-flank.log, log file</td></tr>
<tr><td>bedflanks</td><td>Makes bed file with exact flank regions of interest (split into inner, outer, 5', 3' and named as such)
Generated from tRNA bed file

python script does this: flank_specific/bed2flanks.py
Inputs:
-h, --help show this help message and exit
-bed regions.bed Path to bed file containing regions for which to get flanking regions (e.g., tRNA genes)
-nflank N Length of desired total flanking region (on either side of region, not combined) in bp
-ninner N Length of desired INNER flanking region (on either side of region, not combined) in bp. Bases between this and nflank will be classified as outer flanking regions
-out out.bed Output file to contain flank regions bed</td><td>_trnas_flanks.bed: BED file with 6 flank regions per input tRNA:
<table><tr><th>Name</th><th>What</th></tr><tr><td>flank_total_5</td><td>Total flank length (in workflow defaults to 40) on 5' end</td></tr><tr><td>flank_inner_5</td><td>inner flank (in workflow defaults to 20) on 5' end</td></tr><tr><td>flank_outer_5</td><td>outer flank (in workflow defaults to 20) from total flank end to inner flank</td></tr><tr><td><same as each of above but _3></td><td><same as each above for 3' end></td></tr></table>
Idea here is that can use this to feed to various summary programs</td></tr>
</table>

flank_specific/flankvariation.R

Transforms data for downstream use; makes flank variant plots (doesn't include gene body)

- Notes on classifications, metrics, etc
  - For splitting tRNAs into active or inactive, active is ones where ANY strain has non-lost alleles. Inactive is where ALL strains have lost alleles (technically, all alleles classified as 'lost'). I don't actually think this is useful; more useful would be to use predicted gene activity a la Thornlow 2020
- Path/location
  - In wormtrna gitrepo
  - wormtrna/trnavcfvars/flankvariation.R
- Inputs (also get by running script with --help)

```
-s, --speciesf     File containing information on all species to
                   process here. Columns infilename (exactly how all
                   files have this species in their name),
                   displayname (name that should be used for plot
                   outputs etc), shortname (no-spaces name for ouptut
                   files, sorting, etc - either shorter than or same
                   as infilename, probably).  In order you'd like
                   plots to be in!
-f, --flankvars    EXAMPLE Path to file with tRNA flank variant info
                   output by get_strain_variants-flank.py.  Where
                   species ID/species specific info is, put SAMP
                   instead
-t, --trnainfoal   Path to file containing tRNA per allele
                   information including original codon & allele as
                   well as (potential) remolded one. -
                   *alleleinfo_counts_worigcodonetc.txt output of
                   show_mutational_variation.R
--trnainfog        EXAMPLE Path to file containing tRNA per gene
                   information including name & strand. -
                   *strain_trnas_info.txt output of
                   build_alt_sequences.py Where species ID/species
                   specific info is, put SAMP instead
-l, --lflank       Start and end of basepairs away from gene used in
                   input for 5' flank, comma-separated. E.g -40,0
                   [default: -40,0]
-r, --rflank       Start and end of basepairs away from gene used in
```

input for 3' flank, comma-separated. E.g 0,40

[default: 0,40]

-b, --baseoutname Base name for all output files [default: out]

-o, --outdir Output directory path. if getwd(), current will be

used [default: out]

- Outputs
  - *_perpositiontRNAvarianceprop.txt.gz: key data (proportion of tRNAs with variation at each position) in looooong format. Columns:

Name	description
displayname	species name
strand	strand restriction (or not) for data in these rows - e.g., all tRNAs on + strand
genes	Genes restriction (or not) for data in these rows - all, active, or inactive (inactive =all alleles classified as lost)
varfreq	this is done for all variants (subset as other columns suggest) AND for only vars with allele freq < 0.05
mutclass	all, any.SNV, other [indels], or specific base change this data is for
. alleles	- for multi-mutation classes
. relpos	relative position in flank. **All tRNAs are now in terms of + direction - 3' or 5' call came from original strandedness, but relative position is for + direction**
flank	3' or 5'. **All tRNAs are now in terms of + direction - 3' or 5' call came from original strandedness, but relative position is for + direction**
n.tRNAs.var,	# tRNAs with variation at this position
n.tRNAs.invar,	# tRNAs withOUT variation at this position. NB missing not taken into account here - it's on a per tRNA not allele freq/per strain basis
p.tRNAs.var	proportion of tRNAs (missing excluded) that have variant at this position
low95ci.tRNAs.var	binomial 95% CI on proportion here - lower bound
high05ci.tRNAs.var	binomial 95% CI on proportion here - upper bound

  - *_locVpropvartRNAs_*<species>.pdf - Location relative to tRNA gene vs. proportion of tRNA genes that have mutation in that specific region. (Corrected for strandedness - all are plotted here where relevant). Each page has tRNA gene sets (all, active, inactive i.e. NO non-pseud alleles) x All mutations or mutations with MAF < 0.05. Then, each page is a different strand and/or different mutation class/way of coloring mutation classes

Analyses of tRNA gene and allele data

analysis/show_mutational_variation.R

Big 'starter' script that does a lot of summarizing and plotting of tRNA mutational variation

- Inputs (also get by running with --help)

```
-s, --speciesf     File containing information on all species to
                   process here. Columns infilename (exactly how all
                   files have this species in their name),
                   displayname (name that should be used for plot
                   outputs etc), shortname (no-spaces name for ouptut
                   files, sorting, etc - either shorter than or same
                   as infilename, probably).  In order you'd like
                   plots to be in!
-a, --alleleinfo   EXAMPLE path to *alleleinfo_counts_wmissing.txt
                   output of getstrainxtrnacalls.R. Where species
                   ID/species specific info is, put SAMP instead
-t, --trnasgen     EXAMPLE path to *strain_trnas_gen.txt output of
                   build_alt_sequences.py. Where species ID/species
                   specific info is, put SAMP instead
--trnainfo         EXAMPLE path to
                   *_bygeneup_tRNA_counts_wmissing.txt output of
                   getstrainxtrnacalls.R. (Columns about tRNA; counts
                   of alleles w/ various classifications.)  Where
                   species ID/species specific info is, put SAMP
                   instead
-c, --chrlens      File containing lengths of chromosomes to use for
                   plotting; chromosomes/contigs not included from
                   those analyses/plots. Columns displayname
                   (matching species info in speciesf), Chr, Length
-b, --baseoutname  Base name for all output files [default: out]
-o, --outdir       Outer output directory. Sub-directories will be
                   created internally. **NB: if you provide getwd()
                   here (quote wrapped), current directory will be
                   used
```

- Outputs
  - Updated data

- *_alleleinfo_counts_worigcodonetc.txt: Same as --alleleinfo inputs, but with added columns to map the codon to AlleleCM instead of AA (so can do plots grouping things by estimated original codon/AA, not what the DNA seq codes for). New columns:
  -

freq.ofcalled	Frequency this allele is of all called alleles
freq.ofallinclmissing	Frequency this allele is of all strains
Codon.orig	what the codon was at the original amino acid this tRNA codes for - usually same as Codon, but in case of isotype switches, this helps group it with AlleleCM
AlleleCM.orig	what is the likely AlleleCM in original amino acid this tRNA codes for - so that all alleles from same gene can be grouped even in the (rare, I think) cases where AlleleCM differs
all.altered	usually F; or T - all alleles were flagged as isotype switches here; useful to know about this. CAN be this AND have classification be 'Lost'

  - (5/12/25 considering updating to make sure ones with no codon ID'd by tRNAscan-SE are classified as Lost…but that might should happen upstream….yep, added this upstream)
- Plots showing allele information
  - *All in subdirectory /alleleinfoplots*
  - *tRNA_allele_barplots.pdf: stacked bar plot where each gene is a bar; faceted by codon and amino acid
    - ***note this is for 'coded for' AA, not alleleCM/gene group
  - *<species>_tRNA_allele_stackedpies_<AASplit or AAPlusCodonSplit - are genes grouped within amino acid only/merged across codon, or faceted down to codon level>.pdf: Stacked pie charts, one pie per gene. Pages have different plot versions, gene sets as specified in titles and by looking at plot; each PDF has all for one species.
    - Here, genes are grouped with the 'original' codon they should go to and the 'original' **AlleleCM** for that gene.
    - if prop is < 0.03, pie slice is inflated (and total pie inflated with it) - just for visualization!
- Plots showing tRNA location (+ possibly other info)
  - *All in subdirectory /genelocplusplots*
  - These are made using *strain_trnas_gen.txt output of build_alt_sequences.py: number of alleles etc from there at least at first!! [not currently intersecting with classification etc]]
  - *_tRNA_locVsNAllelesEtc.pdf: location on chromosome of tRNA vs. number of alleles that tRNA has. Faceted by species. 16 versions - with and without

histogram showing N genes pileup; faceted & free y axes; normal and log scale y axes; pseudogenes noted and not
        - In areas of chromosome that one species has but not other - at the end- the species that doesn't go that long has grayed out region
    - *_tRNA_locByAA.pdf- Shows counts in 500kb along chromosomes split by amino acid. So can compare locations. 4 plots per species: all genes; excluding genes pseudo in ref; highlighting genes pseudo in ref; and highlighting genes with any pseudo alleles
        - Note: this is just COUNTS - harder to compare amino acids by eye since they have different n genes, but doing the proportion within amino acid is harder than worth the time *right now*
        - Right now, there's an 'NA' amino acid row added - this is just to get the coordinates right, didn't want to add counts of fake genes to real data. *Gotta be a better way…*
    - *Testing (and plotting) if tRNA genomic location distributions differ based on multiple characteristics*
        - Characteristics tested:

Pseudo. in ref.?
Any pseudo/lost alleles?
Variable in population?
Any strains have any missing genotype calls?

- *_tRNA_groupings_location_KStests.txt: results from Kolmogorov-Smirnov tests of genomic position distributions for each characteristic above, per chromosome and overall (overall = genomic coordinates; chromosomes added together). *P-values are also noted on plots.* Specific columns:
    -

displayname	species
Chr	chromosome or "all_combined" for overall genome distribution
tested	description of characteristic tRNA genes were split on - to test T and F location distributions against each other
ks.D.stat	test statistic
ks.p.value	p value

- *_tRNA_groupings_location_KStests.pdf: two plots per characteristic, faceted by chromosome and species, showing the distributions of genes TRUE for the characteristic vs. FALSE for the characteristic. KS per-chromosome p values annotated on facets; overall p-values are in subtitles. *Dark gray rectangles denote areas of the chromosome that don't exist in that species - i.e., where another species has a longer chromosome*
- Related to examining if pseudogene alleles in reference are functional elsewhere
    - *_pseudoinref_genes.txt - one row per gene that is pseudo in ref, contains total number alleles and number that might be functional and strain numbers for different allele classifications

- *_pseudoinref_summary.txt - per-species summary: n_pseudo_in_ref (# genes pseudo in ref), n_pseudo_in_ref_mult_alleles (# of these that are variable in population), n_pseudo_in_ref_has_fn_allele (# of these that have putatively functional allele)
- *_pseudoinref_details_putativelyfunc.txt -input (alleleinfo file) info for each allele from genes that are pseudo in ref but also have putatively functional alleles

analysis/genebodyvariation.R

Does some data re-formatting and plotting of variants internal to gene body (developed with secondary structure aware methods).

- Notes
  - First wanted to get prop tRNAs with variant each bp, but with different lengths this gets quite gnarly. Instead, going to get rate in each *substructure* [normalized to # *canonical* bases in that substructure

    # Get each site....a little tricky here...do within each structure instead, perhaps...

    # [if convert to certain # bps, how deal with when one tRNA has longer/shorter structure than canonical? average it out across the rest of the bases?]

    # kinda cleaner to do per structure then later (when plotting) assign 'canonical' len to each structure??
- Inputs (also get by running script with --help)

```
-s, --speciesf     File containing information on all species to
                   process here. Columns infilename (exactly how all
                   files have this species in their name),
                   displayname (name that should be used for plot
                   outputs etc), shortname (no-spaces name for ouptut
                   files, sorting, etc - either shorter than or same
                   as infilename, probably).  In order you'd like
                   plots to be in!
-g, --genevars     EXAMPLE Path to file with tRNA gene body info
                   output by get_strain_variants_relpos.py Where
                   species ID/species specific info is, put SAMP
                   instead
-t, --trnainfoal   Path to file containing tRNA per allele
                   information including original codon & allele as
                   well as (potential) remolded one. -
                   *alleleinfo_counts_worigcodonetc.txt output of
                   show_mutational_variation.R
--trnainfog        EXAMPLE Path to file containing tRNA per gene
                   information including name & strand. -
                   *strain_trnas_info.txt output of
                   build_alt_sequences.py Where species ID/species
                   specific info is, put SAMP instead
--ssflags          Path to file containing per-allele information
                   categorizing how well sec structure assignment
```

went ('info' file output of
secstruct2seqpieces.py) Where species ID/species
specific info is, put SAMP instead

--trnasecstruct Path to file with information on how tRNA sec
structures in --genevars input are arranged, etc.
Columns: structure (as in --genevars),
substructure (as in --genevars), nSubStructure (as
in --genevars), canonicalnbp - length of this
structure in a 72/73 bp tRNA [for plotting],
structure_plot and substructure_plot: categories &
names prettified/simplified for how you'd like to
plot them; structure_plot_level &
substructure_plot_level: RANKINGS of unique
structure_plot and substructure_plot for ordering
in plot

-b, --baseoutname Base name for all output files [default: out]

-o, --outdir Output directory path. if getwd(), current will be
used [default: out]

- Outputs
  - *_persecstructtRNAvarianceprop.txt.gz: Key data. Long! All structures x all species x all mutation types information. Columns:
    - displayname, species info
    - strand, Strand analysis is narrowed to for tRNA genes (or all)
    - genes, all, active, or inactive tRNAs
    - secstructflags, Was this analysis restricted to genes without any 'issues' in the python script calling their secondary structure? (Sometimes issues are no big, like a mismatch in acceptor stem; other times they could in theory throw off more of the structure)
    - mutclass - name of mutclass here
    - alleles - "-" if doesn't make a difference; Ref > Alt or Major > Minor for specific mutation classes
    - structure, substructure, nSubStructure, canonicalnbp - from --trnasecstruct input; information about which tRNA secondary structure piece is described in this row
    - n.tRNAs.var, # tRNAs with one or more variants in this substructure
    - n.variants, absolute # variants observed in this substructure
    - n.tRNAs.invar, # tRNAs with no variants recorded in this substructure
    - <p, low95ci, high95ci>.tRNAs.var.any - proportion & 95% CI (binomial) for number variant tRNAs over all. Not corrected by structure length.
    - <p, low95ci, high95ci>.varsPerBp - proportion & 95% CI (binomial) for number variants in region normalized to length of that region from input (may be

imperfect, but closer to per-bp number). Multiple variants from same tRNA count where relevant.

- plots
    - *NB: not perfect yet - showing 1bp is tricky & depends on how the thing works…*
    - 1 per species per any variants vs. variants corrected per bp. Pages and facets have all other varieties!
    - when mutations shown combined, 95% Cis plotted - as thin black lines above/below the main line

analysis/flankandgenebodyvar_visualize.R

Puts flank region variation and gene body variation together – downstream of analysis/genebodyvariation.R and flank_specific/flankvariation.R

- Inputs (also get by running script with --help)

```
-f, --flankpertrna     Flank variant location, proportion of tRNAs
                       info: *_perpositiontRNAvarianceprop.txt.gz
                       output of flankvariation.R
-g, --genevarspertrna  Within gene body variant info/proportion of
                       tRNAs: *_persecstructtRNAvarianceprop.txt.gz
                       output of genebodyvariation.R
-t, --trnasecstruct    Path to file with information on how tRNA sec
                       structures are arranged as used in
                       genebodyvariation.R call. Columns: structure
                       (as in --genevars), substructure (as in
                       --genevars), nSubStructure (as in --genevars),
                       canonicalnbp - length of this structure in a
                       72/73 bp tRNA [for plotting], structure_plot
                       and substructure_plot: categories & names
                       prettified/simplified for how you'd like to
                       plot them; structure_plot_level &
                       substructure_plot_level: RANKINGS of unique
                       structure_plot and substructure_plot for
                       ordering in plot
-s, --speciesf         File containing information on all species
                       processed in preceeding script. Columns
                       infilename (exactly how all files have this
                       species in their name), displayname (name that
                       should be used for plot outputs etc - in other
                       input files), shortname (no-spaces name for
                       ouptut files, sorting, etc - either shorter
                       than or same as infilename, probably).  In
                       order you'd like plots to be in!
-b, --baseoutname      Base name for all output files [default: out]
-o, --outdir           Output directory path. if getwd(), current
                         will be used [default: out]
```

- Outputs
  - A bunch of plots!
  - All tRNA gene body values are normalized to canonical n bp in that structure

- All genes included regardless of secondary structure flagging

analysis/trna_location_analyses.R

Performs various analyses of tRNA location with regard to protein coding gene location; genomic recombination domains; other tRNAs; etc

- Inputs (also get by running with --help)

```
-s, --speciesf    File containing information on all species to
                  process here. Columns infilename (exactly how all
                  files have this species in their name),
                  displayname (name that should be used for plot
                  outputs etc), shortname (no-spaces name for ouptut
                  files, sorting, etc - either shorter than or same
                  as infilename, probably).  In order you'd like
                  plots to be in!. **If not all files exist for each
                  species, only does analyses that it can for each
                  species**
-t, --trnasgen    EXAMPLE path to *strain_trnas_gen.txt output of
                  build_alt_sequences.py. Where species ID/species
                  specific info is, put SAMP instead
--trnainfo        EXAMPLE path to
                  *_bygeneup_tRNA_counts_wmissing.txt output of
                  getstrainxtrnacalls.R. (Columns about tRNA; counts
                  of alleles w/ various classifications.)  Where
                  species ID/species specific info is, put SAMP
                  instead
-g, --genelocs    EXAMPLE path to no-header, 3 column file
                  specifying all genes' locations for a species.
                  Columns: chr, start, end (NOT header'ed!)  Where
                  species ID/species specific info is, put SAMP
                  instead
-c, --chrdomains  EXAMPLE path to chromosome domain info files (tip,
                  arm, center) - rename them if you need to for this
                  convention to work.  Should have columns chr,
                  domain, subdomain, start, end.  Where species
                  ID/species specific info is, put SAMP instead
--chrlens         File containing lengths of chromosomes to use for
                  plotting; chromosomes/contigs not included from
                  those analyses/plots. Columns displayname
                  (matching species info in speciesf), Chr, Length
-b, --baseoutname Base name for all output files [default: out]
-o, --outdir      Outer output directory. Sub-directories will be
                  created internally. **NB: if you provide getwd()
                  here (quote wrapped), current directory will be
                  used
```

- Outputs
  - Misc - not location realted, but made here - possibly useful for other scripts
    - *_identicalseqgroupinfo.txt: Information on all seqeunces ***from input fastas*** *(may not match other inputs)*:
      - displayname, species
      - tRNA, tRNAscan-SE tRNA sequence ID, with 'chr' stripped if it was present
      - seq, sequence of that tRNA
      - seqlabel, unique sequence number - same sequences have the same number
      - nwseqlabel, number tRNAs with this same sequence in ref genome
    - *_tRNAgeneinfo_combined.txt: summary of all tRNA info pulled in or generated here. *NA when a tRNA didn't have that info, e.g. some seqs for elegans and domain for tropicalis*. Columns:

displayname: species

tRNA: tRNA name
seqlabel: **new,** what sequence GROUP is this tRNA in (all with this same label have same sequence)
nwseqlabel: **new,** how many other tRNAs in this ref genome have the same sequence
AA: amino acid *in reference* (from strain_gens)
Codon: codon *in reference* (from strain_gens)
Chr: chromosome of gene
Start: start pos of gene
End: end pos of gene
Strand: strand of gene
pos: midpoint of gene - for plotting one pos
domain: chromosomal recomb domain of midpoint where available
subdomain: "" but subdomain - L, r, etc
VariableInPop: multiple alleles, T or F
nAlleles: # alleles, from my summary
numStrains: # strains per allele
anyMissingCalls: any strains have missing calls
Best: # best alleles
Functional: # functional alleles
Altered: # switch alleles (PRELIM, may not match with full analysis later)
Lost: # pseud/lost alleles (PRELIM, may not match with full analysis later)

- *_tRNAgenecounts_refsequniqueness.txt: Summary of how many genes of different categories there are - total, with ref seq in fasta, UNIQUE gene sequences in genome, and this for genes with no pseud alleles, not all pseud alleles (*from early calling*)

- General gene location related
  - *_genecounts_per500kb_trnasandall.txt: For 500kb in the genome *(last bin on chr can be from 0.5-1.5 this bin length),* how many (number AND proportion) tRNA, non-tRNA genes are there. Columns:
    - displayname, species
    - description, which tRNA genes used here : gene set - all possible tRNAs (including pseud), functional tRNAs (at least one non-lost allele), Pseudo tRNAs (no non-pseud alleles)
    - genes, which genes are counted in this row? all (from background input geneset - prot coding genes expected), tRNA (tRNAs that also meet criteria in description),
    - chr, bin chromosome
    - start, bin start
    - end, bin end
    - bin.mid, bin midpoint
    - n, number of given genes here
    - p, proportion of all genes of the given category that are in this bin
    - p_low95ci, lower 95% binomail CI on p
    - p_high95ci, upper 95% binomial CI on p
  - *_allGsvstRNA_per500kb_plots.pdf: plots of above - line plots showing genes per 500kb; various gene sets, with and without 95% CI. Has domains overlaid.
- *Chromosome recombination domain related*

- All in subdirectory domain_analyses. **Only generated for species that had all necessary inputs.**
- *_geneCounts_subdomains_withperkb.txt: for multiple gene sets, counts of genes and number genes per Kb for each genomic subdomain. Columns:
    -

description	gene set - all possible tRNAs (including pseud), functional tRNAs (at least one non-lost allele), Pseudo tRNAs (no non-pseud alleles)
displayname	species nice format
shortname	species short format
chr	
domain	tip, arm, or center
subdomain	left or right, if appropriate
level	"subdomain" (can ignore)
genes	which genes are counted in this row? all (from background input geneset - protein coding genes in general), tRNA (tRNAs that also meet criteria in description),
N	number of genes in this region for these characteristics
N_per_kb	N/kb length of this region

- *_geneCounts_domainsup_withperkb.txt: for multiple gene sets, counts of genes in domains combined across chromosomes and genome wide (so, all arms on that chromosome or all arms genome wide; on a chromosome, center is the same for here or subdomain). *Columns are a subset of those above.*
- *_tRNAGeneProportion_subdomains.txt: what proportion of genes in each region are tRNAs. As above, tRNAs are also subset by *description* column. Number columns here are p (proportion), p_low95ci and p_high95ci (binomial 95% confidence intervals on proportion)
- *_tRNAGeneProportion_domainsup.txt: as above, but what proportion of genes in larger/summarized windows (e.g. all arms summed) are tRNAs. *Probably less useful.*
- *_withingeneset_regionchisq_allchrs.txt: chi-sq result for distribution of genes *within each category* across chromosomal domains, summed each chromosome. I.e., do arm and center proportion differ *within tRNAs*
- *_withingeneset_regionchisq_perchr.txt: chi-sq results as above (*within each category*), but done for each chromosome domain/subdomain set separately
- *_betweengeneset_regionchisq.txt: chi-sq results **testing proportion tRNAs in each domain vs. non-tRNAs in each domain.** Overall and per chromosome. For all tRNAs, functional only, pseudol only (*not testing pseudo & not against each other though)*
    - ***this is what would use to annotate plots, for example***
- *_genestrnas_perkbprops_indomains.pdf: shows genes per kb of tRNA, non-tRNA genes; proportion of genes on each chromosome that are in each arm

○ Related to distance among genes

- *Notes*
  - NOTE: codon here is from first file (_strain_trnas_gen.txt): identity of amino acid assigned by tRNAscan-SE *in reference genome*
  - NOTE, this is for all the seqs tRNAscan-SE IDs - NOT cleaned up narrower number from gtrnadb (for *C. elegans*, this means its my FASTA, not trnascan-se's)
- All in directory /pairwisedistance/
- *_tRNApairdistances.txt.gz: allll the data: *all* pairwise distances between tRNAs in input here (>816k currently!!), with characterizations of the 'type' of pair it is for splitting/analysis. Columns:
  - displayname, species
  - genepair, tRNA gene names paired (- separated, sorted) - so can match with any OTHER info about genes you might decide to add later…
  - pseud, 0, 1, or 2 - how many of the genes compared here had all alleles classified as Lost after first pass. ***early/imperfect pseud call***
  - AA, amino acid pair (from AA column of input) [sorted]
  - Codon, codon pair (from AA column of input) [sorted]
  - chr, chromosome (one if same chr, 2 sorted if not)
  - isoacceptor, T if pair is same AA, F if not
  - isodecoder, T if pair is same Codon, F if not
  - sameseq, T if pair has same sequence (seqlabel), F if not
  - lonelyseq, 0, 1, or 2 - how many of the genes are the only one of their classification that has it's seq ID (nwseqlabel = 1)
  - samechr, T or F: was pair on same chromosome
  - distance, distance in bp between genes. *arbitrarily 30Mb here if on diff chromosomes*
- *_tRNApairtypes_nsumm.txt: summary of the number of pairs with different classifications of interest. For all tRNAs (pseud & not) and excluding those with all pseud alleles.
  -

Different isoacceptor	not same AA; obligately also different isodecoder and different sequence
Same isoacceptor	any two genes with same AA
Same isoacceptor, different isodecoder	two genes have same AA, different codons [subset of isoacceptor]
Same isodecoder	same codon [subset of isoacceptor]
Same isodecoder, different sequence	same codon, sequence otherwise differs [subset of isodecoder]
Same sequence	genes have identical sequence [subset of isodecoder]

- *_tRNApairdistance_propsamechr_stats.txt: Results from proportion test comparing proportion of one set of gene pairs that are on the same chromosome vs another. Species, genes, sets, then stat test info (estimates, p values) shown. Comparisons done:
  -

set1name	set2name
Different isoacceptor	Same isoacceptor
Same isoacceptor, different isodecoder	Same isodecoder
Same isodecoder, different sequence	Same sequence
Different sequence	Same sequence

- *_tRNApairdistance_MWcompare_stats.txt: results from Mann-Whitney/Wilcox test testing distance distribution for tRNA pairs in different sets (see above - 'Comparisons done'). For all chromosomes combined but excluding chromosome-different pairs; each chromosome individually (to make sure chromosomal distribution not underlying differences); all combined where diff chromosome pairs arbitrarily assigned 30Mb apart [can mess up medians!]
- _tRNApairdistance_propsamechr_plot.pdf: for each of the gene sets/subsets (set1 or 2), plots proportion of genes on the same chromosome. *No Cis for now.*
- *_tRNApairdistance_vsseqdiv.pdf: hex tile plot of sequence divergence (new-ish input) between alignments vs. pairwise distance (for pairs where both had structure alignments). Many pages: all together vs split by chr; log scale distance vs not; different seq divergence metrics
- *Modeling distance on difference between sequence alignments*
  - *_tRNApairdistance_seqdiffs_glms_gammalog.RData: Quick save out of all the actual model objects in case want details long term. Named by all the info about them (ugly), described in other outputs. Outer list is of species - multiple models in each species.
  - *_tRNApairdistance_seqdiffs_glmsummary_gammalog.txt: The models run (descriptions, formula, etc) and key summary criteria - dispersion, AIC, and effects for the seq diff parameter, where included. Columns:
    - Displayname, species these results are for

      seqdiffdescrip: how are tRNA sequence alignment differences counted
      seqdiffmeasure: name of data column corresponding to above
      description: description of this model
      data.slice: any data reduction done (same chromosome; pseud in or out)
      use.formula: formula used for model
      model.AIC: AIC of model
      model.dispersion: dispersion of model (resid deviance squared / resid, close to 1 suggests good fit)
      seqdiff.beta: Estimate of seqdiff effect (raw/as is)
      seqdiff.beta.se: SE of seq diff effect
      seqdiff.tval: T value of seq diff effect
      seqdiff.pval: p value of seq diff effect
      seqdiff.foldchange: Each unit change in seq diff has this *multiplicative* effect on pairwise distance (exp(beta, which was in log scale))
      seqdiff.percentchange: percent greater pairwise diff is for each seq diff unit change (in % format - 0.1% means 0.1%, not 10%)
- *_tRNApairdistance_seqdiffs_glm_gammalog_plots_<species>.pdf: plots of the GLM models: *ignore for super nulls!*

  # Residuals vs Fitted: Look for patterns (should be random scatter).

Normal Q-Q: Deviance residuals should roughly follow a straight line.
Scale-Location: Checks homoscedasticity.
Residuals vs Leverage: Identifies influential points.

- *_tRNApairdistance_seqdiffs_glmgammalog_mcfaddensr2.txt: McFadden's pseudo-R2 : proportion deviance explained by fuller model vs. nuller model. Some rows are for isolating contribution of seq diff, others for chrom vs null
- *_tRNApairdistance_seqdiffs_glms_gammalog_anovas.RData: ANOVAs for JUST the ones that include seq diff (since terms added sequentially anyway). In case want the full info at some point; summaries & stats etc in next outputs
- *_tRNApairdistance_seqdiffs_glmgamma_anovapdev.txt: ANOVA results and prop/percent deviance explained for each model where seq divergence was one of the factors. Columns:
    - displayname: species
    - seqdiffdescrip: which measure of sequence difference used here
    - description: model description, including if pseud included or not
    - Df: Df from ANOVA (on glm)
    - Deviance: from ANOVA (on glm)
    - Resid. Df: from ANOVA (on glm)
    - Resid. Dev: from ANOVA (on glm)
    - Pr(>Chi): from ANOVA (on glm)
    - prop.modeltermdeviance: Proportion of the deviance explained by the terms in the model that this term explains (*not super useful)*
    - percent.modeltermdeviance: Percent of the deviance explained by the terms in the model that this term explains (*not super useful)*
    - prop.totaldeviance: Proportion of the TOTAL deviance that this term explains
    - percent.totaldeviance: Percent of the TOTAL deviance that this term explains

analysis/pairwisediffsrefseq.R

Script that gets sequence 'distance' between reference genome sequence alignments (number of nucleotides or gaps that don't match and number unique stretches of same). Needed for location analysis script; a bit circular as currently it also uses an output of that script.

- *Inputs (also get by running script with --help)*

```
-s, --speciesf     File containing information on all species
                   processed in preceeding script(s). Columns
                   infilename (exactly how all files have this
                   species in their name), displayname (name that
                   should be used for plot outputs etc - in other
                   input files), shortname (no-spaces name for ouptut
                   files, sorting, etc - either shorter than or same
                   as infilename, probably).  In order you'd like
                   plots to be in!
-x, --xfasta       Foursale alignment xfasta for ALL tRNA alleles in
                   ALL species of interest
-i, --introninfo   EXAMPLE path to *_tRNA2struct_info.txt.gz output
                   file of secstruct2seqpieces.py, used here to ID
                   genes with suspected introns vs not.  where sample
                   info goes, put SAMP instead (as in speciesf)
--seqgroupinfo     *_identicalseqgroupinfo.txt output of
                   trna_location_analyses.R: information on all
                   seqeunces from ref genome - columns: displayname,
                   species tRNA, tRNAscan-SE tRNA sequence ID, with
                   'chr' stripped if it was present seq, sequence of
                   that tRNA seqlabel, unique sequence number - same
                   sequences have the same number nwseqlabel, number
                   tRNAs with this same sequence in ref genome
-b, --baseoutname  Base name for all output files [default: out]
-o, --outdir       Output directory path. if getwd(), current will be
                   used [default: out]
```

- Outputs
  - *_alignedseqdiffs_uniqueseqpairs.txt.gz: the relevant data! For each pair of *unique reference* tRNA alignments (after 4sale etc) in each species, provides:
    - genepair, tRNA genes used here, '-' separated
    - seqlabelpair, seq label of the genes compared here (so can match with other genes), '-' separated
    - **intron,** *added later,* number of genes in pair with intron [introns are excluded from alignment so any diffs in intron won't be captured here]

        - ntNoMatch, number of elements in alignment [nts] that were NOT identical
        - ntMatch, number of elements in alignment [nts] that were identical
        - stretchNoMatch, number of distinct STRETCHES of alignments that don't match - i.e., if elements are F F F T, this is 1 stretch non-match, 3 nts non match
        - stretchMatch, number of distinct STRETCHES of alignments that match...not sure this will be used, but might as well have
- **NOTE** this is for all reference seqs where I was able to assign all tRNA secondary structures - NOT all reference sequences in the seqgroupinfo input
- *_alignedseqdiffs_uniqueseqpairs_hists.pdf: histograms of number of gene pairs with different numbers/measures of mis-matches

analysis/trna_trees_init.R

Script that makes neighbor-joining trees of tRNA allele multiple sequence alignments (made in secondary structure aware fashion), messes around with plotting. Unlike most scripts, Right now plenty in here is **hard coded** for current species. Would have to update for others.

- Inputs (get with --help, mostly just doing it w/in Rstudio though)

```
-s, --speciesf     File containing information on all species
                   processed in preceeding script. Columns infilename
                   (exactly how all files have this species in their
                   name), displayname (name that should be used for
                   plot outputs etc - in other input files),
                   shortname (no-spaces name for ouptut files,
                   sorting, etc - either shorter than or same as
                   infilename, probably).  In order you'd like plots
                   to be in!
-a, --alleleplus   *alleleinfo_counts_worigcodonetc.txt output of
                   show_mutational_variation.R, with all allele info
                   including Codon.orig, AlleleCM.orig,
                   VariableInPop, Classification, etc
-p, --pseudref     *_pseudoinref_genes.txt output of
                   show_mutational_variation.R, with info on if each
                   gene is all pseud
-x, --xfasta       Foursale alignment xfasta for ALL tRNA alleles in
                   ALL species of interest
-b, --baseoutname  Base name for all output files [default: out]
-o, --outdir       Output directory path. if getwd(), current will be
                   used [default: out]
```

- Outputs
  - *_trees.Rdata: saves workspace every time a big computation is done (and so can use the NJ trees again another time - generating them is somewhat slow)
  - *Plots - info*
    - Allele frequency when shown is number with this allele/total number strains (including those with missing), not proportion of called strains
  - *Plots - file names (?)*
    - *_phylotrees_allspeciesgenes.pdf: all variants of phylo (branches unequal lengths) including all species and all amino acids. Different colorways, different inclusion/exclusion of pseudogenes.
    - *_cladotrees_allspeciesgenes.pdf: as above, but trees are cladograms (branch lengths not considered - all go to edge)

- *phylotrees_<species>.pdf - tree for alleles across Aas in each individual species. Gene, colorway subsets noted in title.
- *cladotrees_<species>.pdf - cladograms for alleles across AAs in each individual species. Gene, colorway subsets noted in title.
- *_trees_<specific amino acid>.pdf: Trees for specific amino acids (by that a.a.'s BACKBONE - all originally a leucine plotted, etc; excluding genes where all alleles are pseudo). First page is tree, second is cladogram.
    - **These for 4 that I picked specifically, later made for all amino acids**
- (added later) *_phylotrees_eachAA.pdf, Trees for specific amino acids (by that a.a.'s BACKBONE - all originally a leucine plotted, etc; excluding genes where all alleles are pseudo). Typical tree.
- (added later) *_cladotrees_eachAA.pdf, Trees for specific amino acids (by that a.a.'s BACKBONE - all originally a leucine plotted, etc; excluding genes where all alleles are pseudo). Cladogram!