**Data Dictionary**

**File name: AMLdata\_WashU\_05oct17.xls**

**Created on: Thursday, October 5, 2017**

**Using: dataprep\_r5.sas**

**Programmer: Brenda Birmann**

**Variable Format Label† Value(s)/definition(s)**

pid i3 Fake ID Based on original sequence of participant in list order

coll i1 Collection 1=1st blood collection (NHS or HPFS), 2=2nd blood collection (NHS only)

aliasid a4 Alias ID (from key file)

chr a6 Chr Chromosome

start i10 Start Mutation start position

end i10 End Mutation end position

ref a20 Ref Reference (major) allele

alt a4 Alt Alternate (minor) allele(s)

gene a7 Gene Name of the gene the mutation is annotated to fall in. This includes introns/exons/UTRs

func a9 Func When the region of that gene is annotated as exonic, intronic, or a splice site.

ensgene a18 EnsGene Ensembl gene ID for that the annotations are based on.

ncenstrans a18 NC.ensTrans Ensembl transcript ID for the specific isoform that splicing or UTR variants fall within.

ncannot a18 NC.annot The effect of the splicing or UTR variant at the DNA level for the transcript listed.

exonicfunc a18 ExonicFunc The predicted AA change for exonic mutations (nonsynonymous, stopgain, etc…)

aaenstrans a18 AA.ensTrans The Ensembl transcript ID that corresponds to the predicted annotation for exonic mutations.

aaexon a8 AA.exon The exon number based on the Ensembl transcript ID for each coding mutation.

aacdna a18 AA.cDNA The coding DNA sequence change predicted for each exonic mutation given Ensembl transID

aacodon a8 AA.Codon The predicted AA change for each exonic mutation given Ensembl transcript ID.

cosmic77 a12 cosmic77 Catalog of Somatic Mutations in Cancer (COSMIC) ID for annotations previous seen in cancer.

\_1000g2015augall a12 1000g2015aug\_all The MAF for clonal mutations identified in the population based on the 1000 Genomes Project.

cadd13raw a10 CADD13\_Raw Predicted consequence of mutation based on CADD score (http://cadd.gs.washington.edu/)

cadd13phred a8 CADD13\_PHRED Equivalent Phred score for given CADD score (http://www.phrap.com/phred/).

vaf11 f7.4 VAF1.1 Variant allele fraction for 1st tech repeat of 1st collection sample

vaf12 f7.4 VAF1.2 Variant allele fraction for 2nd tech repeat of 1st collection sample

vaf21 f7.4 VAF2.1 Variant allele fraction for 1st tech repeat of 2nd collection sample (NHS only)

vaf22 f7.4 VAF2.2 Variant allele fraction for 2nd tech repeat of 2nd collection sample (NHS only)

quest a2 Questionable The “V” denotes additional samples validated after we sent you the data. “?” denotes the interesting sample 48 which had three splicing mutations in SRSF2 and CALR that were >0.4 VAF and should be filtered per the parameters as likely germline heterozygous mutations, but are actually likely true clonal mutations. And a single mutation in sample 24A, which is at VAF 0.34, but is actually probably a real heterozygous mutation because it falls in an intron.

avgvaf1 f5.3 Average of the vafs from tech repeats 1 and 2 for collection 1; BLANK if one or both sample 1 repeats were blank (i.e. if we consider sample 1 to be negative for this variant).

avgvaf2 f5.3 Average of the vafs from tech repeats 1 and 2 for collection 2; BLANK if one or both sample 2 repeats were blank (i.e. if we consider sample 2 to be negative for this variant).

avgany f5.3 Average of the non-zero vafs for collection 1 and 2; BLANK if neither sample was non-zero.

avgearly f5.3 Average of the vafs from the 2 technical repeats for the earliest collected sample in the study with a non-zero value for avgvaf for this variant (i.e. uses sample 1 if included in the study and tested positive in both tech repeats, or sample 2 if there was no sample 1 in the study or the corresponding sample 1 was zero for this variant); BLANK if both samples were negative for this variant.

avglater f5.3 Average of the vafs from the 2 technical repeats for the latest collected sample in the study with a non-zero value for avgvaf for this variant (i.e. uses sample 2 if included in the study and tested positive in both tech repeats, or sample 1 if there was no sample 2 in the study or the corresponding sample 2 was zero for this variant); BLANK if both samples were negative for this variant.

cohort i1 1=NHS, 2=HPFS

caco i1 1=case, 2=control

gender i1 1=female, 2=male

cohname a4 word version of source cohort (“NHS “ v. “HPFS”)

genderfm a1 Character version of gender variable: F, M

matchset i3 Matched set number‡ (<100=NHS; ≥100=HPFS)

age\_bld1 i2 age(y) at 1st bld Age (years) at first blood collection, rounded to nearest year (blank for NHS participants with only a second blood sample in the study)

age\_bld2 i2 age(y) at 2nd bld Age (years) at second blood collection, rounded to nearest year (blank for HPFS participants and for NHS participants with only a first blood sample in the study)

age\_casedx i2 age(y) at case dx Age (years) at date of case diagnosis (based on matched case diagnosis date for controls)

newcaco a7 word version of case-control status (“case ” v. “control”)

caco\_ab a1 A=case, B=control

idlabel a1 A=case, B=control #1, C=control #2

† Left blank if no label was used in SAS

‡ When data set was sorted (separately by cohort) by matched case ID number and case-control status.

**Data File Notes:**

All records from AliasID=19A and 34A were removed. Records for variants in *CALR* and *SRSF2* for AliasID=048B (N=3 records) were removed because one or more VAF values were ≥0.4. No additional records were found with one or more VAF values ≥0.4.

After the above exclusions the file was reduced from an original 611 to 598 records. Subsequent inclusion of one record each for three men in HPFS who did not test positive for any mutations (aliasIDs 49A, 53A and 83A) raised the file back to a total of 601 records.

For sample 18B, the VAF fields for each of the observed variants were recoded as second collection results (originally were labeled as first collection results), e.g. all vaf11s were renamed as vaf21s and all vaf12s were relabeled as vaf22s.