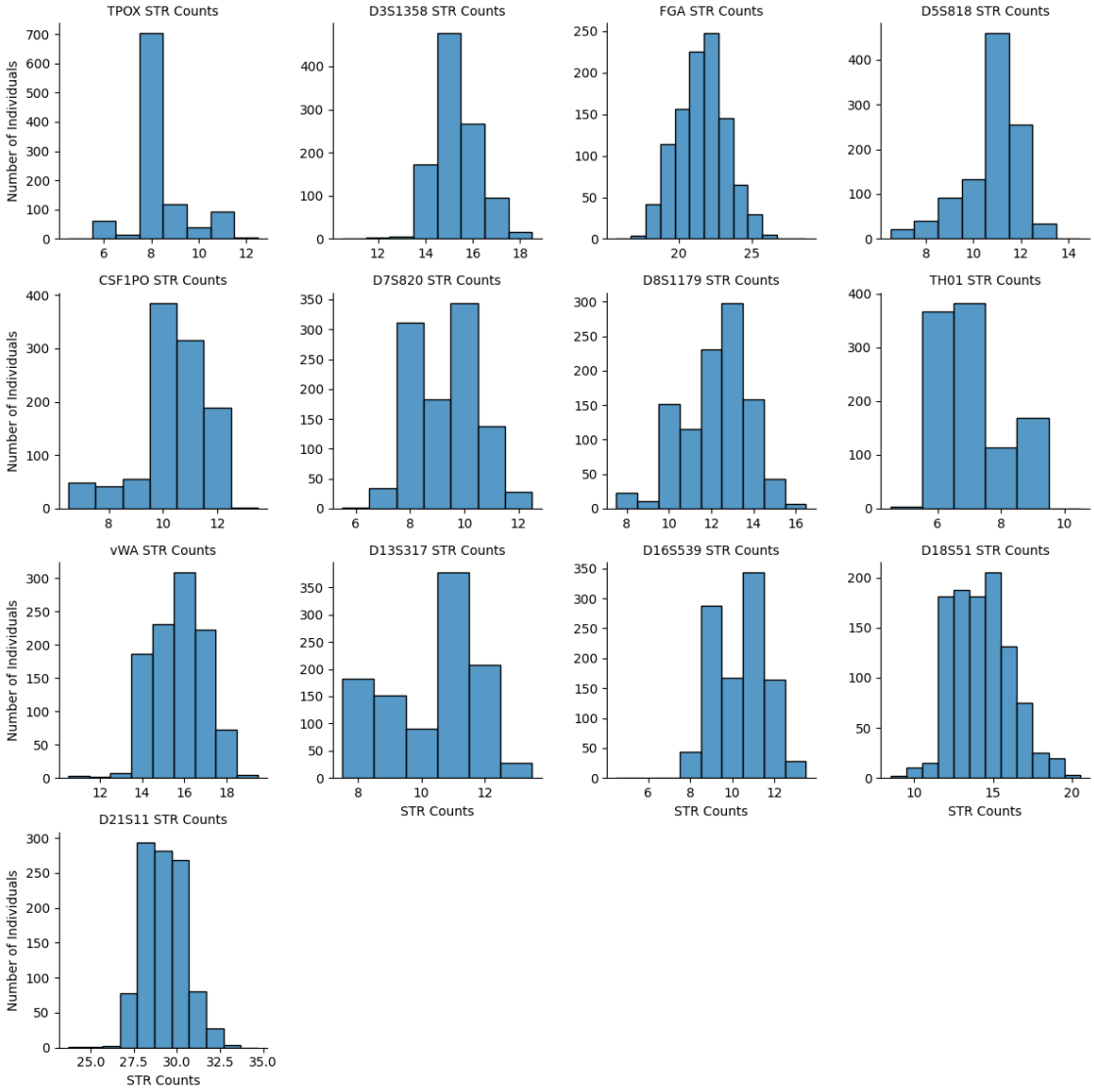
QUESTION 2: Fill in the code below to create separate bar charts for each CODIS locus, including TPOX, to get a sense of the allelic variation and how concentrated the distributions might be. Make sure each bar chart has an appropriate title and axis labels. >Hint: You can use f-strings to substitute variables within strings. >Hint: You can use sns.FacetGrid to create subplots.

In [8]:
```
# Reshape data into long format
df_long = CODIS_profiles.melt(var_name="Locus", value_name="STR Count")

# Create FacetGrid for subplots
g = sns.FacetGrid(df_long, col="Locus", col_wrap=4, sharex=False, sharey=False)

# Map histograms to each facet
g.map_dataframe(sns.histplot, x="STR Count", discrete=True)

# Adjust titles
g.set_titles(col_template="{col_name} STR Counts")

# Adjust labels
g.set_axis_labels("STR Counts", "Number of Individuals")
```

Out[8]:
```
<seaborn.axisgrid.FacetGrid at 0x7fa233391190>
```


TPOX STR Counts
D3S1358 STR Counts
FGA STR Counts
D5S818 STR Counts
CSF1PO STR Counts
D7S820 STR Counts
D8S1179 STR Counts
TH01 STR Counts
vWA STR Counts
D13S317 STR Counts
D16S539 STR Counts
D18S51 STR Counts
D21S11 STR Counts
Number of Individuals
STR Counts

QUESTION 3c: List at least two ways in which the probabilities you calculated above are not fully reflective of reality.

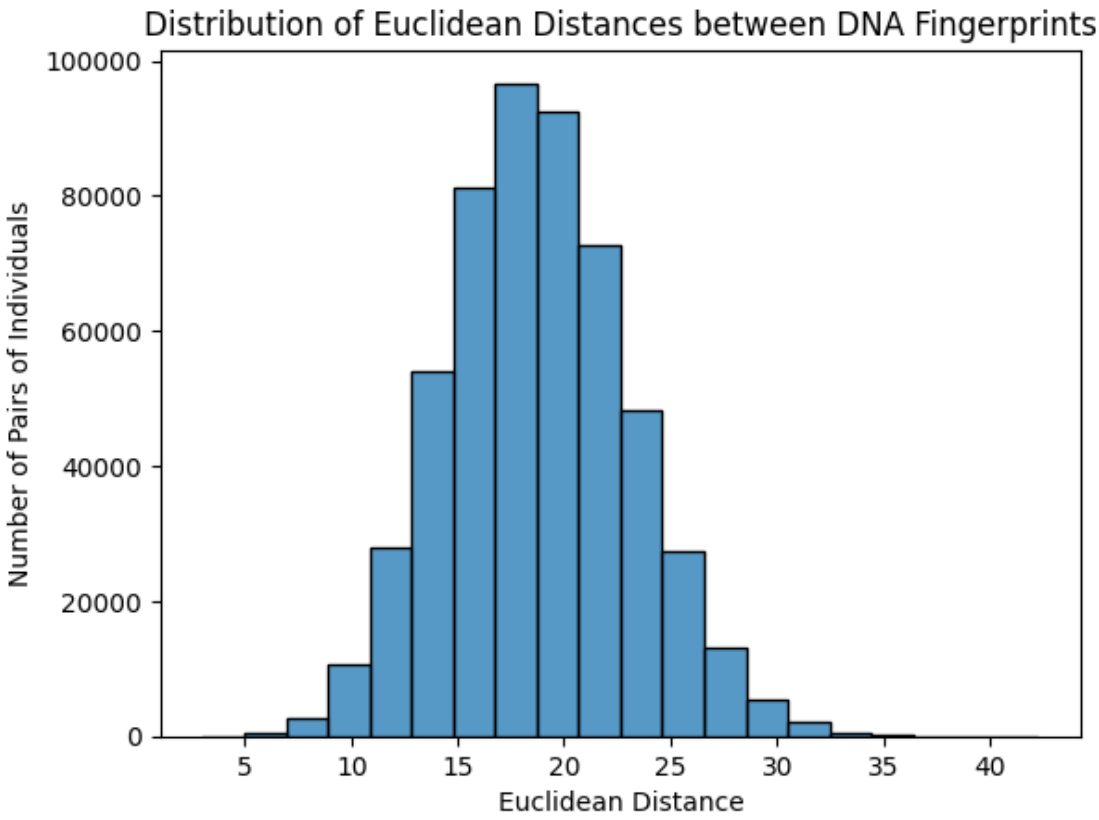
ANSWER:

1. We are assuming independence between loci. The assumption that the STR locus is indpendent is an approximation as in reality, there could be weak genetic correlation among loci due to population substructure. For example, if a subpopulation has a distinct ancestry, certain alleles may be more prevalent in that group of people relative to the general population.
2. As we talked about in class, we also have to consider the sampling bias in CODIS Data. Since the CODIS allele frequencies are derived from specific populations samples, it might not reflect the genetic diversity of an entire population. This would lead to underrepresentation or even overrepresentation of specific ethnic groups

QUESTION 5b: Display a bar chart showing the distribution of the pairwise distances. Set the number of bins to 20, and include an appropriate title and axes labels.

In [28]:
```
#Problem 5b
ax = sns.histplot(distances, bins=20)
ax.set_title("Distribution of Euclidean Distances between DNA Fingerprints")
ax.set_xlabel("Euclidean Distance")
ax.set_ylabel("Number of Pairs of Individuals")
```

Out[28]:
```
Text(0, 0.5, 'Number of Pairs of Individuals')
```



QUESTION 8: Create an overlaid bar chart that shows the STR count distribution of both TPOX and TPOX.1. Make sure to set an appropriate binwidth, title, legend, and axes labels.

```
In [37]: ax = sns.histplot(diploid_profiles[["TPOX", "TPOX.1"]], binwidth=20, discrete=True)
         ax.set_title("STR count distributution TPOX and TPOX.1")
         ax.set_xlabel("Allele")
         ax.set_ylabel("Str Count")
```

```
Out[37]: Text(0, 0.5, 'Str Count')
```