Formatting data for 5-HTTLPR script (v8.1)
(Instructions are unchanged from v8.0.3)
Updated 2014-05-30

Notes:

1. Data files should be stratified by ancestral group (e.g. European ancestry, Admixed European and African).
   Each distinct ancestral group should be in a separate file.

2. Each data file must be in csv (comma separated values) format.

3. The first line of the data file is a header line containing an ordered list of the variables in the file.

4. All variable names begin with an upper case letter followed by all lower case characters.

5. Id, Sex, Age, and Gen5 (the 5-HTTLPR L vs S genotype) are required.
   Please omit any other variables that are missing for all subjects.

6. **First depressive episode** during the subject's lifetime is our primary depression phenotype.
   If information about this episode is available, please include variable values based on this information.
   If ONLY current depression was assessed, please use the "current" variables in your dataset.
   If BOTH were assessed **AND Life stress variables (exposure/quantitative) differ for the 2**, please create 2 datasets
   (If both were assessed, but Life stress variables can be shared, only one dataset is needed).

7. The test data file (TEST_ALLVARS.csv) included in the download package illustrates the required format.

Below is the complete list of variables and instructions for coding.
Please email Amy Horton (achorton@wustl.edu) if you have any questions

	Variable	Coding	Description of variable
1	Id	alpha numeric variable	Unique identification value for each subject
	Demographic Variables		
2	Sex	0 = female 1 = male -9 = missing	Sex of subject
3	Age	> 0 (decimal years) -9 = missing	Age of subject at interview in years
4	Birth_decade	4 digit decade of birth e.g. 1950 for birthdates from Jan 1, 1950 to Dec 31, 1959	4 digit decade of birth
	Genetic Variables		
5	Gen5	LL = 2 copies of the long allele LS = heterozygote SS = 2 copies of the short allele -9 = missing	Genotype of subject for 5HTTLPR L=long allele S=short allele
6	rs25531	0 = 0 copies of the G allele 1 = 1 copy of G 2 = 2 copies of G -9 = missing	# of G alleles for rs25531 Note: The G allele for rs25531 occurs most frequently with the L allele for 5HTTLPR. If phasing for Gen5 and rs25531 is known, set this variable to the number LG alleles. Otherwise, ambiguous haplotypes (double heterozygotes) will be assumed to be SA-LG
	Depression Variables		
7	Dep_dx_life	0 = no lifetime diagnosis (dx) of major depressive disorder 1 = positive lifetime dx of major depressive disorder	Ever in life had a depression diagnosis Subject has at some time during his or her lifetime qualified for a diagnosis of major depressive disorder

		-9 = missing	NOTE: *Preferred phenotype. Use if available.*
	Variable	**Coding**	**Description of variable**
8	Dep_quant_life	≥ 0 if assessed -9 = missing	Maximum Lifetime quantitative depression scale If subject has a lifetime diagnosis of major depressive disorder, give symptom count when first met diagnosis. Otherwise, give maximum lifetime value. **Use DSM-IV symptom count if it is available.** *Otherwise, use your own scale, **BUT** ensure that all scored values are ≥ 0, to avoid conflict with the missing value, -9. (Add a constant to all values if needed.)* **NOTE: *Preferred phenotype. Use if available.***
9	Dep_dx_curr	0 = no current diagnosis of major depressive disorder 1 = current diagnosis of major depressive disorder -9 = missing	Current depression diagnosis Subject has a diagnosis of major depression at the time of interview (e.g. within the last 6 months or last year)
10	Dep_quant_curr	≥ 0 if assessed -9 = missing	Current quantitative depression measure **Use current DSM-IV symptom count if it is available.** *Otherwise, use your own scale, **BUT** if necessary add a constant to all values so all scored values ≥ 0 (This avoids a conflict with using -9 for missing values.)*
11	Dep_life_ao	≥ 0 if known (decimal years) -9 = missing	Age of onset (decimal) of first major depressive episode during lifetime If never depressed = age at interview+1[1] **NOTE: *Preferred phenotype. Use if available.***
12	Dep_curr_ao	≥ 0 if known (decimal years) -9 = missing	Age of onset of current depression episode If not currently depressed = age at interview+1[1]
	Childhood Maltreatment Variables		
13	CTQ_EA	≥ 5 if assessed -9 if missing	CTQ[2] sub-score for emotional abuse *For those assessed, this value ranges from 5 to 25*
14	CTQ_PA	≥ 5 if assessed -9 = missing	CTQ sub-score for physical abuse *For those assessed, this value ranges from 5 to 25*
15	CTQ_SA	≥ 5 if assessed -9 if missing	CTQ sub-score for sexual abuse *For those assessed, this value ranges from 5 to 25*
16	CTQ_EN	≥ 5 if assessed -9 = missing	CTQ sub-score for emotional neglect *For those assessed, this value ranges from 5 to 25*
17	CTQ_PN	≥ 5 if assessed -9 if missing	CTQ sub-score for physical neglect *For those assessed, this value ranges from 5 to 25*
18	Child_mal_exp	0 = not exposed to childhood maltreatment 1 = exposed to childhood maltreatment -9 = unknown	**Exposure** to Childhood Maltreatment ***ONLY create this variable IF*** at least one of CTQ subscales for Physical Abuse (v14), Sexual Abuse (v15), or Physical Neglect (v17) is not available for most (or all) of the dataset[3]
19	Child_mal_q	≥ 0 if assessed -9 = missing	Quantitative Childhood Maltreatment score ***ONLY create this variable IF*** at least one of CTQ subscales for Physical Abuse (v14), Sexual Abuse (v15), or Physical Neglect (v17) is not available for most (or all) of the dataset[3]

	Variable	Coding	Description of variable

<table>
<tr><th></th><th>Variable</th><th>Coding</th><th>Description of variable</th></tr>
<tr><td colspan="4">Life Stress Variables (stress other than childhood maltreatment)
NOTE: If your data contains multiple assessments of Life stress (e.g. from longitudinal data) that differ relative to first depressive episode and current depressive episode, please create 2 separate datasets: one with information about the first depressive episode during lifetime (and related life stress), the other with information about current depression (and related life stress).</td></tr>
<tr><td>20</td><td>Life_stress_exp</td><td>0 = not exposed to LIFE stress during period assessed

1 = exposed

-9 = missing</td><td>LIFE stress exposure (dichotomous) (NOT including childhood maltreatment)
NOTE: The group's consensus was that Life stress "exposure" using the LTE-Q (Brugha criteria) should require ≥ 2 events.

Studies with other LIFE stress measures must each define stress "exposure" for their data.</td></tr>
<tr><td>21</td><td>Life_stress_q</td><td>≥ 0 if assessed

-9 = missing</td><td>Use LTE-Q score, if available
<i>Otherwise, use your own scale, BUT ensure all values are ≥ 0 to avoid conflict with the missing value, -9. (Add a constant to all values if needed.)</i></td></tr>
<tr><td>22</td><td>Age_last_stress_before_dep1</td><td>-1 = no LIFE stress events during period assessed

≥ 0 = age of most recent LIFE stress event before first depressive episode (if EVER depressed)
OR
age at most recent LIFE stress event before interview (if NEVER depressed)

-9 = exposure unknown or age of exposure unknown</td><td>Age during most recent LIFE stress event <i>before</i> FIRST major depressive episode
NOTE 1: LIFE stress is stress OTHER THAN childhood maltreatment
NOTE 2: If no major depressive episodes, list age of most recent LIFE stress before interview</td></tr>
<tr><td>23</td><td>Age_last_stress_before_dep_curr</td><td>-1 = no LIFE stress events during period assessed

≥ 0 = age of most recent LIFE stress event before current depressive episode (if CURRENTLY depressed)
OR
age at most recent LIFE stress event before interview (if NOT currently depressed)

-9 = exposure unknown or age of exposure unknown</td><td>Age during most recent LIFE stress event <i>before</i> CURRENT major depressive episode
NOTE 1: LIFE stress is stress OTHER THAN childhood maltreatment
NOTE 2: If no current major depressive episode, list age of most recent LIFE stress before interview</td></tr>
<tr><td colspan="4">24-26: Additional timing variables for lifetime depression (create ONLY if one of variables 11 or 22 is missing)</td></tr>
<tr><td>24[4]</td><td>Stress_before_dep1</td><td>0 = First major depressive episode occurred before first LIFE stress event

1 = LIFE stress occurred before first depression,
    OR
    no LIFE stress during assessed period
    OR
    no lifetime depression diagnosis

-9 = unknown timing</td><td>Assessed LIFE stress occurred before first depressive episode.

<i>ONLY create this variable if</i>
variable 7 or variable 8 is present
<i>For reference</i>
<i>v7 = lifetime depression diagnosis</i>
<i>v8 = lifetime quant depression</i></td></tr>
</table>

	Variable	Coding	Description of variable
25[4]	Dep1_5yr_stress_exp	0 = Not exposed to LIFE stress in the 5 years prior to *first depression* (if ever depressed) OR *interview* (if no lifetime depression dx[5]) 1 = Exposed to LIFE stress in the 5 years prior to *first depression* (if ever depressed) OR *interview* (if no lifetime depression dx) -9 = unknown 5-year timing	LIFE stress before, but not more than 5 years prior to first depressive episode ***ONLY create this variable if*** variable 7 or variable 8 is present *For reference* *v7 = lifetime depression diagnosis* *v8 = lifetime quant depression*
26[4]	Dep1_5yr_stress_q	≥ 0 if assessed for the 5 years prior to *first depression* (if ever depressed) OR *interview* (if no lifetime depression) -9 = unknown 5-year timing	Quantitative measure of LIFE stress during the 5 years prior to first depressive episode ***ONLY create this variable if*** variable 7 or variable 8 is present *For reference* *v7 = lifetime depression diagnosis* *v8 = lifetime quant depression*
27-29: Additional timing variables for current depression (create ONLY if one of variables 12 or 23 is missing)			
27[4]	Stress_before_dep_curr	0 = Current depression occurred before first LIFE stress event 1 = LIFE stress occurred before current depression **OR** no LIFE stress during the period assessed **OR** no current depression diagnosis -9 = unknown (stress exposure unknown or timing unknown)	Assessed LIFE stress occurred before current depressive episode ***ONLY create this variable if*** variable 9 or variable 10 is present *For reference* *v9 = current depression diagnosis* *v10 = current quantitative depression*
28[4]	Dep_curr_5yr_stress_exp	0 = Not exposed to LIFE stress in the 5 years prior to *current depression* (if depressed) OR *interview* (if no current depression) 1 = Subject was exposed to LIFE stress in the 5 years prior to *current depression* (if depressed) OR *interview* (if no current depression) -9 = unknown (exposure unknown or 5 year timing unknown)	LIFE stress occurred before, but not more than 5 years prior to current depressive episode ***ONLY create this variable if*** variable 9 or variable 10 is present *For reference* *v9 = current depression diagnosis* *v10 = current quantitative depression*
29[4]	Dep_curr_5yr_stress_q	≥ 0 if assessed for the 5 years prior to *current depression* (if depressed) OR *interview* (if no current depression) -9 = unknown if quantitative LIFE stress assessed for the 5-year window	Quantitative measure of LIFE stress during the 5 years prior to current depressive episode ***ONLY create this variable if*** variable 7 or variable 8 is present *For reference* *v7 = lifetime depression diagnosis* *v8 = lifetime quant depression*

[1] This is so that R will not eliminate such subjects from appropriate analyses
[2] CTQ = Childhood Trauma Questionnaire
[3] Primary childhood maltreatment variable derived from the CTQ is based on these three variables
[4] Variables 24-29 ONLY apply to stress OTHER THAN childhood maltreatment (here called LIFE stress)
[5] dx = diagnosis