

Creating file for NeighborNet Treebuilding

Grace Saville

16/05/2021

1. Loading the data

```
data <- read.csv("./data/RStudio/ratsSNPs_clean.csv")  
dim(data) # 370 rows 298 columns
```

```
## [1] 370 298
```

2. Testing the dat2phylip() function

```
# x <- data.frame(a = c("asdfghjkl1", "asdf2ghjkl", "asdf3ghjkl", "asdfghjkl4"),  
# b = c("CTAGTGACCCGTAG", "TGACCCGTTAGAAC", "TGACTCTTTAGAAC", "TTTCACGTTGAGAC"))  
# dummy df  
  
# dat2phylip(x, outfile = "test.phy") # testing this function that saves phylip files
```

After some experimenting, for Phylip files...

- spaces in the base strings aren't necessary
- 10 character length names are needed
- dashes "-" are allowed
- no differentiation between upper and lower case
- only letters from bases or amino acids allowed (?, e.g acdefghiklmnpqrstvwyz)

3. Replacing the SNP symbols with IUPAC Ambiguity code symbols

IUPAC Ambiguity code:

Symbol	SNP bases
A	AA
T	TT
C	CC
G	GG

Symbol	SNP bases
R (purine)	AG
Y (pyrimidine)	CT
W (weak)	AT
S (strong)	CG
M (amino)	AC
K (keto)	GT

```
copy <- data # making a copy
kable(table(unlist(copy[, -c(1:16)])))
```

Var1	Freq
?	10571
A:A	15051
A:C	478
A:G	3246
A:T	410
C:C	24963
C:G	841
C:T	2618
G:G	30849
T:G	427
T:T	14886

```
copy[copy == "A:A"] <- "A"
copy[copy == "T:T"] <- "T"
copy[copy == "C:C"] <- "C"
copy[copy == "G:G"] <- "G"
copy[copy == "A:G"] <- "R"
copy[copy == "C:T"] <- "Y"
copy[copy == "A:T"] <- "W"
copy[copy == "C:G"] <- "S"
copy[copy == "A:C"] <- "M"
copy[copy == "T:G"] <- "K"
```

```
kable(table(unlist(copy[, -c(1:16)])))
```

Var1	Freq
?	10571
A	15051
C	24963
G	30849
K	427
M	478
R	3246
S	841
T	14886
W	410
Y	2618

4. Adjusting the data frame for Phylip output function

```
names(copy)
```

```
## [1] "island" "registration.number" "genus"
## [4] "species" "sex" "country"
## [7] "state_province" "island.1" "locality"
## [10] "site" "geo_lat" "geo_long"
## [13] "collector" "collecting.date" "field.number"
## [16] "Populatie" "X299_CHR1_114679736" "X13_CHR1_116614092"
## [19] "X14_CHR1_124857905" "X15_CHR1_134869867" "X16_CHR1_137314938"
## [22] "X18_CHR1_185979552" "X19_CHR1_192708191" "X20_CHR1_198383739"
## [25] "X21_CHR1_201964872" "X22_CHR1_209547552" "X23_CHR1_211919559"
## [28] "X262_CHR1_212322960" "X24_CHR1_216451585" "X25_CHR1_220057345"
## [31] "X26_CHR1_231126749" "X27_CHR1_255622475" "X300_CHR1_262011841"
## [34] "X301_CHR1_262011844" "X28_CHR1_265508390" "X29_CHR1_281235485"
## [37] "X260_CHR1_32362358" "X2_CHR1_34868240" "X3_CHR1_53720156"
## [40] "X296_CHR1_57745953" "X4_CHR1_57815638" "X5_CHR1_65742365"
## [43] "X6_CHR1_72363888" "X7_CHR1_80135877" "X297_CHR1_80533719"
## [46] "X298_CHR1_82579836" "X8_CHR1_85381970" "X9_CHR1_89039311"
## [49] "X10_CHR1_94071374" "X174_CHR10_100628723" "X175_CHR10_106721530"
## [52] "X157_CHR10_13158451" "X276_CHR10_15702323" "X158_CHR10_18222467"
## [55] "X159_CHR10_29319370" "X160_CHR10_38717342" "X277_CHR10_41408367"
## [58] "X278_CHR10_45900147" "X161_CHR10_46560158" "X163_CHR10_56443762"
## [61] "X279_CHR10_56614639" "X164_CHR10_59037416" "X165_CHR10_63678613"
## [64] "X166_CHR10_66786945" "X156_CHR10_6843845" "X167_CHR10_71822645"
## [67] "X168_CHR10_74081533" "X169_CHR10_77114932" "X170_CHR10_83332247"
## [70] "X171_CHR10_86303384" "X172_CHR10_91514023" "X173_CHR10_94603088"
## [73] "X176_CHR11_14441692" "X177_CHR11_22397127" "X178_CHR11_31463063"
## [76] "X179_CHR11_44808465" "X180_CHR11_53037550" "X181_CHR11_62063970"
## [79] "X183_CHR11_79227284" "X184_CHR11_81741813" "X185_CHR12_13111258"
## [82] "X186_CHR12_17180963" "X187_CHR12_22138482" "X188_CHR12_28111193"
## [85] "X311_CHR12_41236268" "X312_CHR12_41264415" "X190_CHR12_43084794"
## [88] "X191_CHR12_51249426" "X192_CHR13_50199444" "X284_CHR13_57006176"
## [91] "X193_CHR13_63129514" "X194_CHR13_75428433" "X195_CHR13_79469457"
## [94] "X196_CHR13_84181170" "X313_CHR13_91427136" "X198_CHR13_92726132"
## [97] "X285_CHR13_93714043" "X199_CHR13_95717062" "X208_CHR14_105431405"
## [100] "X209_CHR14_110278754" "X210_CHR14_113585203" "X201_CHR14_11536255"
## [103] "X202_CHR14_20823876" "X203_CHR14_42594513" "X200_CHR14_5088615"
## [106] "X204_CHR14_62715105" "X286_CHR14_71970877" "X205_CHR14_74155933"
## [109] "X206_CHR14_79443960" "X207_CHR14_86735968" "X222_CHR15_101663107"
## [112] "X223_CHR15_108810163" "X224_CHR15_111076172" "X211_CHR15_1633950"
## [115] "X212_CHR15_28658678" "X213_CHR15_34321957" "X215_CHR15_43580802"
## [118] "X217_CHR15_52430180" "X218_CHR15_77738259" "X219_CHR15_81392057"
## [121] "X220_CHR15_88731621" "X221_CHR15_96822565" "X225_CHR16_1603970"
## [124] "X226_CHR16_22347879" "X227_CHR16_37436064" "X228_CHR16_44068904"
## [127] "X229_CHR16_47717933" "X230_CHR16_69357902" "X287_CHR16_72228092"
## [130] "X234_CHR17_11396298" "X315_CHR17_18333453" "X232_CHR17_2009955"
## [133] "X288_CHR17_20165912" "X235_CHR17_20259016" "X236_CHR17_30683947"
## [136] "X237_CHR17_33086513" "X238_CHR17_37527057" "X239_CHR17_49956902"
## [139] "X233_CHR17_6670575" "X241_CHR17_76277887" "X290_CHR17_84174938"
## [142] "X317_CHR17_90440327" "X243_CHR18_24619428" "X291_CHR18_29561318"
```

## [145]	"X245_CHR18_32607221"	"X292_CHR18_37095675"	"X246_CHR18_41341746"
## [148]	"X293_CHR18_58447620"	"X242_CHR18_6474943"	"X247_CHR18_77747130"
## [151]	"X248_CHR18_81418040"	"X249_CHR19_1335965"	"X250_CHR19_16421590"
## [154]	"X251_CHR19_21394944"	"X252_CHR19_25766701"	"X253_CHR19_38113806"
## [157]	"X38_CHR2_117541471"	"X39_CHR2_121218337"	"X31_CHR2_12917122"
## [160]	"X40_CHR2_147793172"	"X41_CHR2_150754842"	"X43_CHR2_174928002"
## [163]	"X44_CHR2_178680587"	"X45_CHR2_185767298"	"X46_CHR2_188988798"
## [166]	"X266_CHR2_189602081"	"X47_CHR2_196126115"	"X48_CHR2_199336146"
## [169]	"X32_CHR2_20334241"	"X49_CHR2_210088383"	"X50_CHR2_221654675"
## [172]	"X51_CHR2_224367757"	"X302_CHR2_248289141"	"X52_CHR2_249800861"
## [175]	"X33_CHR2_25437163"	"X263_CHR2_29685824"	"X34_CHR2_39163286"
## [178]	"X35_CHR2_43068057"	"X36_CHR2_48863798"	"X37_CHR2_52531394"
## [181]	"X30_CHR2_8296972"	"X254_CHR20_19358518"	"X295_CHR20_20650236"
## [184]	"X255_CHR20_22824916"	"X256_CHR20_26801189"	"X258_CHR20_47292294"
## [187]	"X259_CHR20_55569799"	"X61_CHR3_103752695"	"X62_CHR3_107203912"
## [190]	"X63_CHR3_111699929"	"X64_CHR3_118186992"	"X267_CHR3_123119438"
## [193]	"X65_CHR3_123178700"	"X268_CHR3_130083043"	"X66_CHR3_142244207"
## [196]	"X67_CHR3_146704502"	"X68_CHR3_149033562"	"X304_CHR3_153450115"
## [199]	"X269_CHR3_160845101"	"X69_CHR3_161467231"	"X54_CHR3_22396329"
## [202]	"X55_CHR3_35931140"	"X56_CHR3_43121389"	"X57_CHR3_46846440"
## [205]	"X58_CHR3_58151966"	"X53_CHR3_6113698"	"X59_CHR3_61666827"
## [208]	"X60_CHR3_95847397"	"X303_CHR3_98089694"	"X76_CHR4_108302183"
## [211]	"X77_CHR4_115625977"	"X78_CHR4_121415856"	"X79_CHR4_151448626"
## [214]	"X305_CHR4_152925705"	"X80_CHR4_178759194"	"X306_CHR4_181117834"
## [217]	"X81_CHR4_183519538"	"X72_CHR4_40383606"	"X73_CHR4_56114004"
## [220]	"X74_CHR4_63012274"	"X70_CHR4_7204678"	"X75_CHR4_81949632"
## [223]	"X89_CHR5_101553257"	"X90_CHR5_108836834"	"X91_CHR5_112971787"
## [226]	"X92_CHR5_131389039"	"X93_CHR5_134940547"	"X272_CHR5_135658619"
## [229]	"X94_CHR5_148231291"	"X95_CHR5_151977774"	"X96_CHR5_165773835"
## [232]	"X82_CHR5_16842603"	"X97_CHR5_169293182"	"X84_CHR5_33163922"
## [235]	"X85_CHR5_36463932"	"X86_CHR5_50002491"	"X87_CHR5_58018936"
## [238]	"X88_CHR5_64849757"	"X111_CHR6_107657404"	"X112_CHR6_110666683"
## [241]	"X308_CHR6_12527061"	"X113_CHR6_133889824"	"X99_CHR6_14055789"
## [244]	"X114_CHR6_146773076"	"X100_CHR6_22398928"	"X101_CHR6_28733150"
## [247]	"X102_CHR6_36270438"	"X103_CHR6_48130972"	"X104_CHR6_58168931"
## [250]	"X105_CHR6_61442912"	"X106_CHR6_67084810"	"X107_CHR6_70874449"
## [253]	"X108_CHR6_74178666"	"X109_CHR6_80827771"	"X110_CHR6_91961599"
## [256]	"X123_CHR7_105399751"	"X124_CHR7_110179324"	"X125_CHR7_116865918"
## [259]	"X126_CHR7_122091528"	"X309_CHR7_12639745"	"X115_CHR7_12667929"
## [262]	"X127_CHR7_137830770"	"X128_CHR7_140342721"	"X129_CHR7_144614527"
## [265]	"X116_CHR7_48960243"	"X117_CHR7_58371316"	"X118_CHR7_70578025"
## [268]	"X119_CHR7_73663555"	"X120_CHR7_79272051"	"X121_CHR7_90755489"
## [271]	"X142_CHR8_103189899"	"X143_CHR8_109479154"	"X275_CHR8_117307690"
## [274]	"X144_CHR8_119566148"	"X145_CHR8_132167298"	"X130_CHR8_36538451"
## [277]	"X131_CHR8_46002216"	"X273_CHR8_48588797"	"X132_CHR8_49021356"
## [280]	"X133_CHR8_56570758"	"X274_CHR8_62930894"	"X134_CHR8_63637034"
## [283]	"X135_CHR8_66425804"	"X136_CHR8_68815138"	"X137_CHR8_75418426"
## [286]	"X138_CHR8_81061024"	"X139_CHR8_89233698"	"X140_CHR8_94192380"
## [289]	"X141_CHR8_98839713"	"X155_CHR9_101269288"	"X147_CHR9_16472226"
## [292]	"X146_CHR9_2345960"	"X148_CHR9_25096861"	"X149_CHR9_41768036"
## [295]	"X150_CHR9_49497106"	"X152_CHR9_70081740"	"X153_CHR9_82647453"
## [298]	"X154_CHR9_95290356"		

```
copy <- copy[,-c(2:16)] # removing the rows with information other than the  
# species key and codes  
copy <- unite(copy, bases, -1, sep = "", remove = TRUE) # merging all the base  
# columns into one
```

```
dat2phylip(copy, outfile = "./data/SplitsTree/ratsSNPs_SplitsTree_input.phy")  
# can be opened by splitstree when configured as proteins/amino acids
```