Genome Informatics assignment 3: SRY

303034908

Abstract

In humans, female sex is the outcome of undisturbed sex determination. The expression of the SRY (sex-determining region Y) mammalian gene, instead, induces male sex. It does so by encoding a transcription factor, the testes-determining factor (TDF), which binds to and increases the expression of a paralog autosomal gene, SOX9 (Sry-related HMG box 9). This binding is mediated by a highly conserved HMG (high mobility group) box domain. Mutations within this HMG domain often lead to gonadal dysgenesis. The mechanisms of sex determination are highly variable in the animal kingdom. Phylogeny is, therefore, a useful method for understanding the evolutionary basis of structure and function of the human SRY gene.

Introduction

The human SRY gene was first identified by searching for conserved sequences among translocated Y chromosomal DNA from XX male patients (Sinclair et al. 1990). In 1993, a 887 nucleotide long intron-less gene was found.

The expression of the SRY gene is both temporally and spatially specific. It is localized to bi-potential gonadal cells during embryogenesis. These are cells that differentiate into ovaries or testes, indirectly determining secondary sexual development as well as primary. As shown in figure 1, from Kashimada and Koopman 2010, its expression is also very short lived. Once it is able to increase the expression of SOX9 to an appropriate threshold, a negative feedback loop blocks further SRY transcription.

The exact mechanism of function of the SRY protein was discovered gradually. Sekido 2010, showed, using mutation, co-transfection, sex reversal studies and CHiP analysis, that TDF requires synergistic action with the transcription factor SF1 to bind to enhancer element TESCO of the autosomal SOX9 gene and increase its expression in Sertoli cells.

This binding is mediated by the HMG domain, is sequence and location specific, and induces a 60-85°bend in the DNA once bound (V R Harley and Goodfellow 1994). Both DNA-binding and bending are essential for SRY function. Analysis of human SRY protein from XY females shows that the mutations are often in the HMG domain, reflecting the importance of this domain for the function of SRY.

While sex determination pathways are fairly conserved across the animal kingdom, what triggers them is vastly different. A phylogenetic overview of the SRY gene, therefore, may be a useful insight to explain the evolutionary origin for its current structure and function.

Methods

Functional annotation and homology

Nucleotide and protein sequences for *Homo sapiens* were downloaded from Ensembl and Uniprot. Sequences were run through BLAST and BLASTp on online PHAMMER against NCBI, UniProt, Swissprot and Pfam databases. PHAMMER BLAST was repeated with separately only the HMG box domain sequence, the N-terminal domain of the human SRY and the C-terminal domain, giving similarity results for each domain across homologues.

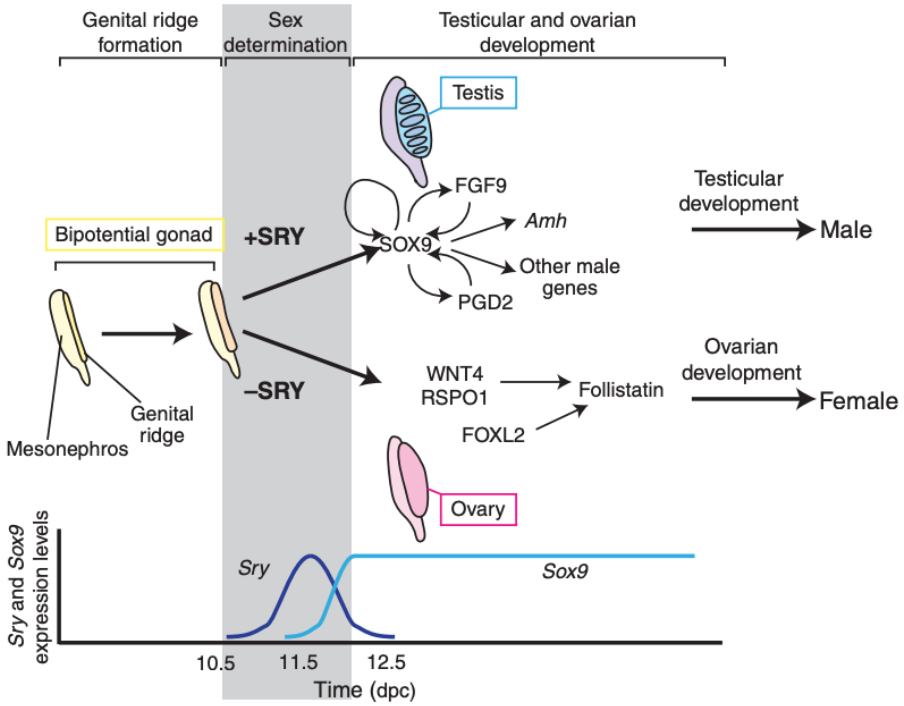


Figure 1: Overview of sex determination in mice. Taken from Kashimada and Koopman 2010

Variations and mutation

Variation data was downloaded from Gnomad, dbSNP, NCBI and Ensembl. They all contained variant IDs, and Ensembl and Gnomad contained further information such as allele changes, mutation effects and predicted clinical outcomes.

Ten SNPs from Ensembl were inputted into the Variant Effect Predictor (VEP) online in the format: chromosome, start, end, allele, strand. These ten SNPs were chosen as they were already annotated by Ensembl, and had a mixture of pathogenic and benign outcomes, thus providing a chance for comparison both of the method of prediction and the outcome.

Phylogeny and evolution

The top 200 unque BLASTp hits were inputted for phylogenetic tree reconstruction via multiple sequence alignment (MUSCLE), curation (GBlocks), phylogeny (PhyML), and tree rendering (TreeDyn). This pipeline was run via Phylogeny.fr online (A Dereeper et al. 2008, Alexis Dereeper et al. 2010). This phylogenetic tree can be compared to both the one produced by Ensembl/NCBI as well as one of the most recent papers on evolution of the SRY gene, by Katsura et al. 2018. Parts of these phylogenetic trees are shown in figures 7, 8, and 9.

Results and Discussion

Functional annotation and homology

Alignment with PHAMMER of the human SRY protein sequence to the Pfam database showed one functional domain: the HMG box, $\sim$ 70 amino acids long, present near the middle of the 204 amino acid long protein, as shown in figure 2.

After aligning the HMG box domain with PHAMMER to the proteomes databases, the top 6 unique hits other than human were primate proteins (bitscores from 455 to 409, e-values between e^{-137} and e^{-119}) . The next inflection point occurred at a bitscore of 130. All hits above that threshold were within the phylum Chordata, and included animals such as

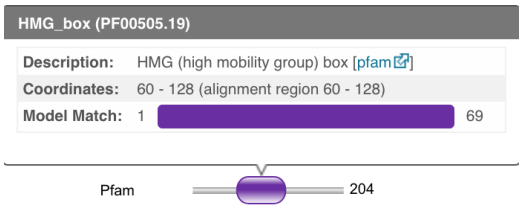


Figure 2: HMG box domain in human SRY protein

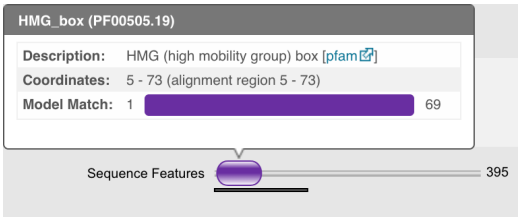
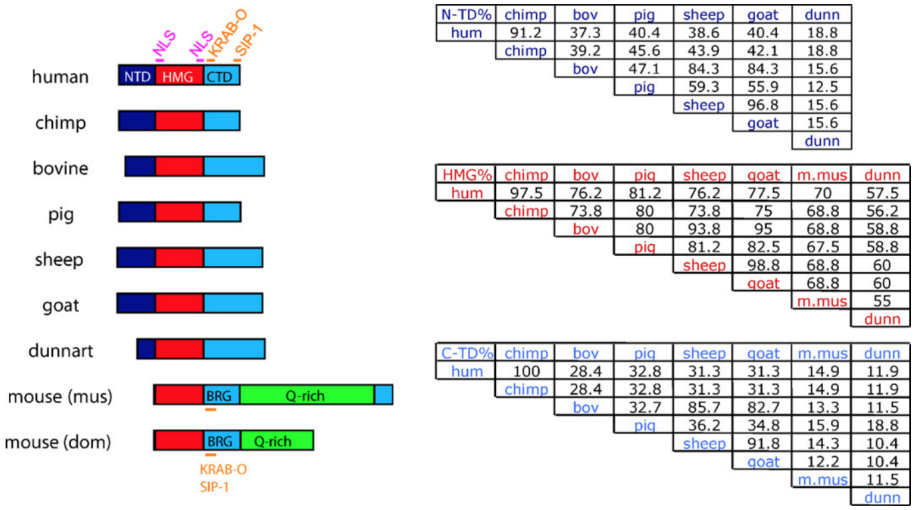


Figure 3: HMG box domain in mouse SRY protein



N-TD%	chimp	bov	pig	sheep	goat	dunn
hum	91.2	37.3	40.4	38.6	40.4	18.8
	chimp	39.2	45.6	43.9	42.1	18.8
		bov	47.1	84.3	84.3	15.6
			pig	59.3	55.9	12.5
				sheep	96.8	15.6
					goat	15.6
						dunn

HMG%	chimp	bov	pig	sheep	goat	m.mus	dunn
hum	97.5	76.2	81.2	76.2	77.5	70	57.5
	chimp	73.8	80	73.8	75	68.8	56.2
		bov	80	93.8	95	68.8	58.8
			pig	81.2	82.5	67.5	58.8
				sheep	98.8	68.8	60
					goat	68.8	60
						m.mus	55
							dunn

C-TD%	chimp	bov	pig	sheep	goat	m.mus	dunn
hum	100	28.4	32.8	31.3	31.3	14.9	11.9
	chimp	28.4	32.8	31.3	31.3	14.9	11.9
		bov	32.7	85.7	82.7	13.3	11.5
			pig	36.2	34.8	15.9	18.8
				sheep	91.8	14.3	10.4
					goat	12.2	10.4
						m.mus	11.5
							dunn

Figure 4: taken from Sekido 2010. "Schematic SRY protein structures and domain comparisons between various mammalian species. SRY is composed of three major domains; the N-terminal domain (N-TD, dark blue), HMG domain (HMG, red) and C-terminal domain (C-TD, light blue). Mouse SRY contains the bridge domain (BRG) and a glutamine-rich domain (Q-rich, green) instead of usual C-TD, but lacks N-TD. The pink and orange bars represent nuclear localization signal (NLS) and SIP-1/KRAB-O interacting domains, respectively. Amino acid sequence similarities of each domain are shown in tables."

dog, cat, sheep, pig, horse, goat, bovine, and mouse. For most of these species, the HMG box aligned around the same location as in the human SRY protein. However, for the mouse, the HMG box was located at the beginning of the 395-amino-acid-long protein, as shown in figure 3.

The human SRY protein can be further subdivided into its N-terminal domain (NTD) on the left of the HMG box and the C-terminal domain (CDT) on the right according to the diagram in figure 2. Rerunning alignment over each of these sections shows that only the HMG domain is conserved across species. The top 7 unique alignments of the human SRY CTD using PHAMMER had bitscores between 161 and 130, with e-values between e^{-35} to e^{-46}. The top 7 hits when aligning the NTD also showed bitscores between 119 and 99, with e-values between e^{-32} and e^{-26}. In both cases, the top 7 unique hits were all primate species. A full pairwise species to species comparison between the three parts of the human SRY protein was run by Sekido 2010. This is shown in figure 4.

Variations and mutation

The results of the VEP analysis are shown in figure 5. A representation of these effects is shown in figure 6, where the HMG box is coloured in blue, the pathogenic mutations are highlighted in pink and the benign mutations are highlighted in green. The results are the same as the Ensembl annotation. Two out of the three mutations within the HMG box domain were classified as pathogenic, whilst only one out of the six benign SNPs was within the HMG box domain.

Uploaded_variation	Consequence	Protein_position	Amino_acids	Existing_variation	SIFT	PolyPhen
Y_2787015_G/A	missense_variant	197	R/C	rs756606002	tolerated(0.3)	benign(0)
Y_2787044_T/C	missense_variant	187	N/S	rs780561417	tolerated(0.22)	benign(0.021)
Y_2787053_G/A	missense_variant	184	P/L	rs1019354171	tolerated(0.17)	benign(0.001)
Y_2787060_G/A	missense_variant	182	H/Y	rs1194771063	tolerated(1)	benign(0.147)
Y_2787113_T/C	missense_variant	164	Y/C	rs748958243	tolerated(0.06)	benign(0.003)
Y_2787136_T/G	missense_variant	156	E/D	rs754623064	tolerated(0.49)	benign(0.006)
Y_2787207_G/A	missense_variant	133	R/W	rs104894976	deleterious(0)	probably_damaging(0.987)
Y_2787221_T/C	missense_variant	128	K/R	rs375342012	deleterious(0.02)	possibly_damaging(0.81)
Y_2787224_T/A	missense_variant	127	Y/F	rs104894973	deleterious(0.01)	probably_damaging(0.999)
Y_2787240_C/T	missense_variant	122	E/K	rs771449441	tolerated(0.79)	benign(0.274)

Figure 5: Results from Variant Effect Predictor analysis

MKFRWDCYLSCFNSDDYSPAVQENIPALRRSSSFLCTESCNSKYQCETGENSKGNVQDRVKRPM
NAFIVWSRDQRRKMALENPRMRNSEISKQLGYQWKMLTEAEKWPFFQEAQKLQAMHREKYPN
FRYRPRWKAKMLPKNCSLLPADPASVLCSDVQLDNRLCRDDCTKATHSRMEHQLGYLLPISAASS
PQQRDCYSHWTKL

Figure 6: Example of human SRY with SNPs annotated by colour: pink highlight if pathogenic, green highlight if benign, and blue letters for the HMG box domain

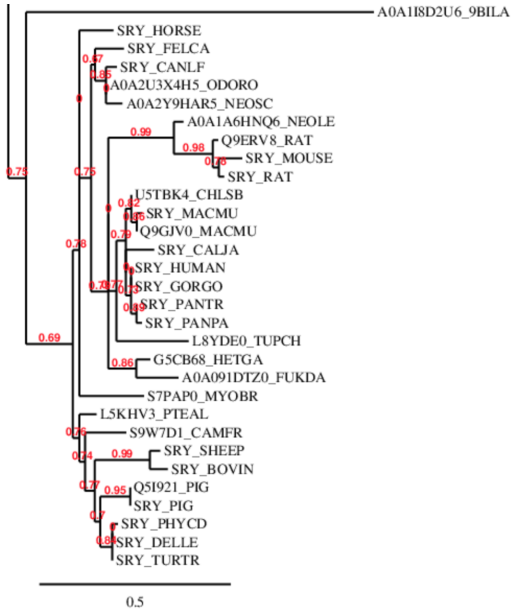


Figure 7: Partial SRY phylogenetic tree from Phylogeny.fr, A Dereeper et al. 2008

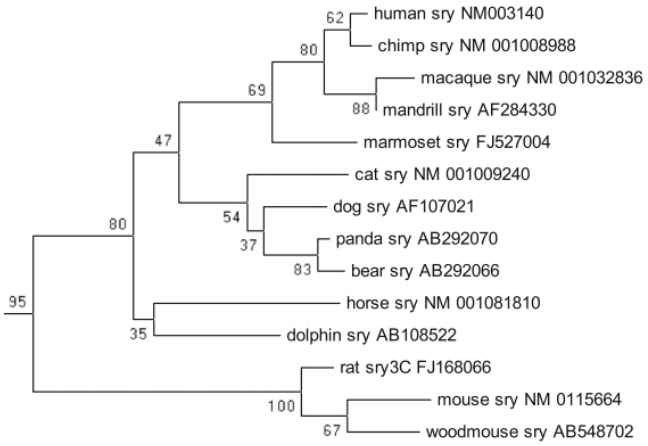


Figure 8: Partial SRY phylogenetic tree from Katsura et al. 2018

Phylogeny and evolution

The phylogenetic trees show some common features. Primates (gorilla, chimpanzee, etc.) are the closest to the human SRY. In both the phylogeny.fr tree and the Ensembl tree, the mouse species are closer to the human than the group of mammals including sheep, bovine, pig, horse, cat, dog. This is not the case in the phylogenetic tree from Katsura et al. 2018, which instead shows the mouse SRY as of the first diversions of the tree. This makes more more sense, as we have seen above that the mouse Sry sequence is structurally more different than for example the bovine to the human. These differences in results may be due to the algorithm used to build these trees.

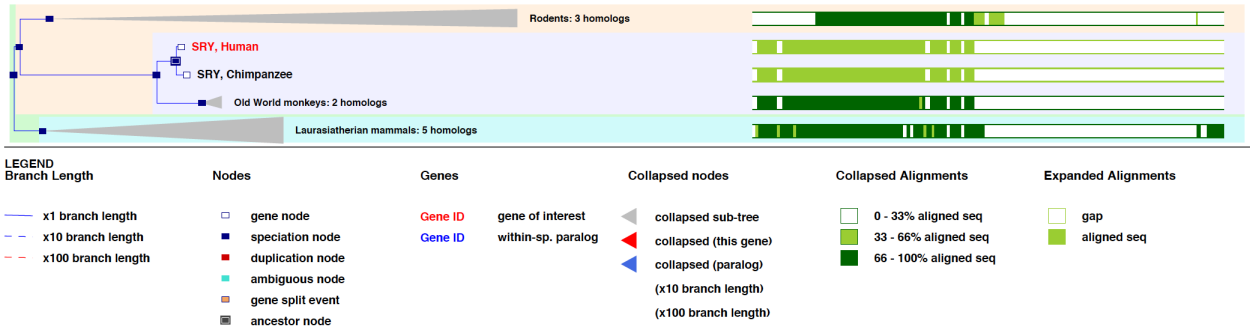


Figure 9: SRY phylogenetic tree from Ensembl

Discussion

Functional annotation and homology

The HMG-box domain is the most conserved aspect of the human SRY protein. It is conserved both across SRY proteins in different species and SOX proteins within humans. The N- and the C- terminal domains instead are highly variable. The question then becomes if the N- and C- termini have any role in the function of the protein. Mouse Sry is different from other mammals, as it lacks an NTD and containing an unusual C terminus comprising a bridge domain and a polyglutamine (polyQ) tract. By trancation, immunoflurescence and expression studies, Zhao et al. 2014, showed that the polyQ domain not only stabilizes mouse Sry protein, but more importantly, functions as a transactivation domain essential to activate Sox9 transcription and effect male sex determination in vivo.

Mutations and SNPs

Gonadal dysgenesis rarely occurs if mutations are outside the HMG box domain. Work by V R Harley and Goodfellow 1994 and Mitchell and Vincent R Harley 2002 showed through biochemical studies that mutations such as Y127F and K128R in the table above affect directly the binding ability of the HMG box as well as the bending of the DNA once bound. Both are key aspects for successful function of the SRY protein.

Phylogeny and evolution

Whilst not shown in the phylogenetic trees above, the most similar autologous gene to SRY is SOX3. This gene is located on the X chromosome and is a key gene for brain development. However, it was shown that ectopic expression of SOX3 in pregonadal cells in transgenic XX mice also leads to male development, suggesting that SRY evolved by truncation and promoter relocation of the SOX3 gene (Sutton et al. 2011). The chromosome which included the newly created SRY gene became the Y chromosome. As explained by Graves 2006, selection of male-specific alleles on the Y chromosome led to repression of recombination, and gradually to loss of all active genes other than those with a male advantage, such as SRY. This high rate of loss and mutation may explain the high variability observed outside of the HMG box domain even in closely related mammals.

Conclusion

The SRY gene is a recent evolutionary addition, and yet an incredibly important one in the initiation of sex differentiation in mammals. Mutations in its HMG box domain lead to gonadal dysgenesis and sex reversal. Alignment of homologous sequences show that the HMG box domain is the only very highly conserved region of the gene. Whilst important, in some cases, such as the mouse, this domain is necessary but not sufficient for correct gene function, reiterating its varied molecular function in the animal kingdom.

References

[1] A H Sinclair et al. “A gene from the human sex-determining region encodes a protein with homology to a conserved DNA-binding motif”. en. In: *Nature* 346.6281 (July 1990), pp. 240–244.

[2] V R Harley and P N Goodfellow. “The biochemical role of SRY in sex determination”. en. In: *Mol. Reprod. Dev.* 39.2 (Oct. 1994), pp. 184–193.

[3] Claire L Mitchell and Vincent R Harley. “Biochemical defects in eight SRY missense mutations causing XY gonadal dysgenesis”. In: *Mol. Genet. Metab.* 77.3 (Nov. 2002), pp. 217–225.

[4] Jennifer A Marshall Graves. “Sex chromosome specialization and degeneration in mammals”. en. In: *Cell* 124.5 (Mar. 2006), pp. 901–914.

[5] A Dereeper et al. “Phylogeny.fr: robust phylogenetic analysis for the non-specialist”. en. In: *Nucleic Acids Res.* 36.Web Server issue (July 2008), W465–9.

[6] Alexis Dereeper et al. “BLAST-EXPLORER helps you building datasets for phylogenetic analysis”. en. In: *BMC Evol. Biol.* 10 (Jan. 2010), p. 8.

[7] Kenichi Kashimada and Peter Koopman. “Sry: the master switch in mammalian sex determination”. en. In: *Development* 137.23 (Dec. 2010), pp. 3921–3930.

[8] Ryohei Sekido. “SRY: A transcriptional activator of mammalian testis determination”. en. In: *Int. J. Biochem. Cell Biol.* 42.3 (Mar. 2010), pp. 417–420.

[9] Edwina Sutton et al. “Identification of SOX3 as an XX male sex reversal gene in mice and humans”. en. In: *J. Clin. Invest.* 121.1 (Jan. 2011), pp. 328–341.

[10] Liang Zhao et al. “Structure-function analysis of mouse Sry reveals dual essential roles of the C-terminal polyglutamine tract in sex determination”. en. In: *Proc. Natl. Acad. Sci. U. S. A.* 111.32 (Aug. 2014), pp. 11768–11773.

[11] Yukako Katsura et al. “The evolutionary process of mammalian sex determination genes focusing on marsupial SRYs”. en. In: *BMC Evol. Biol.* 18.1 (Jan. 2018), p. 3.

```
#
    ##############################################################################

#                                    Functional domains
#
    ##############################################################################


setwd("~/Desktop/code/CompBio MPhil/GenomeInformatics/GIa3/Function and
    sequences/Function/")


human.sry.ppt <- ">SRY-201 peptide: ENSP00000372547 pep:protein_coding
MQSYASAMLSVFNSDDYSPAVQENIPALRRSSSFLCTESCNSKYQCETGENSKGNVQDRVKRPMNAFIVWSRDQRRKMALENPRMRNSEISK
    "

human.sry.ppt.only <- strsplit(human.sry.ppt, split="\n")[[1]][2]

human.sry.ppt.NTD <- substr(human.sry.ppt.only, 0, 59)
human.sry.ppt.CTD <- substr(human.sry.ppt.only, 129, 204)

write.table(human.sry.ppt.NTD, "human.sry.ppt.NTD.txt",
            row.names = FALSE, quote = FALSE, col.names = FALSE)
write.table(human.sry.ppt.CTD, "human.sry.ppt.CTD.txt",
            row.names = FALSE, quote = FALSE, col.names = FALSE)



#### remove duplicated entries on full fasta sequences
setwd("~/Desktop/code/CompBio MPhil/GenomeInformatics/GIa3/
      Function and sequences/Phylogeny/")

library("seqinr")
homologues.fasta <- read.fasta("sryblastfullfasta")
hits <- names(homologues.fasta)
length(hits)
length(unique(hits))

#select top 200 fasta sequences
twohundred.homologues.fasta <- homologues.fasta[1:200]
length(unique(names(twohundred.homologues.fasta))) #check if duplicates
#save each fasta as 1 string
names(twohundred.homologues.fasta) <- paste0(">", names(twohundred.homologues.
    fasta))
for (name in names(twohundred.homologues.fasta)) {
  twohundred.homologues.fasta[[name]] <- toupper(paste(
    twohundred.homologues.fasta[[name]], collapse=""))
  twohundred.homologues.fasta[[name]] <- paste(
    name, twohundred.homologues.fasta[[name]], sep ="\n" )
}

twohundred.homologues.fasta <- unlist(twohundred.homologues.fasta)

write.table(twohundred.homologues.fasta, "twohundred.homologues.fasta.txt",sep="
    \n",
            row.names = FALSE, col.names = FALSE, quote = FALSE)


#
    ##############################################################################
```

```
#                                    Variation
#
    ##############################################################################


setwd("~/Desktop/code/CompBio MPhil/GenomeInformatics/GIa3/Variation")

#from https://www.ncbi.nlm.nih.gov/gene/6736
sry.start.GRCh37 <- 2654896
sry.start.GRCh38 <- 2786855
HMGbox.start <- 60*3 #starts at 60th AA * 3 nt per amino acid
HMGbox.end <- 128*3
SRY.length <- 204*3

#downloaded from ftp://ftp.ncbi.nlm.nih.gov/snp/organisms/human_9606_b151_
    GRCh38p7/chr_rpts/
#chr.Y <- read.delim("chr_Y.txt", header=TRUE, stringsAsFactors = FALSE)
#sry.snps <- chr.Y[grepl('SRY', chr.Y$local), ]
#sry.snps.select <- sry.snps[, c(12, 17, 24, 25)]
#sry.snps.select$chr.2 <- as.numeric(sry.snps.select$chr.2) - sry.start.GRCh38
#sry.snps.select <- sry.snps.select[sry.snps.select$vali. >0, ]

#downloaded from http://gnomad.broadinstitute.org/gene/ENSG00000184895
# gnomad data: GRCh37, not GRCh38...
gnomad.sry.snps <- read.csv("gnomAD_v2.1_ENSG00000184895_2018_12_18_21_58_03.csv
    ",
                            header=TRUE)
gnomad.sry.snps$Position <- gnomad.sry.snps$Position - sry.start.GRCh37

#polyphen compares to GRCh37
polyphen.input <- gnomad.sry.snps$rsID
write.table(polyphen.input,"polyphen.input.txt",sep="\n",
            row.names=FALSE, quote = FALSE)


#ensembl data: GRCh38
#downloaded from https://www.ensembl.org/Homo_sapiens/Gene/Variation_Gene/...
#    Table?db=core;g=ENSG00000184895;r=Y:2786855-2787699;t=ENST00000383070
ensembl.sry.snp <- read.csv("ensembl-gene-variations.csv", header = TRUE)
ensembl.sry.snp <- ensembl.sry.snp[grepl("SNP", ensembl.sry.snp$Class), ]
ensembl.sry.snp <- ensembl.sry.snp[order(ensembl.sry.snp$sift_sort,
                                         decreasing = FALSE), ]
ensembl.sry.snp <- ensembl.sry.snp[grepl('missense', ensembl.sry.snp$Conseq..
    Type), ]
ensembl.sry.snp$Location <- gsub("Y:", "", ensembl.sry.snp$Location)

#VEP compares to GRCh38
vep.snps <- cbind("Y", ensembl.sry.snp$Location, ensembl.sry.snp$Location,
                  as.character(ensembl.sry.snp$Alleles), "+")
write.table(vep.snps,"VEP.input.snps.txt",sep="\t",row.names=FALSE,
            quote = FALSE, col.names = FALSE)

ensembl.sry.snp$Location <- as.numeric(ensembl.sry.snp$Location) - sry.start.
    GRCh38


# VEP results

vep.results <- read.delim("VEPresults.txt", header=TRUE)
```

```
vep.results.select <- vep.results[grep("protein_coding", vep.results$BIOTYPE), ]
vep.results.select <- vep.results.select[,c(1,4,17,18, 20, 28, 29, 47)]

write.csv(vep.results.select, "vep.results.select.csv", quote=FALSE,
          col.names = FALSE, row.names = FALSE)


#convert correct PPT sequence to mutated one
locations <- vep.results.select$Protein_position
alleles <- vep.results.select$Amino_acids
for ( i in 1:length(locations)){
  substr(human.sry.ppt.only, locations[i], locations[i]) <- strsplit(as.
    character(alleles),
                                                                      split="/")
    [[i]][2]
}

substr(human.sry.ppt.only, 155, 157 )



#
    ##############################################################################

#                                       Expression
#
    ##############################################################################

setwd("~/Desktop/code/CompBio MPhil/GenomeInformatics/GIa3/Expression/")

#downloaded from Encode
rna.tissues <- read.delim("rna_tissue.tsv")
rna.tissues.sry <- rna.tissues[grep("SRY", rna.tissues$Gene.name), ]
rna.tissues.sry <- rna.tissues.sry[order(rna.tissues.sry$Value, decreasing =
    TRUE),]

library("ggplot2")
ggplot(data=rna.tissues.sry, aes(x = reorder(Sample, -Value), y = Value, fill =
    Gene.name)) +
  geom_bar(stat = "identity") +
  theme(axis.text.x = element_text(angle = 60, hjust = 1))
```