VDJdb summary statistics

```
df <- read.table("../database/vdjdb.slim.txt", header=T, sep="\t")
```

```
library(plyr)
library(knitr)

summarize_complexes <- function(ddf, column) {
  ddply(ddf, column, summarize, records = length(complex.id), paired.records = sum(ifelse(complex.id=="
}

kable(format = "latex", summarize_complexes(df, c("species","gene")))
```

species	gene	records	paired.records
HomoSapiens	TRA	337	228
HomoSapiens	TRB	2026	228
MacacaMulatta	TRA	74	0
MacacaMulatta	TRB	943	0
MusMusculus	TRA	16	16
MusMusculus	TRB	17	17

```
kable(format = "latex", summarize_complexes(subset(df, mhc.class=="MHCI"), c("species", "mhc.a")))
```

species	mhc.a	records	paired.records
HomoSapiens	HLA-A*01:01	141	2
HomoSapiens	HLA-A*02	180	84
HomoSapiens	HLA-A*02:01	676	64
HomoSapiens	HLA-A*02:01,HLA-A*02	13	0
HomoSapiens	HLA-A*02:01,HLA-A*02:01:48,HLA-A*02	1	1
HomoSapiens	HLA-A*02:01:110	2	2
HomoSapiens	HLA-A*02:01:48	44	44
HomoSapiens	HLA-A*02:01:48,HLA-A*02:01:59	2	2
HomoSapiens	HLA-A*02:01:59	7	7
HomoSapiens	HLA-A*02:01:59,HLA-A*02:01:48	1	1
HomoSapiens	HLA-A*02:01:98	4	4
HomoSapiens	HLA-A*02:256	2	2
HomoSapiens	HLA-A*03	19	0
HomoSapiens	HLA-A*11:01	42	0
HomoSapiens	HLA-A*24:02:84	6	6
HomoSapiens	HLA-B*07	3	3
HomoSapiens	HLA-B*07:02	285	67
HomoSapiens	HLA-B*07:02,HLA-B*07	1	1
HomoSapiens	HLA-B*08	155	44
HomoSapiens	HLA-B*08:01	254	3
HomoSapiens	HLA-B*08:01,HLA-B*08:01:29	2	2
HomoSapiens	HLA-B*08:01:29	6	6
HomoSapiens	HLA-B*15	34	0
HomoSapiens	HLA-B*18	9	0
HomoSapiens	HLA-B*27	16	0
HomoSapiens	HLA-B*27:05:31	4	4
HomoSapiens	HLA-B*35:01	154	22
HomoSapiens	HLA-B*35:01,HLA-B*35:08,HLA-B*35:42:01,HLA-B*35:08:01	1	1
HomoSapiens	HLA-B*35:01,HLA-B*35:42:01	2	2
HomoSapiens	HLA-B*35:01,HLA-B*35:42:01,HLA-B*35:08:01	1	1
HomoSapiens	HLA-B*35:08	50	25
HomoSapiens	HLA-B*35:08,HLA-B*35:08:01	4	4
HomoSapiens	HLA-B*35:08:01	3	3
HomoSapiens	HLA-B*35:42:01	2	2
HomoSapiens	HLA-B*44:05	2	2
HomoSapiens	HLA-B*44:05:01	6	6
HomoSapiens	HLA-B*51:193	2	2
HomoSapiens	HLA-B*57	57	0
HomoSapiens	HLA-B*57:01	25	0
HomoSapiens	HLA-B*57:06	2	2
HomoSapiens	HLA-E*01:01:01:03	2	2
MacacaMulatta	Mamu-A*01	1017	0
MusMusculus	48425589	14	14
MusMusculus	48425619	2	2

```
kable(format = "latex", summarize_complexes(df, c("antigen.gene", "antigen.species")))
```

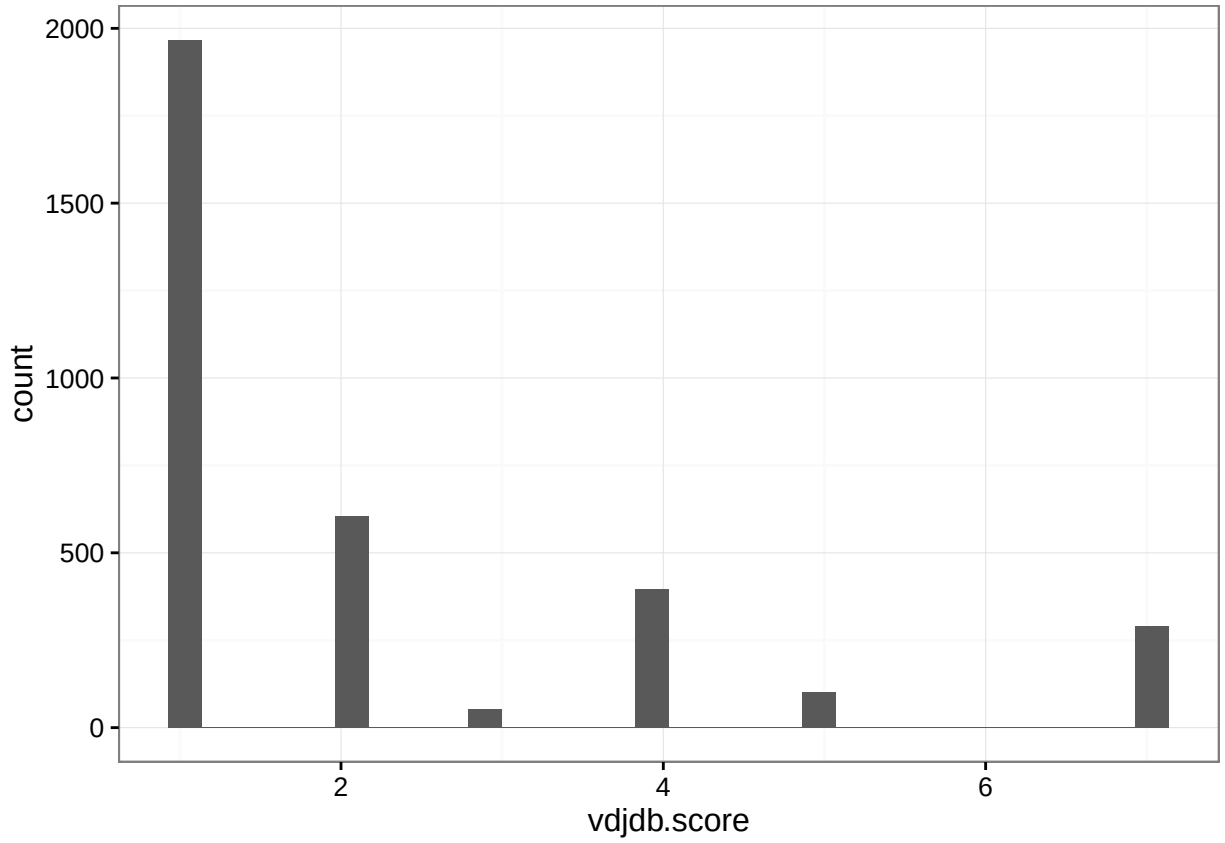
antigen.gene	antigen.species	records	paired.records
ABCD3	HomoSapiens	2	2
ALPHA-GLIADIN	TriticumAestivum	16	16
BMLF1	EBV	243	2
BRLF1	EBV	71	0
BZLF1	EBV	254	43
CTAG1B	HomoSapiens	6	6
EBNA1	EBV	54	13
EBNA3	EBV	8	8
EBNA3a	EBV	27	0
EBNA3A	EBV	65	3
EBNA3b	EBV	15	0
EBNA3B	EBV	52	0
EBNA6	EBV	2	2
ELAVL4	HomoSapiens	2	2
ERBB2	HomoSapiens	2	2
Gag	HIV-1	137	8
Gag	SIV	419	0
Gtpbp1	MusMusculus	2	2
HA	InfluenzaA	113	5
IE1	CMV	54	18
INS	HomoSapiens	6	6
Kctd20	MusMusculus	2	2
LYZ	GallusGallus	2	2
M1	InfluenzaA	3	3
Mbp	MusMusculus	5	5
MBP	HomoSapiens	2	2
MCC	HomoSapiens	6	6
MCC	ManducaSexta	2	2
MFI2	GallusGallus	2	2
MLANA	HomoSapiens	23	23
N	VSIV	2	2
Ndufa4	MusMusculus	2	2
Nef	HIV-1	128	32
ns3	HCV	117	72
NS3	HCV	161	0
p17	CMV	17	0
p24	CMV	72	0
p65	CMV	566	74
PA	InfluenzaA	2	2
Pol	HIV-1	29	0
POLG	HCV	2	2
Rbm5	MusMusculus	2	2
synthetic	synthetic	4	4
Tat	SIV	598	0
TAX	HTLV-1	17	17
Tel1	SaccharomycesCerevisiae	2	2
TERT	HomoSapiens	2	2
TPI	HomoSapiens	6	6
UL40	CMV	2	2
UL83	CMV	6	6
VP22	HSV-2	69	67
WT1	HomoSapiens	10	10

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.2.4
```

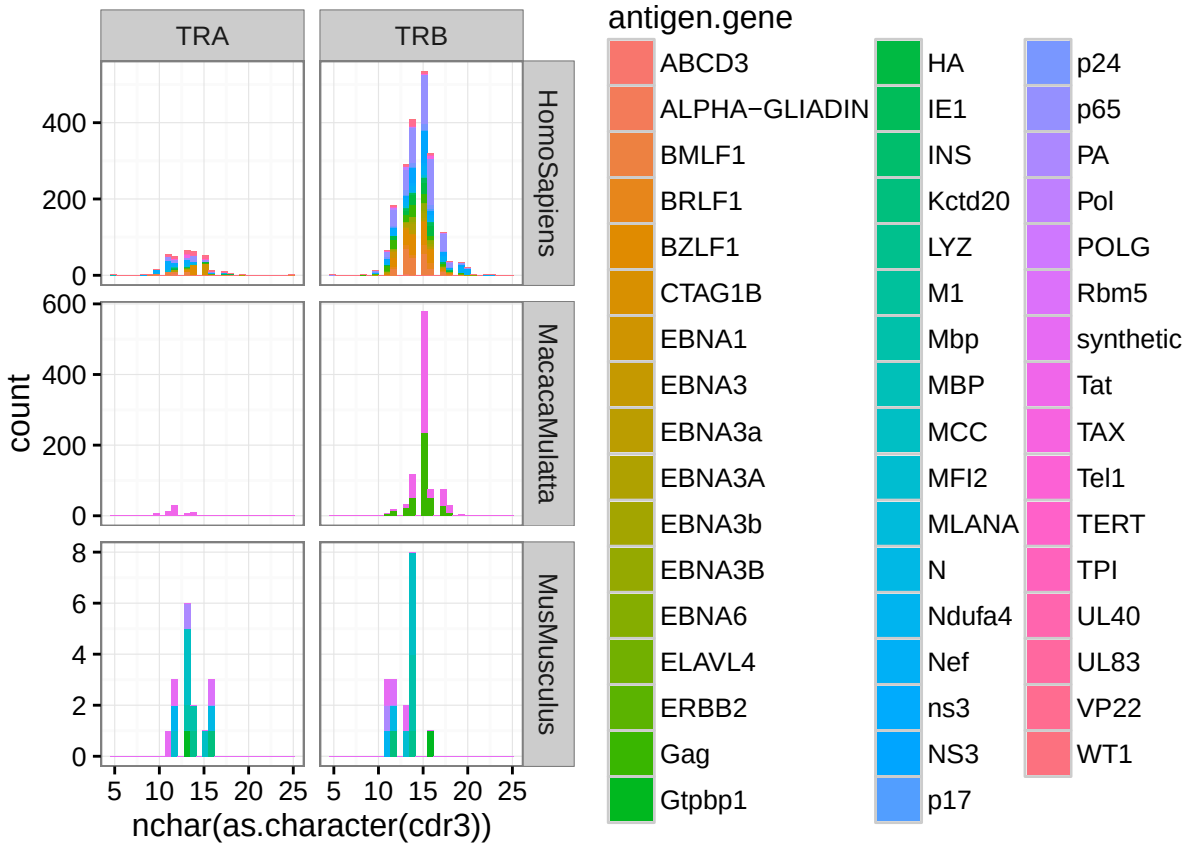
```
ggplot(df, aes(x=vdjdb.score)) + geom_histogram() + theme_bw()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
ggplot(df, aes(x=nchar(as.character(cdr3)), fill = antigen.gene)) +
  geom_histogram() +
  facet_grid(species~gene, scales="free_y") + theme_bw()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```


TRA
TRB
HomoSapiens
MacacaMulatta
MusMusculus
count
nchar(as.character(cdr3))
antigen.gene
ABCD3
ALPHA−GLIADIN
BMLF1
BRLF1
BZLF1
CTAG1B
EBNA1
EBNA3
EBNA3a
EBNA3A
EBNA3b
EBNA3B
EBNA6
ELAVL4
ERBB2
Gag
Gtpbp1
HA
IE1
INS
Kctd20
LYZ
M1
Mbp
MBP
MCC
MFI2
MLANA
N
Ndufa4
Nef
ns3
NS3
p17
p24
p65
PA
Pol
POLG
Rbm5
synthetic
Tat
TAX
Tel1
TERT
TPI
UL40
UL83
VP22
WT1