Severe Hemophilia A (Factor VIII): A Statistical Analysis Report on CDC Genetic Variant Data

Tai Chou-Kudu

Abstract

This study examines genetic variants associated with severe Hemophilia A (Factor VIII deficiency) using data from the CDC's CHAMP database. Statistical analyses revealed significant associations between variant types and mechanisms ($p < 0.001$), with missense substitutions and frameshift deletions being the most common. Coding regions and mixed exon-intron regions were frequently linked to severe cases. Genomic coordinates showed broader distributions in severe cases ($p < 2.2e\text{-}16$). Additional tests confirmed associations with specific gene domains, subtypes, and codons ($p < 0.0005$).

Notably, mixed exon-intron regions were exclusively linked to severe cases, suggesting a distinct functional role. Poly-A regions were frequently observed in severe cases. Additionally, synonymous variants, often associated with minimal functional impact, appeared in several severe cases, indicating a potential area for further exploration. These findings align with existing genetic research on Hemophilia A, while also highlighting underresearched areas for further investigation.

Research Question

What are the most frequently reported causative genetic variants and their genomic locations associated with severe Hemophilia A (Factor VIII deficiency) in this dataset?

This analysis is based on reported unique genetic variants from the dataset, which is structured to highlight the diversity and characteristics of causative variants rather than provide a representative sample of the Hemophilia population.

Background Information

Hemophilia is a bleeding disorder in which the blood does not clot properly because it is missing clotting factors. The condition is inherited genetically, or more rarely acquired. Whether inherited or acquired, a genetic mutation leads to this factor deficiency. People with severe Hemophilia A (less than 1% of clotting factor VIII baseline, compared to normal baselines of 50-150%) must take weekly medicine in order to provide their bodies with the needed factor replacement. This helps prevent bruises and internal bleeding in joints, along with serious and life-threatening internal bleeding.

Cases

What are the cases, and how many are there?

2539 of 4038 records exist for people with severe Hemophilia A (compared to other clinical severity levels). We must note that this dataset excludes those who are "female or had more than one X chromosome or copy of F8 present" as they stated it "prevent[s] clear determination of the variant phenotype".

Data Collection

Describe the method of data collection.

"The original database was developed to support the Hemophilia Inhibitor Research Study (HIRS) at the CDC, which enrolled more than 1,000 people with hemophilia, to allow accurate reporting and record-keeping…The first CHAMP Mutation List was posted online at the CDC.gov website in 2011…The database was compiled from existing literature reports and databases to include the first identifiable report of each novel F8 variant reported to cause hemophilia A…It was compiled from mutations listed originally in the Haemophilia A Mutation, Structure, Test and Resource Site (HAMSTeRS), as well as those from more than 350 additional publications…They are listed as Year 0 in the current database if the date of initial publication is not known. At the time of download, HAMSTeRS included 943 unique variants. The initial CHAMP Mutation List included 2,537 unique variants."

This database is an activity of the Division of Blood Disorders and Public Health Genomics in the National Center on Birth Defects and Developmental Disabilities of the Centers for Disease Control and Prevention.

Type of Study

What type of study is this (observational/experiment)?

This study is observational.

Data Source

CDC Hemophilia Mutation Projects (CHAMP and CHBMP) (linked)

"The CHAMP F8 (factor VIII [8]) mutation list is an Excel database containing more than 4,000 changes in the F8 gene that have been reported to cause hemophilia A...Each mutation has been reviewed and uniquely identified using the Human Genome Variation Society nomenclature for DNA and predicted protein changes, as well as using traditional nomenclature based on the mature processed protein."

It's essential to note that in this dataset: "Each entry represents a single report of the given variant. Multiple reports of each variant are not collected, unless they occurred in the same calendar year, in which case they are incorporated into a single listing as the first report. The CDC Variant Lists therefore do not include a group of people with the same variant and cannot be assumed to be representative of all people with the reported variant. Their use for phenotypic analysis is therefore limited."

HGVS cDNA	DNA change based on cDNA reference sequence NM_000132.4, available at *https://www.ncbi.nlm.nih.gov/nuccore/NM_000132* , in accordance with recommendation of the Human Genome Variation Society (HGVS) at *http://www.hgvs.org/mutnomen/.*
hg 19 Coordinates	Genomic location of variant on the X chromosome according to the hg19 build of Genome Assembly GRCh37 at *https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_000001405.13/.*
HGVS Protein	Protein change based on protein reference sequence NP_000123.1, in accordance with recommendations of the Human Genome Variation Society (HGVS) at *http://www.hgvs.org/mutnomen/.*
Mature Protein	Protein change based on mature processed protein sequence, as used in early databases and publications, referred to as the legacy or traditional sequence .
Variant Type	Type of protein change caused by reported variant.
Mechanism	Type of DNA change caused by the reported variant.
Exon	The exon of *F8* where the variant is located.
Codon	The codon of *F8* where the variant is located, based on the mature protein sequence.
Domain	Domain location of the reported variant based on domains as defined in Saenko et al. *Vox Sanguinis* 2002; 83:89-96.
Subtypes	More detailed information regarding changes to *F8* caused by the reported variant.
In Poly A	Indicates whether or not the variant occurs within a Poly A run.

Figure 1: Field Definitions from CHAMP Data Workbook

Response Variable

What is the response variable, and what type is it (numerical/categorical)?

The response variable is: Reported Clinical Severity and it is categorical.

Explanatory Variable/s

What is the explanatory variable(s), and what type is it (numerical/categorical)?

Multiple explanatory variables exist: HGVS cDNA (categorical), hg 19 Coordinates (numerical), HGVS Protein (categorical), Mature Protein (categorical), Variant Type (categorical), Mechanism (categorical), Exon (categorical), Codon (categorical), Domain (categorical), Subtype (categorical), and in Poly A (categorical). Field definitions are included in the dataset. We also have Year Reported, which is a numerical variable.

We're not including these columns: severe, moderate, mild, no fvIII, history of inhibitor, comments, reference number, newly added in this current version. We'll use the Reported Clinical Severity column.

Data Preparation

```
#Load necessary libraries
library(tidyr)
library(dplyr)
library(ggplot2)
library(readr)
library(readxl)
library(janitor)
library(stringr)
library(skimr)
library(purrr)
library(forcats)
library(car)
```

```
#Load data and check column types
variants <- readxl::read_excel(here::here('CHAMP_Variant_List_2022.xlsx'),
                               sheet = 2, guess_max = 4000)
```

All columns have been loaded as characters, and we'll be performing statistical analyses, so we'll need to convert some columns to factors and numeric data types.

```r
# Clean data: Apply clean_names to the dataset to convert data to snake case
variants <- variants %>%
  clean_names()

# Select relevant columns
variants_clean <- variants %>%
  select(
    hgvs_c_dna, hg19_coordinates, hgvs_protein, mature_protein, variant_type,
    mechanism, exon, codon, domain, subtype, in_poly_a,
    reported_clinical_severity, year_reported
  )

# Convert explanatory variables to factors (categorical)
variants_clean <- variants_clean %>%
  mutate(across(
    c(
    variant_type, mechanism, exon, domain, subtype,
    in_poly_a, reported_clinical_severity
    ), as.factor
  ))

# Make sure that `year_reported` is numeric
variants_clean <- variants_clean %>%
  mutate(year_reported = as.numeric(year_reported))

skim(variants_clean)
```

Table 1: Data summary

Name	variants_clean
Number of rows	4050
Number of columns	13
_____	
Column type frequency:	
character	5
factor	7
numeric	1
_____	
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
hgvs_c_dna	12	1.00	6	59	0	4038	0
hg19_coordinates	22	0.99	9	45	0	3072	0
hgvs_protein	613	0.85	5	38	0	3238	0
mature_protein	602	0.85	1	34	0	3248	0
codon	274	0.93	1	18	0	1783	0

Variable type: factor

skim_variable	n_missing	complete_rate	ordered	n_unique	top_counts
variant_type	12	1.00	FALSE	10	Mis: 1803, Fra: 1052, Non: 460, Spl: 351
mechanism	12	1.00	FALSE	14	Sub: 2586, Del: 995, Dup: 277, Ins: 84
exon	12	1.00	FALSE	198	14: 861, 13: 190, 4: 176, 7: 161
domain	493	0.88	FALSE	13	A1: 750, A2: 727, B: 694, A3: 655
subtype	372	0.91	FALSE	8	Hea: 2064, Lig: 1087, Hea: 164, Lig: 132
in_poly_a	42	0.99	FALSE	3	N: 3831, Y: 176, n: 1
reported_clinical_severity	12	1.00	FALSE	12	Sev: 2353, Mil: 772, Mod: 521, Not: 252

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
year_reported	12	1	1907.17	442.32	0	2004	2010	2016	2022	

Exploratory Data Analysis across all severity levels

We'll explore the broader dataset before we investigate severe Hemophilia, by calculating some frequency tables and proportions.

```
variants_clean %>%
  count(hgvs_c_dna) %>%
  mutate(proportion = n / sum(n)) %>%
  arrange(desc(n))
```

```
variants_clean %>%
  count(hg19_coordinates) %>%
  mutate(proportion = n / sum(n)) %>%
  arrange(desc(n))
```

```
variants_clean %>%
  filter(!is.na(hgvs_protein)) %>%
  count(hgvs_protein) %>%
  mutate(proportion = n / sum(n)) %>%
  arrange(desc(n))
```

I noticed that the highest number of data points for hgvs_protein are p.(=), n = 30. I'll investigate further by checking frequency of mature_protein data.

```
variants_clean %>%
  filter(!is.na(mature_protein)) %>%
  count(mature_protein) %>%
  mutate(proportion = n / sum(n)) %>%
  arrange(desc(n))
```

The value of the most frequent mature_protein is "=". Through further background research I've learned that synonymous variant type (synonymous mutations), mean that the DNA changes but doesn't result in a change to the amino acid sequence of the protein. Often, it's believed that synonymous mutations are silent and do not cause diseases, but Inaba writes "recent studies have demonstrated that synonymous substitutions are not always silent." Inaba discusses the novel and rare variant, p.(Leu40=)/c.120C>A, which is actually included in this dataset as one of thirty synonymous variants. Inaba suggests a synonymous variant genotype is linked to mild clinical severity phenotype, but the CHAMP dataset shows 8 severe individuals and 12 mild individuals with a synonymous variant.

Information on Hemophilia B and synonymous variants seemed more accessible than articles on Hemophilia A and synonymous variants. Perhaps synonymous variants in Hemophilia A are underresearched, understandably, as it points to an even rarer subsect of a rare population.

```
variant_type_count <- variants_clean %>%
  count(variant_type) %>%
  mutate(proportion = n / sum(n)) %>%
  arrange(desc(n))
```

```r
variants_clean <- variants_clean %>%
  mutate(mechanism = str_to_lower(mechanism))
mechanism_count <- variants_clean %>%
  count(mechanism) %>%
  mutate(proportion = n / sum(n)) %>%
  arrange(desc(n))
```

```r
grouped_variant_mechanism_count <- variants_clean %>%
  count(variant_type, mechanism) %>%
  mutate(proportion = n / sum(n)) %>%
  arrange(desc(n))
```

```r
variants_clean %>%
  count(exon) %>%
  mutate(proportion = n / sum(n)) %>%
  arrange(desc(n))
```

```r
variants_clean %>%
  count(codon) %>%
  arrange(desc(n))
```

```r
variants_clean %>%
  count(domain) %>%
  mutate(proportion = n / sum(n)) %>%
  arrange(desc(n))
```

I'm noting that we do not want to standardize case to clean data for Domain, because A1 is shown as different from a1 on the "figures" page of the original dataset workbook. Capitalization makes a difference in the category here.

```r
#Subtype can be cleaned through case standardization
variants_clean <- variants_clean %>%
  mutate(subtype = str_to_lower(subtype))
variants_clean %>%
  count(subtype) %>%
  mutate(proportion = n / sum(n)) %>%
  arrange(desc(n))
```

```r
variants_clean <- variants_clean %>%
  mutate(in_poly_a = str_to_lower(in_poly_a))
variants_clean %>%
  count(in_poly_a) %>%
  mutate(proportion = n / sum(n)) %>%
  arrange(desc(n))
```

```r
variants_clean <- variants_clean %>%
  mutate(reported_clinical_severity = str_to_lower(reported_clinical_severity))
variants_clean %>%
  count(reported_clinical_severity) %>%
  mutate(proportion = n / sum(n)) %>%
  arrange(desc(n))
```

Reported clinical severity shows severe, mild, moderate, not reported, and several categories that show a mix of severities, divided by a / symbol. That could be a data entry issue or uncertainty due to lack of information in a publication. One could explore this more, but we have 2,359 severe cases, which is 58% of the total dataset, and can thus exclude the 72 mixed severe severity cases to simplify analysis.

```r
variants_clean %>%
  count(year_reported) %>%
  mutate(proportion = n / sum(n)) %>%
  arrange(desc(n))
```

```r
#top 5 frequency for each variable
purrr::map_dfr(variants_clean, ~as.data.frame
               (head(sort(table(.), decreasing = TRUE), 5)),
               .id = "variable")
```

```
                   variable                                 . Freq
1                  hgvs_c_dna                       c.-112G>A    1
2                  hgvs_c_dna                 c.-113_-134dupins    1
3                  hgvs_c_dna               c.-1171-?_1271+?del    1
4                  hgvs_c_dna                c.-1171-?_143+?del    1
5                  hgvs_c_dna       c.-1171-?_143+?delins263kb    1
6           hg19_coordinates                         154065972    4
7           hg19_coordinates                         154088865    4
8           hg19_coordinates                         154128172    4
9           hg19_coordinates                         154132663    4
10          hg19_coordinates                         154133280    4
```

11	hgvs_protein	p.(=)	30
12	hgvs_protein	p.(Leu1223*)	4
13	hgvs_protein	p.(His1234Glnfs*2)	3
14	hgvs_protein	p.(Met681Ile)	3
15	hgvs_protein	p.(Phe672del)	3
16	mature_protein	=	30
17	mature_protein	Leu1204*	4
18	mature_protein	His1215Glnfs*2	3
19	mature_protein	Met662Ile	3
20	mature_protein	Phe51*	3
21	variant_type	Missense	1803
22	variant_type	Frameshift	1052
23	variant_type	Nonsense	460
24	variant_type	Splice site change	351
25	variant_type	Large structural change (>50 bp)	236
26	mechanism	substitution	2586
27	mechanism	deletion	997
28	mechanism	duplication	278
29	mechanism	insertion	84
30	mechanism	deletion/insertion	81
31	exon	14	861
32	exon	13	190
33	exon	4	176
34	exon	7	161
35	exon	11	157
36	codon	425	9
37	codon	542	9
38	codon	167	8
39	codon	2229	8
40	codon	282	8
41	domain	A1	750
42	domain	A2	727
43	domain	B	694
44	domain	A3	655
45	domain	C2	304
46	subtype	heavy chain	2228
47	subtype	light chain	1219
48	subtype	multiple domains	126
49	subtype	single domain	104
50	subtype	165	1
51	in_poly_a	n	3832
52	in_poly_a	y	176
53	reported_clinical_severity	severe	2359

```
54 reported_clinical_severity                          mild  772
55 reported_clinical_severity                      moderate  521
56 reported_clinical_severity                  not reported  260
57 reported_clinical_severity                moderate/severe   64
58              year_reported                          2008  406
59              year_reported                          2012  340
60              year_reported                          2018  314
61              year_reported                          2016  237
62              year_reported                          2013  232
```

Exploratory Data Analysis for severe Hemophilia variants

We've gotten a general sense of the dataset. Now, let's start exploring severe Hemophilia by filtering for that clinical severity only.

```
#filter for severe clinical severity only
severe_hemo <- variants_clean %>%
  filter(reported_clinical_severity == "severe")

# Convert explanatory variables to factors (categorical)
severe_hemo <- severe_hemo %>%
  mutate(across(
    c(variant_type, mechanism, exon, domain, subtype, in_poly_a,
    reported_clinical_severity
    ), as.factor
  ))
```

```
#Check for nulls, always an important step to examine (and later, handle)
#null data before performing statistical analysis
skim(severe_hemo)
```

Table 5: Data summary

Name	severe_hemo
Number of rows	2359
Number of columns	13
_____	
Column type frequency:	
character	5
factor	7
numeric	1

Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
hgvs_c_dna	0	1.00	6	59	0	2359	0
hg19_coordinates	5	1.00	9	45	0	2012	0
hgvs_protein	421	0.82	5	28	0	1845	0
mature_protein	411	0.83	1	26	0	1854	0
codon	214	0.91	1	14	0	1332	0

Variable type: factor

skim_variable	n_missing	complete_rate	ordered	n_unique	top_counts
variant_type	0	1.00	FALSE	9	Fra: 861, Mis: 608, Non: 403, Spl: 229
mechanism	0	1.00	FALSE	10	sub: 1201, del: 801, dup: 219, ins: 71
exon	0	1.00	FALSE	179	14: 655, 13: 86, 4: 85, 16: 79
domain	336	0.86	FALSE	12	B: 563, A1: 401, A2: 352, A3: 328
subtype	228	0.90	FALSE	4	hea: 1335, lig: 606, mul: 107, sin: 83
in_poly_a	18	0.99	FALSE	2	n: 2203, y: 138
reported_clinical_severity	0	1.00	FALSE	1	sev: 2359

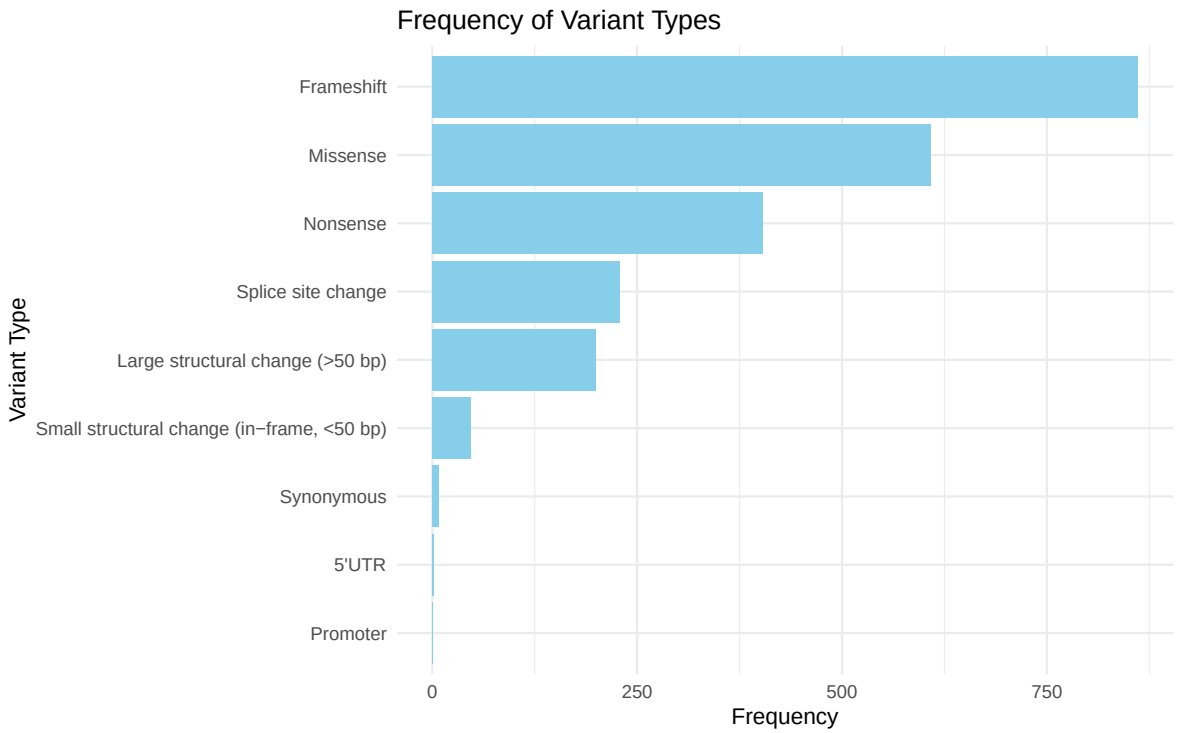
Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
year_reported	0	1	1901.49	453.74	0	2005	2010	2016	2022	

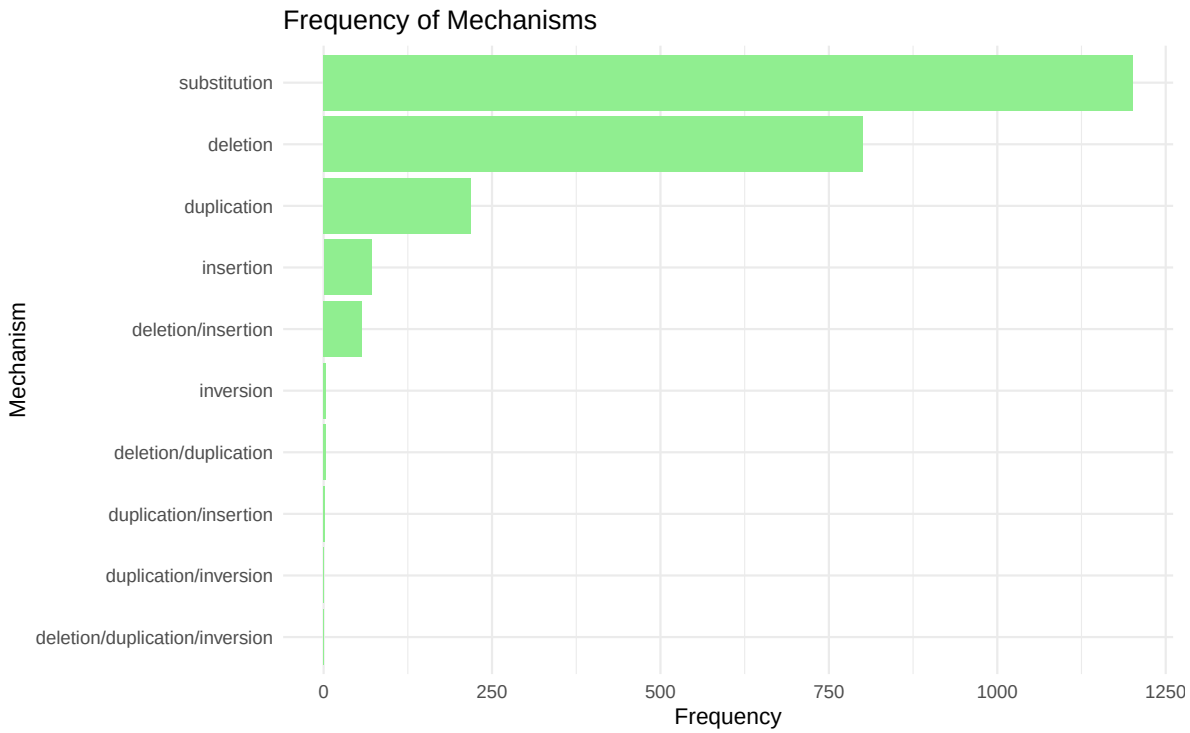
```
#Examine some frequency counts and proportions for severe Hemophilia variants

# Frequency count and proportion for variant_type
variant_type_count_sev <- severe_hemo %>%
  count(variant_type) %>%
```

```
  mutate(proportion = n / sum(n)) %>%
  arrange(desc(n))

# Frequency count and proportion for mechanism
mechanism_count_sev <- severe_hemo %>%
  count(mechanism) %>%
  mutate(proportion = n / sum(n)) %>%
  arrange(desc(n))

# Frequency count and proportion for variant_type and mechanism grouped together
grouped_variant_mechanism_count_sev <- severe_hemo %>%
  count(variant_type, mechanism) %>%
  mutate(proportion = n / sum(n)) %>%
  arrange(desc(n))

# View
variant_type_count_sev
mechanism_count_sev
grouped_variant_mechanism_count_sev
```

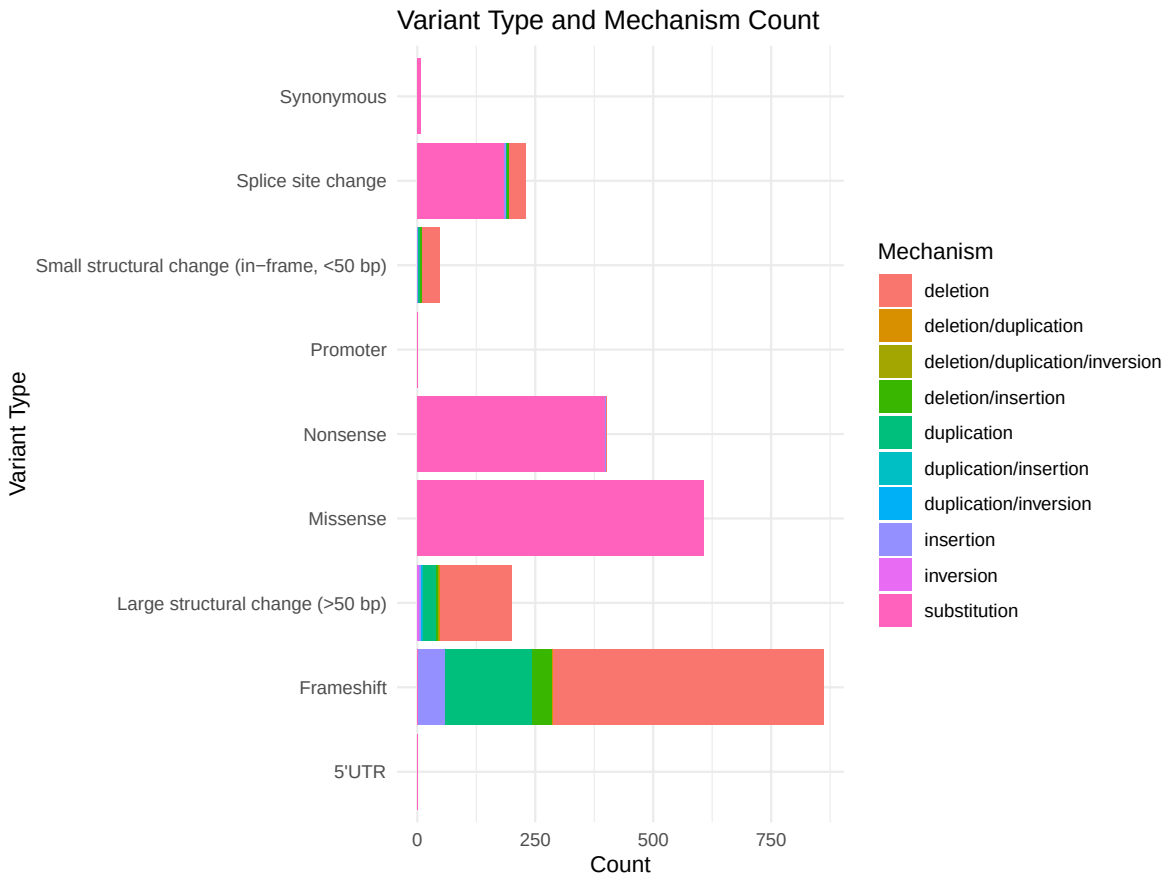
```
#Visualize most common variant types for severe Hemophilia variants

# Bar plot for variant_type frequency
ggplot(variant_type_count_sev, aes(x = reorder(variant_type, n), y = n)) +
  geom_bar(stat = "identity", fill = "skyblue") +
  coord_flip() +
  labs(title = "Frequency of Variant Types",
       x = "Variant Type",
       y = "Frequency") +
  theme(plot.title = element_text(hjust = 0.5)) +
  theme_minimal()
```




```
#Visualize most frequent mechanisms for severe Hemophilia variants

# Bar plot for mechanism frequency
ggplot(mechanism_count_sev, aes(x = reorder(mechanism, n), y = n)) +
  geom_bar(stat = "identity", fill = "lightgreen") +
  coord_flip() +
  labs(title = "Frequency of Mechanisms",
       x = "Mechanism",
       y = "Frequency") +
  theme(plot.title = element_text(hjust = 0.5)) +
  theme_minimal()
```




```
#Visualize most common variant types and mechanisms for severe Hemophilia variants

# Grouped bar plot for variant_type and mechanism
ggplot(grouped_variant_mechanism_count_sev,
       aes(x = variant_type, y = n, fill = mechanism)) +
  geom_bar(stat = "identity", position = "stack") +
  labs(title = "Variant Type and Mechanism Count",
       x = "Variant Type",
       y = "Count",
       fill = "Mechanism") +
  theme(plot.title = element_text(hjust = 0.5)) +
  theme_minimal() +
  coord_flip()
```




```
# Secondary grouping: Explore how exon, codon, domain, subtype, and in_poly_a
# relate to variant_type and mechanism

# Replace NA values in factor columns with "Missing" level
# to prep for data visualization and analysis
# This method allows us to keep all rows, not deleting any valuable scientific data

severe_hemo_cleaned <- severe_hemo %>%
  mutate(
    exon = fct_na_value_to_level(exon, "Missing"),
    codon = fct_na_value_to_level(codon, "Missing"),
    domain = fct_na_value_to_level(domain, "Missing"),
    subtype = fct_na_value_to_level(subtype, "Missing"),
    in_poly_a = fct_na_value_to_level(in_poly_a, "Missing")
  )

# All counts below are grouped by variant_type and mechanism
```

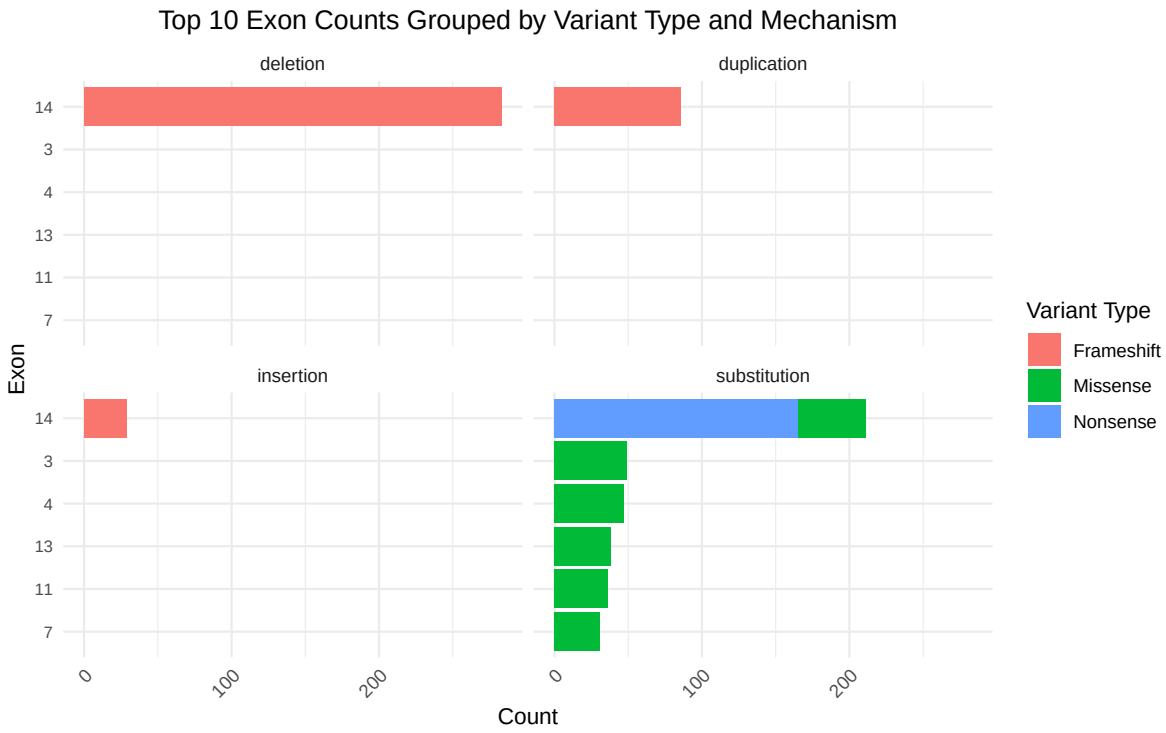
```r
# And are severe Hemophilia variants

# Explore exon counts
exon_count <- severe_hemo_cleaned %>%
  count(variant_type, mechanism, exon) %>%
  arrange(desc(n))

# Explore codon counts
codon_count <- severe_hemo_cleaned %>%
  count(variant_type, mechanism, codon) %>%
  arrange(desc(n))

# Explore domain counts
domain_count <- severe_hemo_cleaned %>%
  count(variant_type, mechanism, domain) %>%
  arrange(desc(n))

# Explore subtype counts
subtype_count <- severe_hemo_cleaned %>%
  count(variant_type, mechanism, subtype) %>%
  arrange(desc(n))

# Explore in_poly_a counts
in_poly_a_count <- severe_hemo_cleaned %>%
  count(variant_type, mechanism, in_poly_a) %>%
  arrange(desc(n))

# View some of these counts to examine top patterns
exon_count
codon_count
domain_count
subtype_count
in_poly_a_count
```

I'll group the top 10 frequent counts per variable, so that it's easier to visualize and examine patterns.

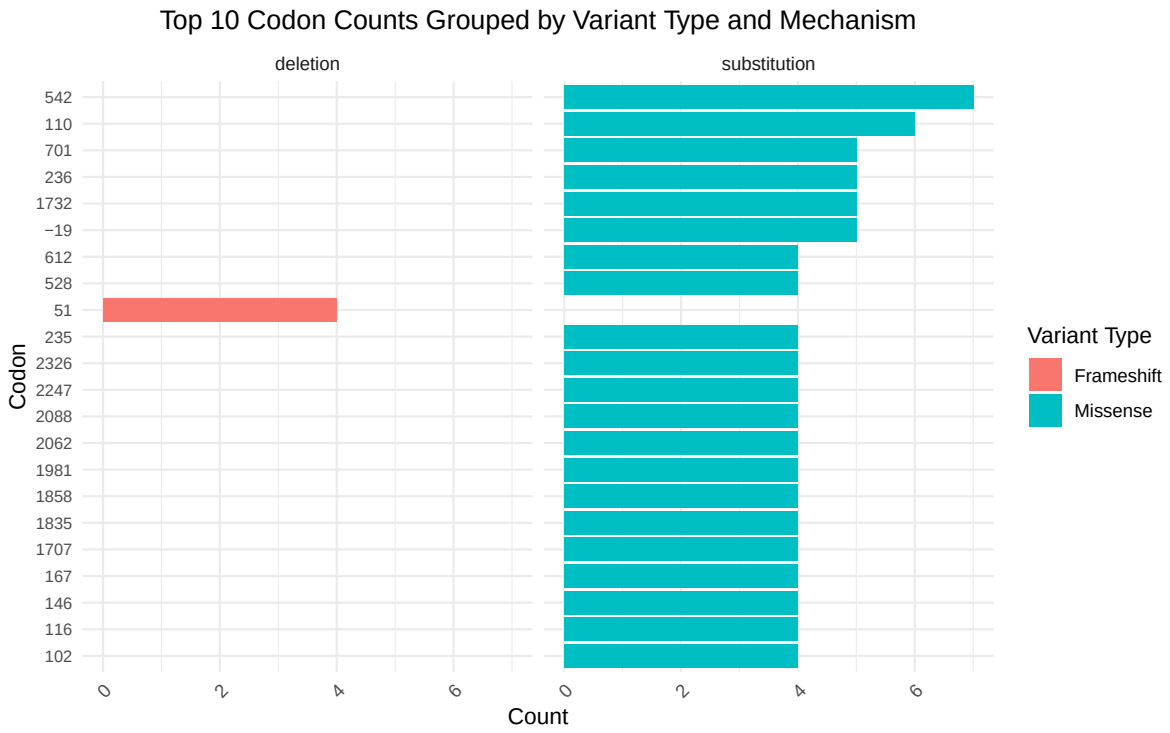
```r
exon_count_top10 <- exon_count %>%
  top_n(10, n)

# filtered out missing values because they were dominating the visualization
codon_count_top10 <- codon_count %>%
```

```
  filter(codon != "Missing") %>%
  top_n(10, n)

# filtered out missing values because they were dominating the visualization
domain_count_top10 <- domain_count %>%
  filter(domain != "Missing") %>%
  top_n(10, n)

subtype_count_top10 <- subtype_count %>%
  top_n(10, n)

in_poly_a_count_top10 <- in_poly_a_count %>%
  top_n(10, n)
```

```
# Grouped bar plot for exon_count_top10
ggplot(exon_count_top10, aes(x = reorder(exon, n), y = n, fill = variant_type)) +
  geom_bar(stat = "identity", position = "stack") +
  facet_wrap(~mechanism) +
  labs(title = "Top 10 Exon Counts Grouped by Variant Type and Mechanism",
       x = "Exon",
       y = "Count",
       fill = "Variant Type") +
  theme_minimal() +
  coord_flip() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1),
        axis.text.y = element_text(size = 8),
        plot.title = element_text(hjust = 0.5))
```




```r
# Grouped bar plot for codon_count_top10
ggplot(codon_count_top10, aes(x = reorder(codon, n), y = n, fill = variant_type)) +
  geom_bar(stat = "identity", position = "stack") +
  facet_wrap(~mechanism) +
  labs(title = "Top 10 Codon Counts Grouped by Variant Type and Mechanism",
       x = "Codon",
       y = "Count",
       fill = "Variant Type") +
  theme_minimal() +
  coord_flip() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1),
        axis.text.y = element_text(size = 8),
        plot.title = element_text(hjust = 0.5))
```



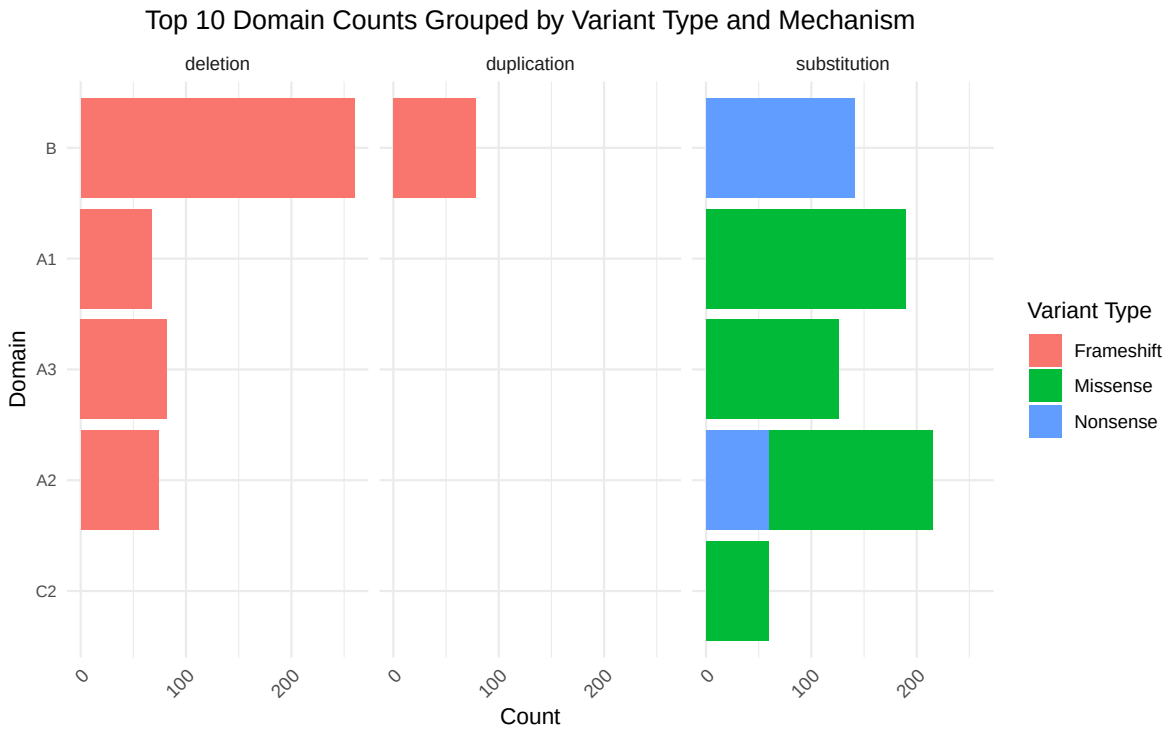
The codon graph above filters out null rows, which would otherwise dominate the graph.

I found valuable patterns through exploring missing codon data (not shown above). Deletion counts for rows missing codon data are above 150, composed of mostly large structural change variants, but also ~ 25 splice site changes. Duplication counts are ~ 25 and composed of large structural change variants. There are a few insertion and deletion/insertion entries, all large structural changes.

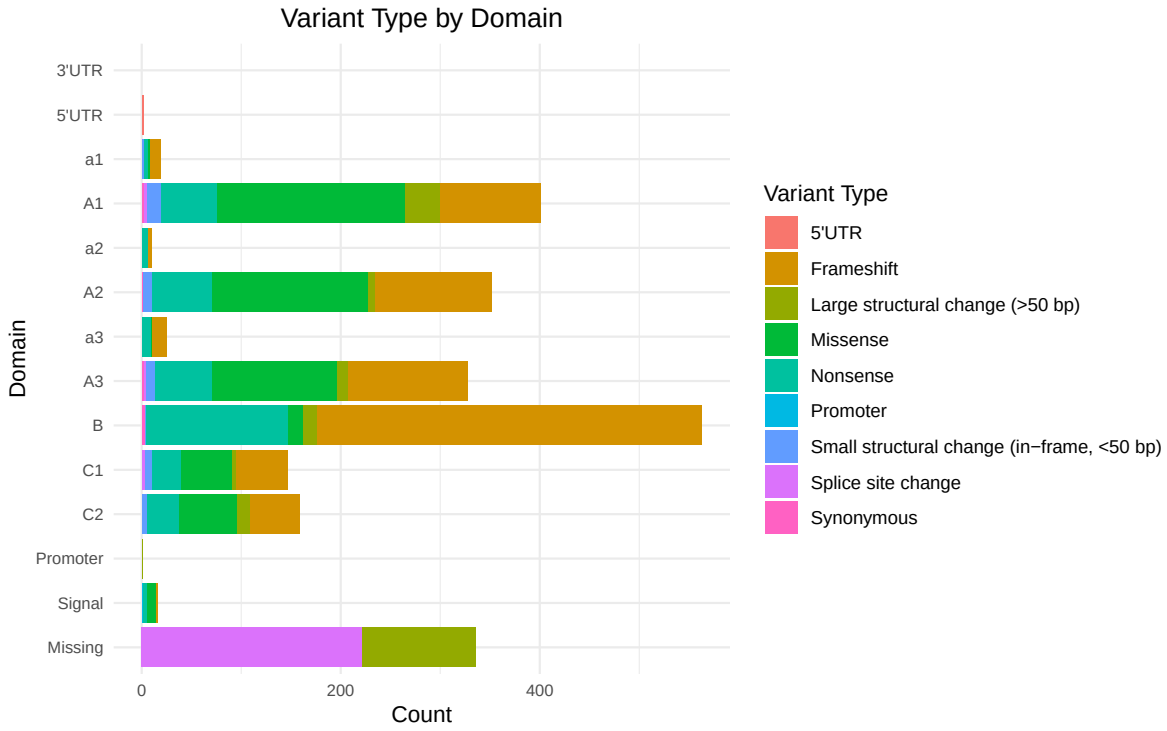
All exon data with an "intron" value has corresponding missing codon data. I am making an assumption that intron entries (but not "intron - exon" value) would reference non-coding regions. Only ~15 rows contain intron values, while over 200 rows have missing codon data. This tells me that much of the missing codon data may not be caused by non-coding regions. Perhaps data collected through scholarly articles led to this missing codon data.

```
# Grouped bar plot for domain_count_top10
ggplot(domain_count_top10, aes(x = reorder(domain, n), y = n, fill = variant_type)) +
  geom_bar(stat = "identity", position = "stack") +
  facet_wrap(~mechanism) +
  labs(title = "Top 10 Domain Counts Grouped by Variant Type and Mechanism",
       x = "Domain",
       y = "Count",
       fill = "Variant Type") +
```

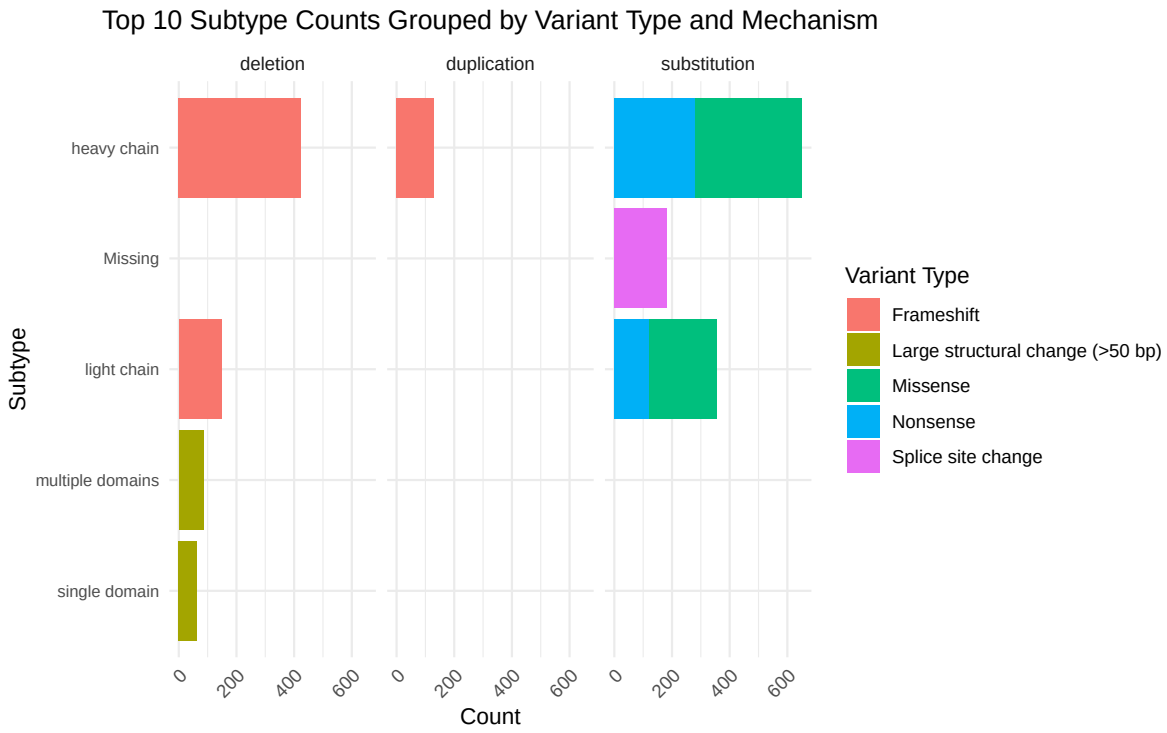
```
  theme_minimal() +
  coord_flip() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1),
        axis.text.y = element_text(size = 8),
        plot.title = element_text(hjust = 0.5))
```




```
# Grouped bar plot: All domains on y-axis, color by variant_type
ggplot(domain_count, aes(x = n, y = domain, fill = variant_type)) +
  geom_bar(stat = "identity", position = "stack") +
  labs(title = "Variant Type by Domain",
       x = "Count",
       y = "Domain",
       fill = "Variant Type") +
  theme_minimal() +
  theme(axis.text.x = element_text(size = 8),
        axis.text.y = element_text(size = 8),
        plot.title = element_text(hjust = 0.5)) +
  scale_y_discrete(limits = rev(levels(domain_count$domain)))
```




```
# Grouped bar plot for subtype_count_top10
ggplot(subtype_count_top10, aes(x = reorder(subtype, n), y = n, fill = variant_type)) +
  geom_bar(stat = "identity", position = "stack") +
  facet_wrap(~mechanism) +
  labs(title = "Top 10 Subtype Counts Grouped by Variant Type and Mechanism",
       x = "Subtype",
       y = "Count",
       fill = "Variant Type") +
  theme_minimal() +
  coord_flip() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1),
        axis.text.y = element_text(size = 8),
        plot.title = element_text(hjust = 0.5))
```

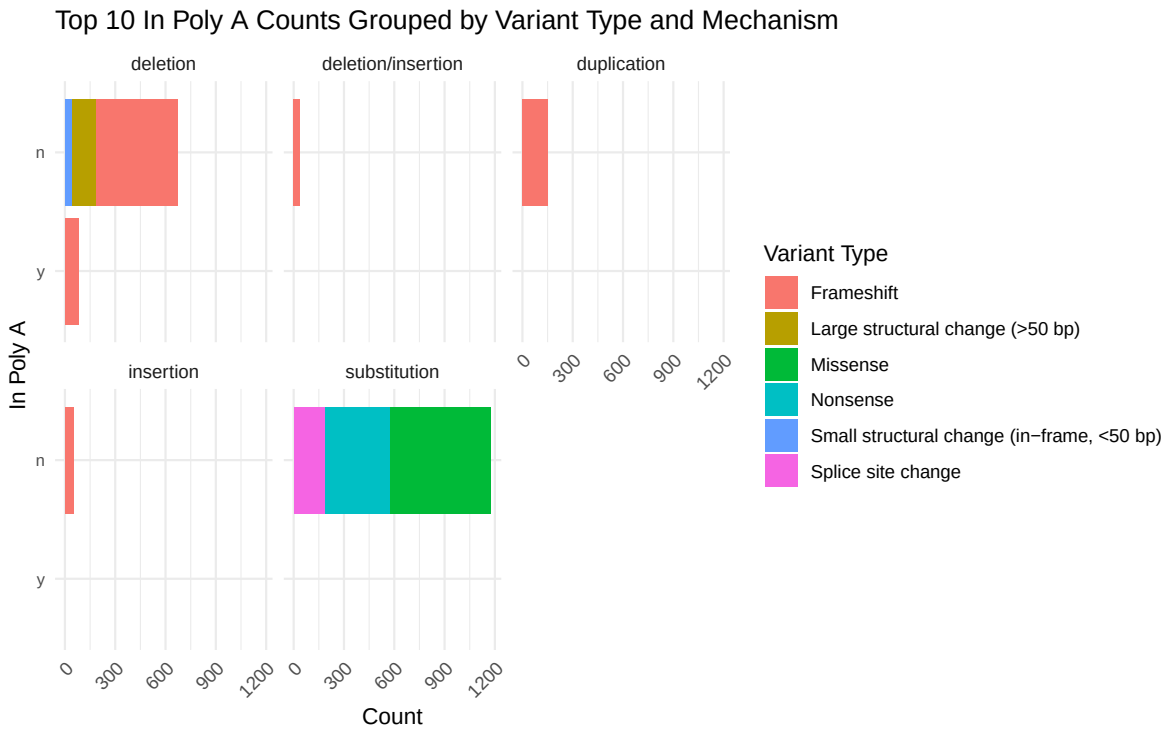


I'm noticing a pattern of missing secondary data when the variant type is a splice site change, I've noticed this with domain and subtype. Further investigation might include external research on scholarly articles, striving to understant what missing data says for each of the secondary variables. For example:

Why would domain data be missing for splicing mutations in a genetic variant dataset?

Why would subtype data be missing for splicing mutations in a genetic variant dataset?

```
# Grouped bar plot for in_poly_a_count_top10
ggplot(in_poly_a_count_top10, aes(x = reorder(in_poly_a, n), y = n, fill = variant_type)) +
  geom_bar(stat = "identity", position = "stack") +
  facet_wrap(~mechanism) +
  labs(title = "Top 10 In Poly A Counts Grouped by Variant Type and Mechanism",
       x = "In Poly A",
       y = "Count",
       fill = "Variant Type") +
  theme(plot.title = element_text(hjust = 0.5)) +
  theme_minimal() +
  coord_flip() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1),
        axis.text.y = element_text(size = 8))
```

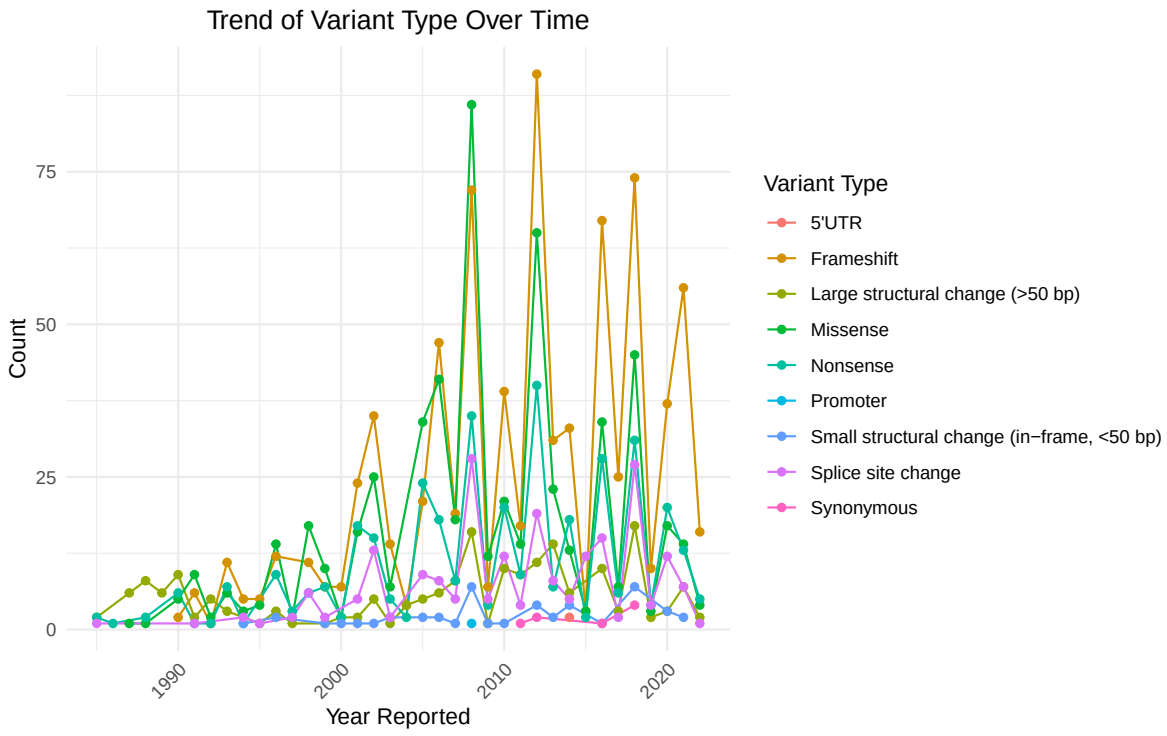



```
# Frequency of combinations: Primary (variant_type, mechanism) and Secondary Variables
combination_count <- severe_hemo_cleaned %>%
  group_by(variant_type, mechanism, exon, codon, domain, subtype, in_poly_a) %>%
  tally() %>%
  ungroup() %>%
  arrange(desc(n))  # Sort by most frequent combinations

# View the top 10 most frequent combinations
head(combination_count, 10)
```

```
# A tibble: 10 x 8
   variant_type mechanism    exon  codon domain subtype     in_poly_a     n
   <fct>        <fct>        <fct> <fct> <fct>  <fct>       <fct>     <int>
 1 Missense     substitution 11    542   A2     heavy chain n             7
 2 Missense     substitution 3     110   A1     heavy chain n             6
 3 Missense     substitution 1     -19   Signal heavy chain n             5
 4 Missense     substitution 14    701   A2     heavy chain n             5
 5 Missense     substitution 15    1732  A3     light chain n             5
 6 Missense     substitution 6     236   A1     heavy chain n             5
 7 Missense     substitution 11    528   A2     heavy chain n             4
```

```
 8 Missense     substitution 12    612   A2     heavy chain n             4
 9 Missense     substitution 14    1707  A3     light chain n             4
10 Missense     substitution 16    1835  A3     light chain n             4
```

```r
# Group data by year_reported and variant_type
year_variant_trend <- severe_hemo_cleaned %>%
  filter(year_reported != '0') %>%
  count(year_reported, variant_type) %>%
  arrange(year_reported)

# Visualize how variant_type has changed over time
ggplot(year_variant_trend, aes(x = year_reported, y = n, color = variant_type)) +
  geom_line() +
  geom_point() +
  labs(title = "Trend of Variant Type Over Time",
       x = "Year Reported",
       y = "Count",
       color = "Variant Type") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  theme(plot.title = element_text(hjust = 0.5))
```



Frameshift and missense seemed to spike in count various times since 2002. They're noticeable, with much higher counts than other variant types.

I have explored variant type, mechanism, and also gene/protein location categories such as exon, codon, and domain, but I haven't delved into specific genomic location data: HG-19 coordinates are the location on the X chromosome.

In order to begin preparing for hypothesis testing, I'll use the variants_clean data set so that I can eventually compare hg-19 coordinates across severe, mild, and moderate data.

First, I'll remove rows with non-numeric values in order to gain more information about the genetic location of variants.

```
non_numeric_rows <- variants_clean %>%
  filter(str_detect(hg19_coordinates, "[^0-9._-]"))

non_numeric_count <- nrow(non_numeric_rows)

print(paste("Number of rows with non-numeric characters in hg19_coordinates:",
            non_numeric_count))
```

```
[1] "Number of rows with non-numeric characters in hg19_coordinates: 5"
```

```
head(non_numeric_rows)
```

```
# A tibble: 5 x 13
  hgvs_c_dna hg19_coordinates hgvs_protein mature_protein variant_type mechanism
  <chr>      <chr>            <chr>        <chr>          <fct>        <chr>
1 c.1844_18~ 154182209_15418~ p.(Pro615_V~ Pro596_Val601~ Small struc~ deletion
2 c.5346_53~ 154134721_15413~ p.(Ile1782M~ Ile1763Metfs*5 Frameshift   deletion
3 c.[-171-?~ 154124352_15425~ <NA>         <NA>           Large struc~ duplicat~
4 c.143-914~ 154235303-15423~ <NA>         inv1;del; dup  Large struc~ deletion~
5 c.[5220-1~ 154136391_15412~ <NA>         <NA>           Large struc~ deletion
# i 7 more variables: exon <fct>, codon <chr>, domain <fct>, subtype <chr>,
#   in_poly_a <chr>, reported_clinical_severity <chr>, year_reported <dbl>
```

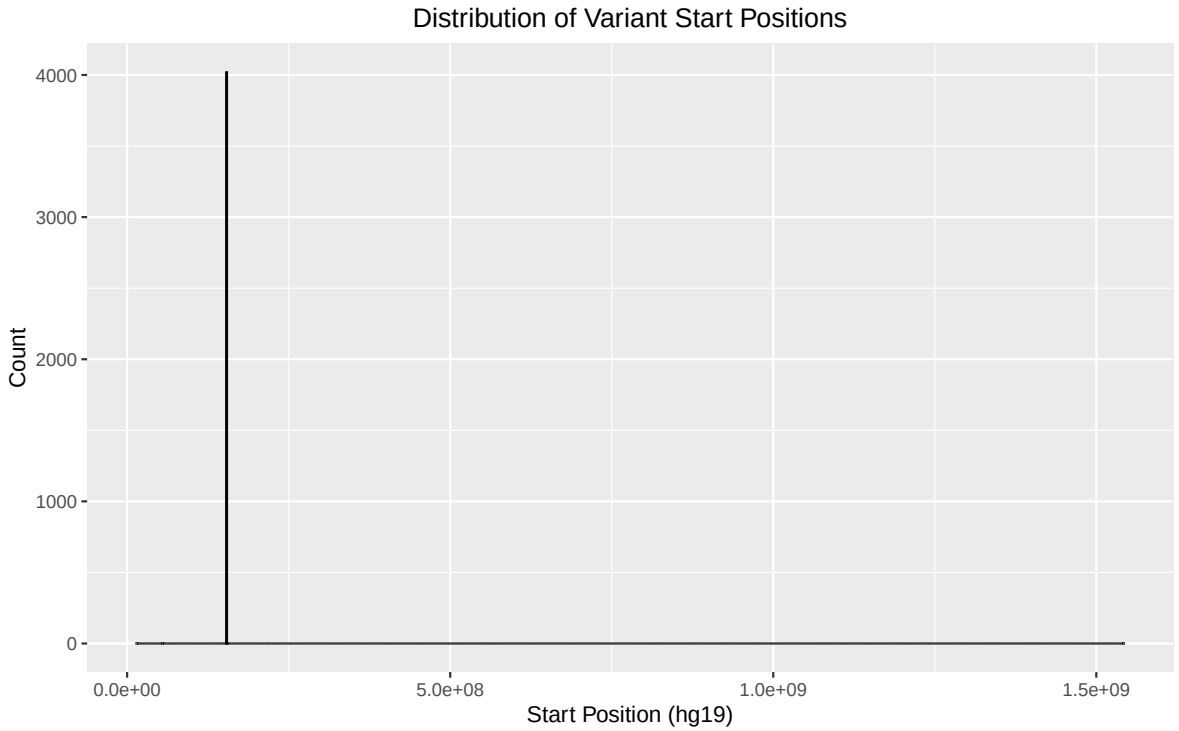
Only 3 rows include letters for the HG-19 coordinates. The letters are "del", and those entries involve the mechanism: deletion. HGVS cDNA data seems to contain more entries with "del", compared to HG-19 coordinates. Because there are only 3 rows for HG-19 coordinates, we can drop the rows in order to do analysis with the coordinates as a numeric variable.

```r
variants_clean <- variants_clean %>%
  filter(!str_detect(hg19_coordinates, "[^0-9._-]"))
```

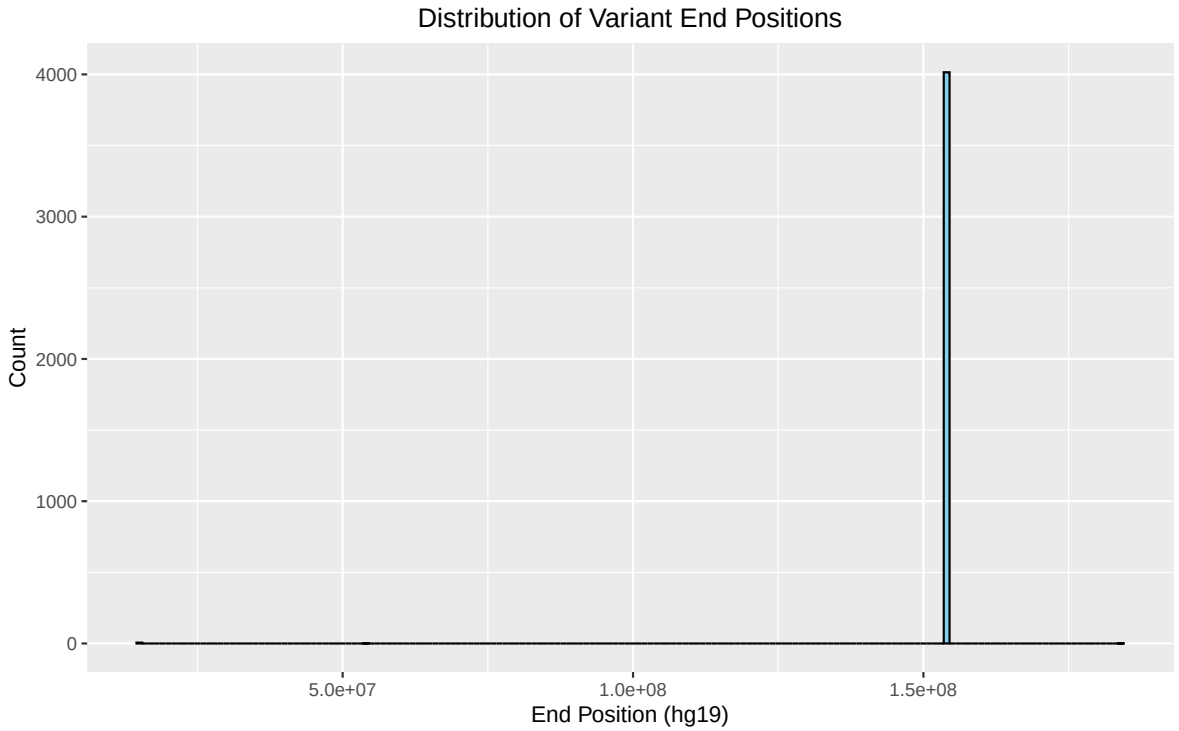
I kept the underscores, as, according to the Genome Assembly, values with an underscore show a range e.g. start_end. I'll transform the variable to read the start and end coordinates.

```r
# Transform coordinates: split into start and end if there's an underscore
variants_clean <- variants_clean %>%
  mutate(
    hg19_start = ifelse(str_detect(hg19_coordinates, "_"),
                        as.numeric(str_extract(hg19_coordinates, "^[^_]+")),
                        as.numeric(hg19_coordinates)),

    hg19_end = ifelse(str_detect(hg19_coordinates, "_"),
                      as.numeric(str_extract(hg19_coordinates, "(?<=_)[^_]+")),
                      NA),  # If no underscore, leave end as NA

    hg19_end = ifelse(is.na(hg19_end), hg19_start, hg19_end)  # Fill NA end with start

  )
```

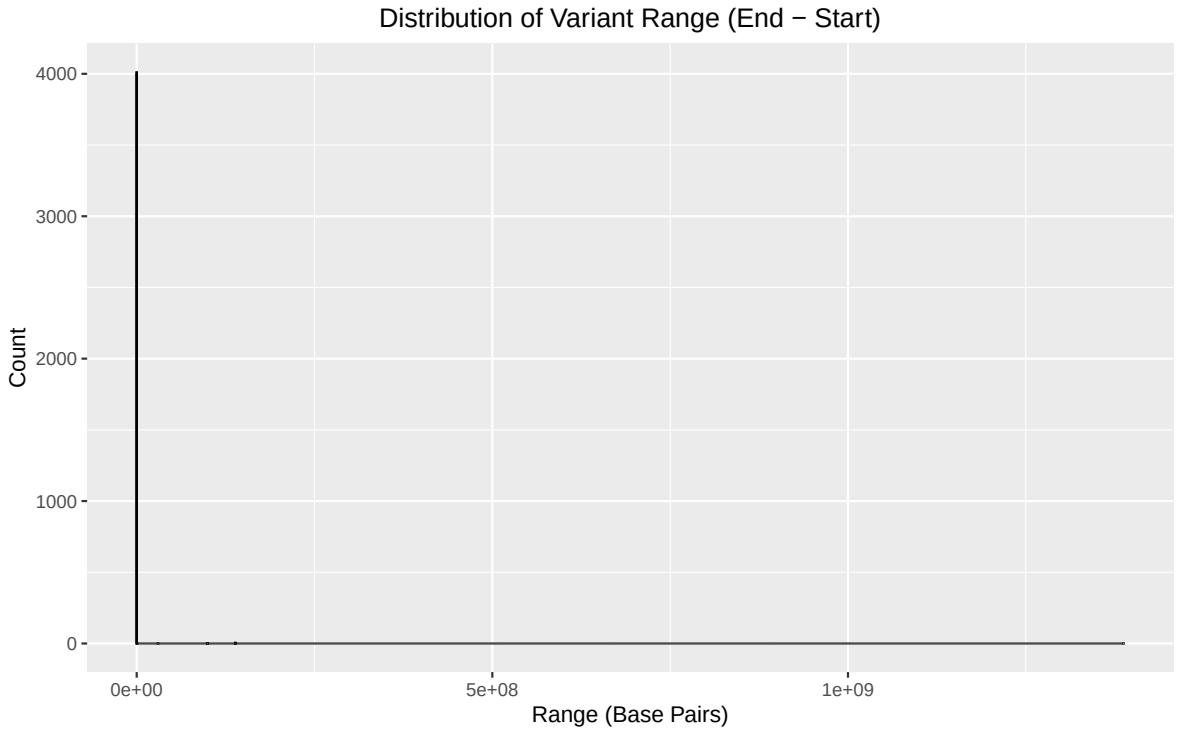
```r
ggplot(variants_clean, aes(x = hg19_start)) +
  geom_histogram(binwidth = 1000000, fill = "skyblue", color = "black") +
  labs(title = "Distribution of Variant Start Positions",
       x = "Start Position (hg19)",
       y = "Count") +
  theme(plot.title = element_text(hjust = 0.5))
```




```
ggplot(variants_clean, aes(x = hg19_end)) +
  geom_histogram(binwidth = 1000000, fill = "skyblue", color = "black") +
  labs(title = "Distribution of Variant End Positions",
       x = "End Position (hg19)",
       y = "Count") +
  theme(plot.title = element_text(hjust = 0.5))
```




```r
# Calculate the range (distance) between start and end
variants_clean <- variants_clean %>%
  mutate(range = abs(hg19_end - hg19_start))

# Plot the distribution of the range between start and end
ggplot(variants_clean, aes(x = range)) +
  geom_histogram(binwidth = 500000, fill = "lightgreen", color = "black") +
  labs(title = "Distribution of Variant Range (End - Start)",
       x = "Range (Base Pairs)",
       y = "Count") +
  theme(plot.title = element_text(hjust = 0.5))
```



The hg19 start positions show a strong concentration of variants between 0 and 5e+08, indicating key mutation hotspots on the X chromosome. The end positions are strongly concentrated around 1.5e+08, following a similar pattern but with slight differences, possibly reflecting distinct mutation types.

The range (hg19_end - hg19_start) is mostly zero, highlighting that most variants are point mutations or small changes. A few larger ranges represent structural variants, which, while rare, could have significant biological impacts.

Next, I'll conduct one final EDA on our formerly discussed exons, introns, and mixed data points.

```
# Mutate to add region_type to the main dataset
variants_clean <- variants_clean %>%
  mutate(region_type = case_when(
    str_detect(exon, "^\\d+[-\\d]*$") ~ "coding",
# Match numbers with or without hyphens
    str_detect(exon, "(?i)intron") & !str_detect(exon, "(?i)exon") ~ "non-coding",
    str_detect(exon, "(?i)intron") & str_detect(exon, "(?i)exon") ~ "mixed",
    str_detect(exon, "(?i)promoter") | str_detect(exon, "(?i)5'utr")
| str_detect(exon, "(?i)3'utr") ~ "regulatory",
    TRUE ~ "other"  # Handle any other cases
```

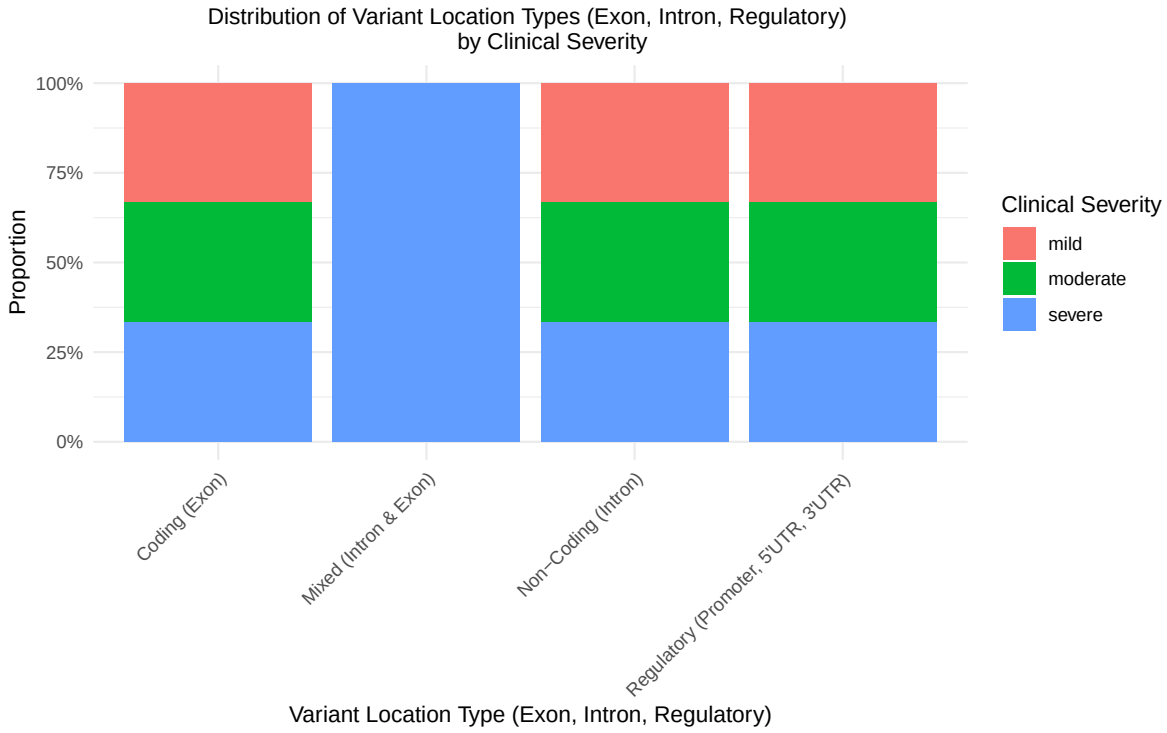
```
  ))

# Now, check the distribution by region type and severity
exon_intron_distribution <- variants_clean %>%
  count(region_type, reported_clinical_severity) %>%
  filter(reported_clinical_severity %in% c("mild", "moderate", "severe"))

# View the distribution
print(exon_intron_distribution)
```

```
# A tibble: 10 x 3
   region_type reported_clinical_severity     n
   <chr>       <chr>                      <int>
 1 coding      mild                         701
 2 coding      moderate                     475
 3 coding      severe                      2100
 4 mixed       severe                        13
 5 non-coding  mild                          60
 6 non-coding  moderate                      44
 7 non-coding  severe                       224
 8 regulatory  mild                          11
 9 regulatory  moderate                       1
10 regulatory  severe                        14
```

```
# View the cases categorized as "other"
other_cases <- variants_clean %>%
  mutate(region_type = case_when(
    str_detect(exon, "^\\d+[-\\d]*$") ~ "coding",
    str_detect(exon, "(?i)intron") & !str_detect(exon, "(?i)exon") ~ "non-coding",
    str_detect(exon, "(?i)intron") & str_detect(exon, "(?i)exon") ~ "mixed",
    str_detect(exon, "(?i)promoter") | str_detect(exon, "(?i)5'utr")
    | str_detect(exon, "(?i)3'utr") ~ "regulatory",
    TRUE ~ "other"
  )) %>%
  filter(region_type == "other")

# View the "other" cases to better understand what is falling under this category
print(other_cases)
```

The other category initially revealed ~26 cases, a mix of promoter rows, 3'UTR, and 5'UTR. All three are connected to gene regulation. I went back and added "regulatory" as a category for exon intron distribution.

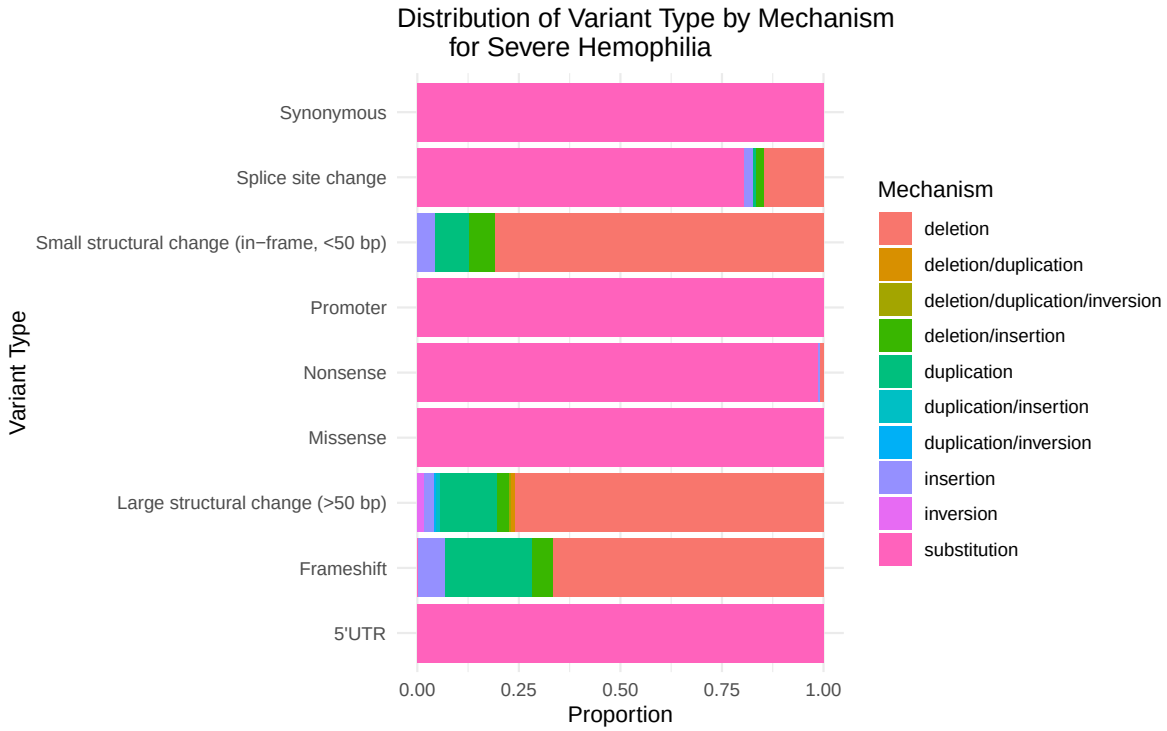
```r
ggplot(exon_intron_distribution, aes(x = region_type, fill = reported_clinical_severity)) +
  geom_bar(position = "fill") +
  labs(title = "Distribution of Variant Location Types (Exon, Intron, Regulatory)
       by Clinical Severity",
       x = "Variant Location Type (Exon, Intron, Regulatory)",
       y = "Proportion",
       fill = "Clinical Severity") +
  scale_x_discrete(labels = c("coding" = "Coding (Exon)",
                              "non-coding" = "Non-Coding (Intron)",
                              "mixed" = "Mixed (Intron & Exon)",
                              "regulatory" = "Regulatory (Promoter, 5'UTR, 3'UTR)")) +
  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.5, size = 11),
    axis.text.x = element_text(angle = 45, hjust = 1),
    plot.title.position = "panel"
  ) +
  scale_y_continuous(labels = scales::percent)
```



Hypothesis Testing

Let's return our focus to our primary explanatory variables and visualize the distribution of variant types by mechanism for severe Hemophilia. We'll focus on severe Hemophilia with the first hypothesis test, then zoom back out to test genomic locations in severe vs. non severe.

```
# Create a grouped bar plot for variant_type and mechanism
ggplot(severe_hemo_cleaned, aes(x = variant_type, fill = mechanism)) +
  geom_bar(position = "fill") +
  labs(title = "Distribution of Variant Type by Mechanism
       for Severe Hemophilia",
       x = "Variant Type",
       y = "Proportion",
       fill = "Mechanism") +
  theme(plot.title = element_text(hjust = 0.2)) +
  theme_minimal() +
  coord_flip()
```



After checking the distribution, I see a high proportion of substitution or deletion for every variant type in severe Hemophilia.

Let's perform a hypothesis test to test independence and statistical significance between variant type and mechanism. I'll use a Fisher's test because it is more reliable than a Chi-Squared test when a contingency table has 0 or <5 counts.

Null Hypothesis: There is no association between variant type and mechanism in severe Hemophilia A cases. Variant type and mechanism are independent of each other.

Alternative Hypothesis: There is an association between variant type and mechanism in severe Hemophilia A cases. Variant type and mechanism are dependent on each other.

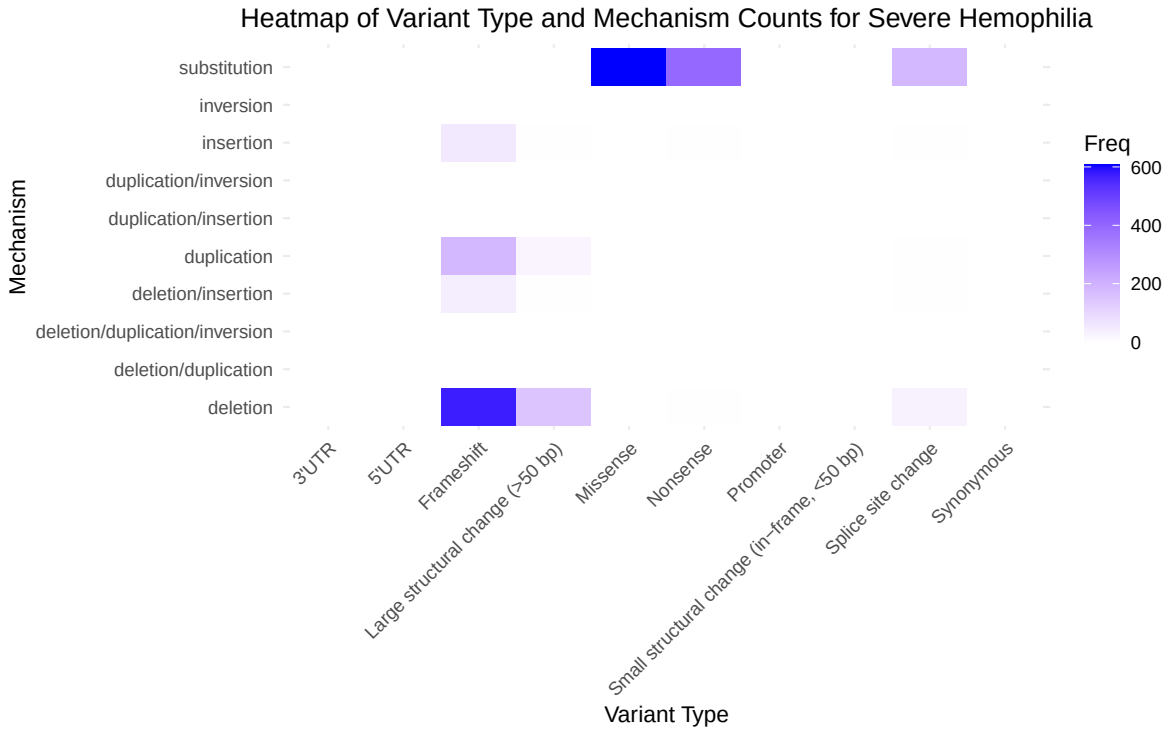
```
top_variant_types <- severe_hemo_cleaned %>%
  count(variant_type) %>%
  top_n(5, n) %>%
  pull(variant_type)

top_mechanisms <- severe_hemo_cleaned %>%
  count(mechanism) %>%
  top_n(5, n) %>%
  pull(mechanism)

# Filter the data to include only the top categories
subset_data <- severe_hemo_cleaned %>%
  filter(variant_type %in% top_variant_types, mechanism %in% top_mechanisms)

# Create the contingency table and run Fisher's test
contingency_table <- table(subset_data$variant_type, subset_data$mechanism)
fisher_result <- fisher.test(contingency_table, simulate.p.value = TRUE)

# Print the result
print(fisher_result)
```

```
    Fisher's Exact Test for Count Data with simulated p-value (based on
    2000 replicates)

data:  contingency_table
p-value = 0.0004998
alternative hypothesis: two.sided
```

With a p-value <0.05, there is a high probability of a statistically significant relationship between variant type and mechanism for severe Hemophilia. We can reject the null hypothesis and accept the alternative hypothesis: There is an association between variant type and mechanism in severe Hemophilia A cases.

I'll create a heat map of the contingency table to better understand the patterns.

```
# Visualize the contingency table using a heatmap
ggplot(as.data.frame(as.table(contingency_table)), aes(Var1, Var2, fill = Freq)) +
  geom_tile() +
  scale_fill_gradient(low = "white", high = "blue") +
  labs(title = "Heatmap of Variant Type and Mechanism Counts for Severe Hemophilia",
       x = "Variant Type",
       y = "Mechanism") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  theme(plot.title = element_text(hjust = 0.5))
```



In severe Hemophilia, the most common Variant Type/Mechanism combinations are: Frameshift/Deletion and Missense/Substitution, followed by Nonsense/Substitution. This aligns with some of the patterns found in EDA. The high frequency of Frameshift/Deletion and Missense/Substitution combinations may indicate their key role in the genomic basis of severe Hemophilia, which could guide further research into specific therapeutic targets or genetic counseling approaches.

Transitioning from focusing on severe Hemophilia hypothesis testing, we'll conduct some hypothesis tests on genomic locations in severe vs. non-severe Hemophilia.

```r
# Filter the variants_clean dataset to only include mild, moderate, and severe cases
variants_clean <- variants_clean %>%
  filter(reported_clinical_severity %in% c("mild", "moderate", "severe"))
```

```r
# Create new variable labeling severe vs. non-severe cases
variants_clean <- variants_clean %>%
  mutate(severity_group = case_when(
    reported_clinical_severity == "severe" ~ "Severe",
    reported_clinical_severity %in% c("mild", "moderate") ~ "Non-Severe"
  ))

severity_counts <- variants_clean %>%
  count(severity_group)

# View the counts
print(severity_counts)
```

```
# A tibble: 2 x 2
  severity_group     n
  <chr>          <int>
1 Non-Severe      1292
2 Severe          2351
```

Double-checking data counts for severe vs. non-severe, we see around 1300 for non-severe and ~2300 for severe. It's important to rely primarily on proportions instead or raw data counts in our analyses to account for the imbalance in counts between groups.

First we'll conduct a hypothesis test on variant location type given by the Exon column, by severity group.

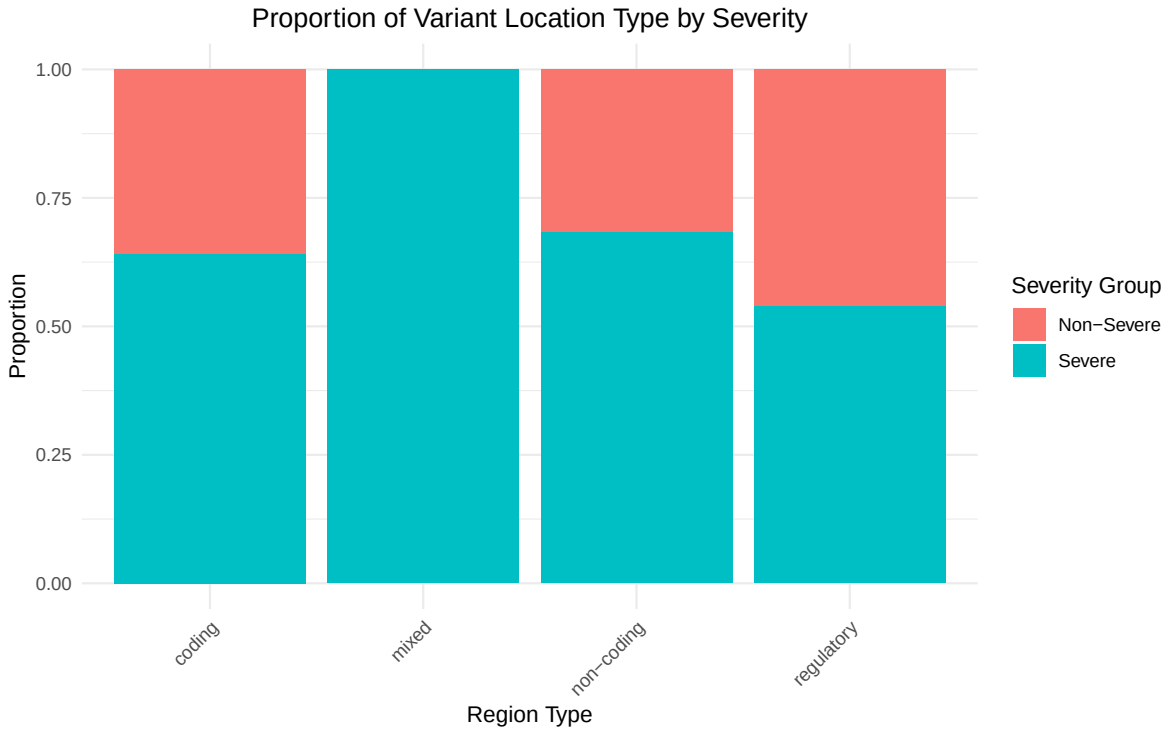
Null Hypothesis: There is no significant relationship between the variant location type (coding, non-coding, mixed, and regulatory) and reported clinical severity (mild, moderate, severe). The distribution of variants by exon location type is independent of the severity of Hemophilia.

Alternative Hypothesis: There is a significant relationship between the variant location type (coding, non-coding, mixed, and regulatory) and reported clinical severity (mild, moderate, severe). The distribution of variants by exon location type is dependent on the severity of Hemophilia.

```r
#Variant Location Type (coding, non-coding, mixed, and regulatory)
#by Severity hypothesis test

contingency_exon <- table(variants_clean$region_type, variants_clean$severity_group)

fisher_result <- fisher.test(contingency_exon)
print(fisher_result)
```

```
    Fisher's Exact Test for Count Data

data:  contingency_exon
p-value = 0.006183
alternative hypothesis: two.sided
```

```r
ggplot(variants_clean, aes(x = region_type, fill = severity_group)) +
  geom_bar(position = "fill") +
  labs(title = "Proportion of Variant Location Type by Severity",
       x = "Region Type",
       y = "Proportion",
       fill = "Severity Group") +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5),
        axis.text.x = element_text(angle = 45, hjust = 1))
```

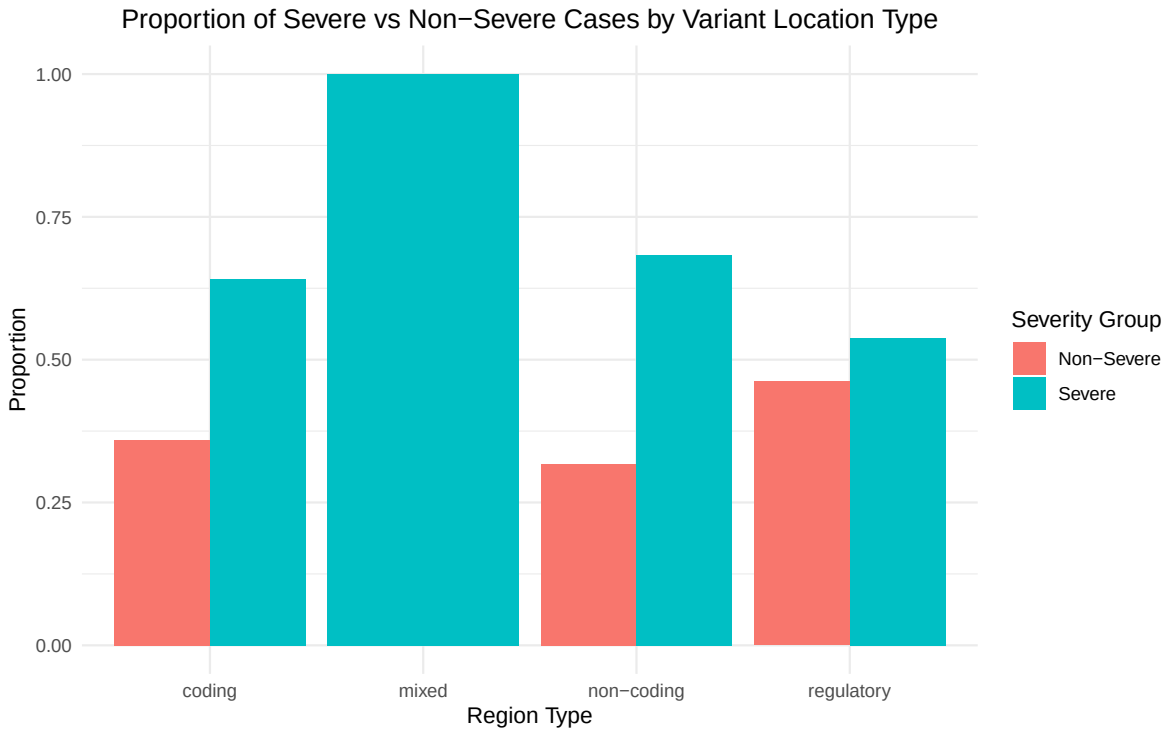


With p-value < 0.05, I'll reject the null hypothesis and accept the alternative hypothesis: There is a significant relationship between the variant location type (coding, non-coding, mixed, and regulatory) and reported clinical severity.

After confirming a significant relationship, visualizing proportions helps clarify the distribution of severe and non-severe cases across variant locations.

```
# Calculate the proportion of severe and non-severe by region type
severity_proportions <- variants_clean %>%
  group_by(region_type, severity_group) %>%
  tally() %>%
  group_by(region_type) %>%
  mutate(proportion = n / sum(n))

# Visualize the proportions
ggplot(severity_proportions, aes(x = region_type, y = proportion, fill = severity_group)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Proportion of Severe vs Non-Severe Cases by Variant Location Type",
       x = "Region Type",
       y = "Proportion",
       fill = "Severity Group") +
```

```
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5))
```

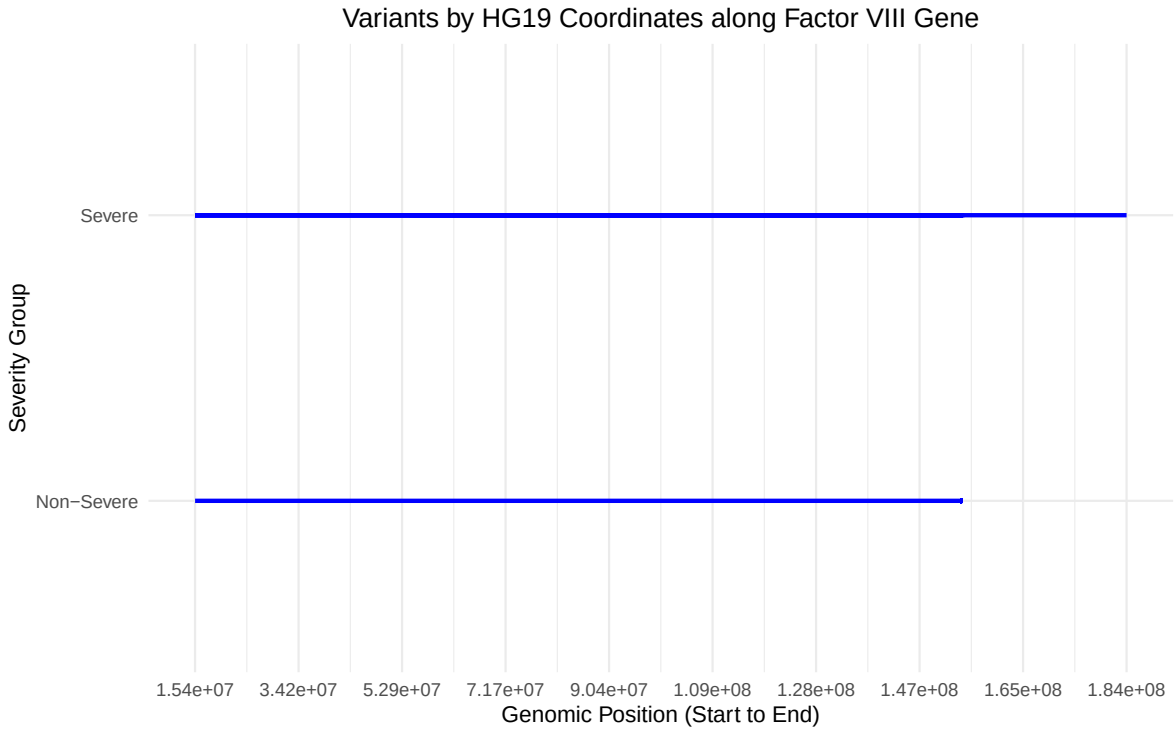


The mixed region is unique in that it only has severe cases, which might suggest that variants at exon-intron boundaries are always associated with severe Hemophilia. The small number, 13, of cases here could impact the accuracy of this finding.

Coding and non-coding regions both show a higher proportion of severe cases than non-severe cases, with non-coding variants showing a slightly lower proportion of severe cases compared to coding regions. Regulatory regions have a small sample size, but show a near-even split between severe and non-severe cases. The small number of cases, 26 total regulatory cases, could impact the accuracy of this finding. Further research and data collection may be needed to reach conclusions here.

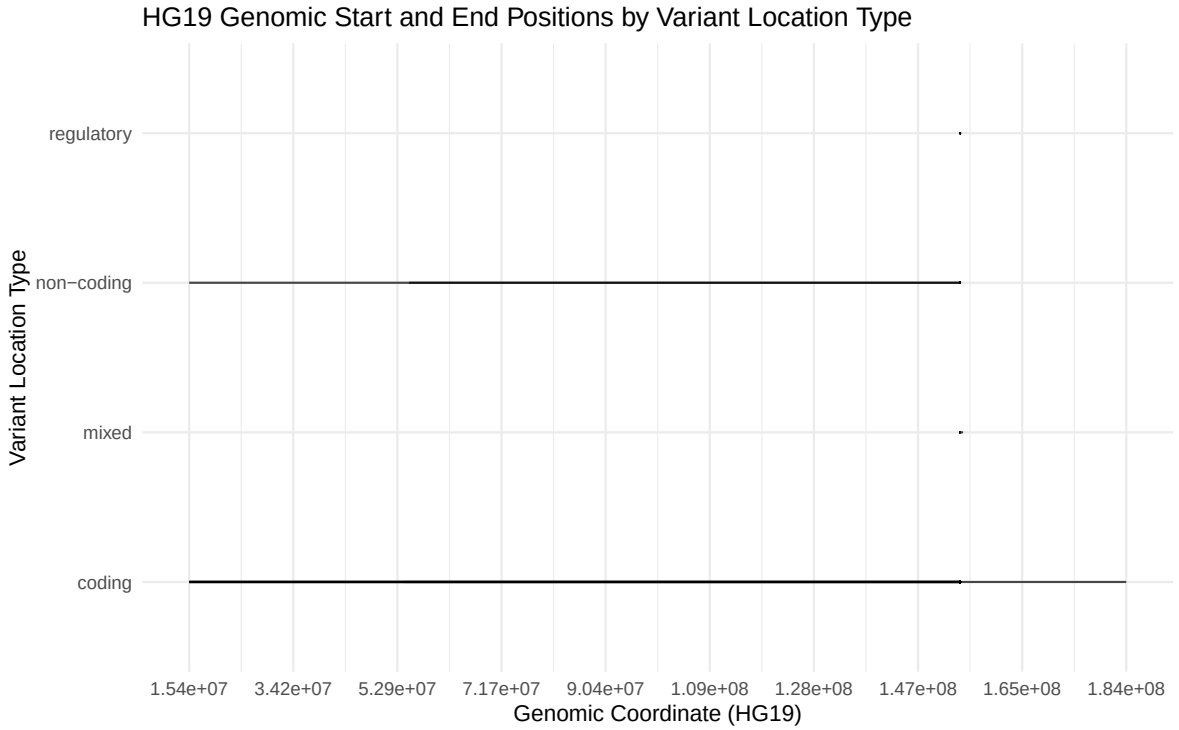
Now that we have seen how the variant location types (coding, non-coding, mixed) relate to clinical severity, let's explore the specific start and end positions of these variants within the Factor VIII gene (hg19 coordinates). This will help us understand the distribution of variants across the gene and whether certain regions are more frequently associated with severe Hemophilia.

```r
ggplot(variants_clean,
       aes(x = hg19_start, xend = hg19_end, y = severity_group)) +
  geom_segment(linewidth = 1, color = "blue") +
  labs(title = "Variants by HG19 Coordinates along Factor VIII Gene",
       x = "Genomic Position (Start to End)",
       y = "Severity Group") +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5)) +
  scale_x_continuous(labels = scales::scientific,
                     breaks = seq(min(variants_clean$hg19_start),
                                  max(variants_clean$hg19_end),
                                  length.out = 10))
```



This plot shows that both severe and non-severe cases have variants starting at approximately 1.54e+07 on the HG19 genomic scale, indicating a shared starting region. However, non-severe variants end around 1.55e+08, while severe variants extend further, ending at 1.84e+08. This suggests that severe cases involve variants distributed across a broader genomic range compared to non-severe cases, potentially reflecting differences in their underlying mechanisms or associated genomic regions.

```r
ggplot(variants_clean,
       aes(x = hg19_start, xend = hg19_end, y = region_type)) +
  geom_segment(aes(yend = region_type), alpha = 0.7) +
  labs(title = "HG19 Genomic Start and End Positions by Variant Location Type",
       x = "Genomic Coordinate (HG19)",
       y = "Variant Location Type") +
  theme_minimal() +
  scale_x_continuous(labels = scales::scientific,
                     breaks = seq(min(variants_clean$hg19_start),
                                  max(variants_clean$hg19_end),
                                  length.out = 10))
```



Coding variants span the broadest genomic range, from 1.54e+07 to 1.84e+08, with a denser concentration at the start. This suggests that coding variants are more evenly distributed across the gene, with some regions having higher variant density.

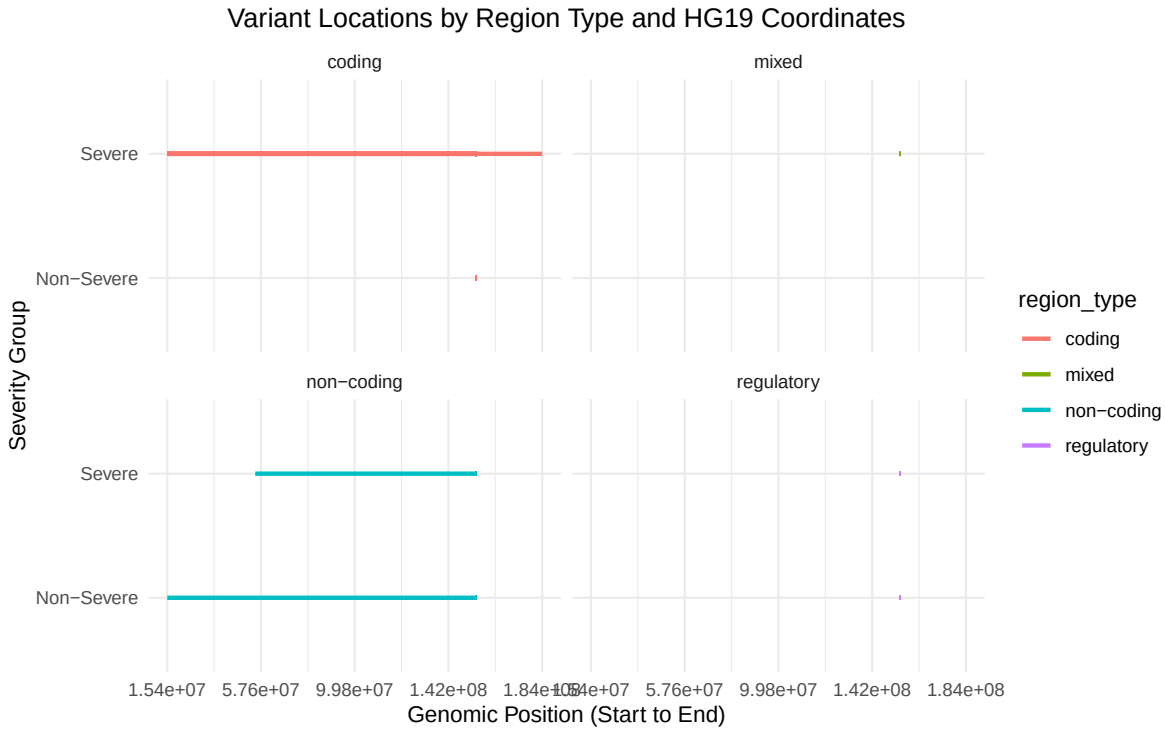
Non-coding variants are concentrated between 1.54e+07 and 1.55e+08, with a noticeable cluster around 5.30e+07. This indicates that non-coding variants are mainly located in a specific region of the gene, with a significant increase in density around this point.

Regulatory variants and mixed variants appear as small, localized peaks around 1.55e+08, suggesting these variants are more restricted to a particular genomic region, possibly linked

to regulatory or intron-exon boundary features that have fewer variant occurrences.

Overall, the plot highlights that coding variants are more spread out across the gene, while non-coding and regulatory variants are more concentrated in specific regions, with non-coding variants showing a distinct clustering pattern.

```
ggplot(variants_clean,
       aes(x = hg19_start, xend = hg19_end, y = severity_group, color = region_type)) +
  geom_segment(linewidth = 1) +
  labs(title = "Variant Locations by Region Type and HG19 Coordinates",
       x = "Genomic Position (Start to End)",
       y = "Severity Group") +
  facet_wrap(~ region_type) +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5)) +
  scale_x_continuous(labels = scales::scientific, breaks = seq(min(variants_clean$hg19_start)
```



The third HG-19 graph shows that both severe and non-severe cases have a significant presence of variants in the non-coding region, with non-severe variants particularly dense between 1.54e+07 and 1.55e+08. This suggests that non-coding variants, which may affect gene regulation rather than protein function, play a role in non-severe cases (Schipper, 2022). In contrast, severe cases show variants spread across a broader genomic range, from 1.54e+07 to 1.84e+08,

indicating that non-coding variants are still present and may contribute to severe outcomes, potentially through regulatory mechanisms.

Interestingly, regulatory and mixed intron-exon boundary variants, both of which were associated with severe cases, appear prominently around 1.55e+08. This suggests that these regions may be crucial for gene regulation or splicing processes, with variants in these areas potentially playing a role in the severity of Hemophilia. While severe cases show a broader distribution of coding variants, which likely disrupt the Factor VIII protein function, the prominence of the 1.55e+08 region in regulatory and mixed variants highlights the significance of these regions in influencing severity through regulatory effects or splicing issues.

Overall, both coding and non-coding variants influence severity, with non-coding and regulatory variants likely affecting gene expression regulation or splicing, while coding variants primarily impact protein function. The association of regulatory and mixed variants with severe cases suggests that these regions could be important for understanding the mechanisms behind severe Hemophilia.

We'll move on to a hypothesis test for HG-19 range and severity group. HG-19 start and HG-19 end both showed no significant relationship with severity (I'm not showing those tests below to reduce repetitiveness– same process as below).

First, we need to check normality assumptions and homogeneity of variance.

```
# Shapiro-Wilk Test for Normality for both groups (Severe vs Non-Severe)
shapiro_severe <- shapiro.test(variants_clean$range
                               [variants_clean$severity_group == "Severe"])
shapiro_non_severe <- shapiro.test(variants_clean$range
                                   [variants_clean$severity_group == "Non-Severe"])

# Levene's Test for Homogeneity of Variance
levene_test <- car::leveneTest(range ~ severity_group, data = variants_clean)

# Check the results
shapiro_severe
```

```
    Shapiro-Wilk normality test

data:  variants_clean$range[variants_clean$severity_group == "Severe"]
W = 0.032701, p-value < 2.2e-16
```

```
shapiro_non_severe
```

```
    Shapiro-Wilk normality test

data:  variants_clean$range[variants_clean$severity_group == "Non-Severe"]
W = 0.0096275, p-value < 2.2e-16
```

```
levene_test
```

```
Levene's Test for Homogeneity of Variance (center = median)
        Df F value Pr(>F)
group    1  2.3597 0.1246
      3641
```

Assumption of normality is not met (p-value <0.05), although homogeneity of variance is met (p-value >0.05), so let's perform a Wilcoxon rank-sum test.

Null Hypothesis: There is no difference in the range of genomic locations (difference between hg19_start and hg19_end) between severe and non-severe cases of Hemophilia A. The distribution of the range is the same for both groups.

Alternative Hypothesis: There is a difference in the HG-19 coordinate range of genomic locations between severe and non-severe cases of Hemophilia A. The distribution of the range differs between the two groups.

```
wilcox_result <- wilcox.test(range ~ severity_group, data = variants_clean)
print(wilcox_result)
```

```
    Wilcoxon rank sum test with continuity correction

data:  range by severity_group
W = 1208235, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
```

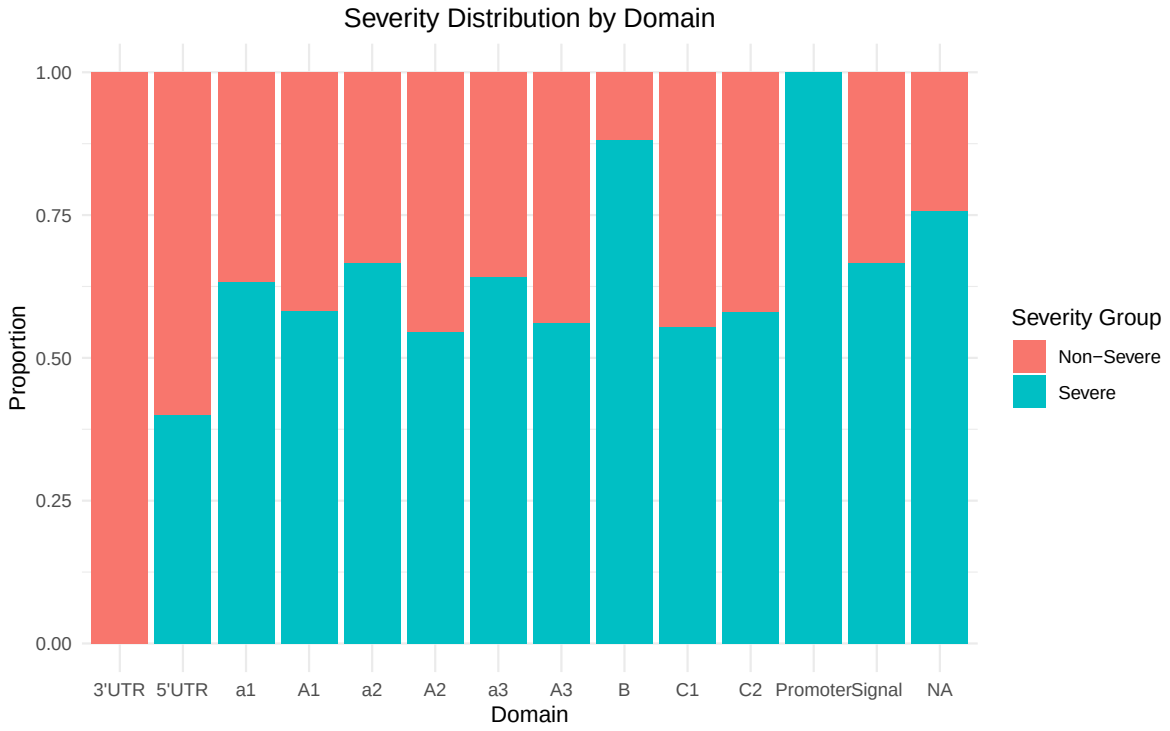
The p-value is less than 2.2e-16, so we reject the null hypothesis and accept the alternative hypothesis. There is a significant difference between the HG-19 coordinate range for severe and non-severe Hemophilia A cases. Severe cases may involve larger or more varied genomic regions, or the presence of variants with larger spans could be more prevalent in the severe cases.

```r
# Fisher's test for domain
contingency_domain <- table(variants_clean$domain, variants_clean$severity_group)
fisher_domain_result <- fisher.test(contingency_domain, simulate.p.value = TRUE)
print(fisher_domain_result)
```

```
    Fisher's Exact Test for Count Data with simulated p-value (based on
    2000 replicates)

data:  contingency_domain
p-value = 0.0004998
alternative hypothesis: two.sided
```

```r
contingency_domain
```

```r
# Plot the distribution of severity across domains
ggplot(variants_clean, aes(x = domain, fill = severity_group)) +
  geom_bar(position = "fill") +
  labs(title = "Severity Distribution by Domain",
       x = "Domain",
       y = "Proportion",
       fill = "Severity Group") +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5))
```



The Fisher's Exact Test demonstrates a significant association between domain and severity group (p-value = 0.0005). Domains such as B and Signal are predominantly linked to severe cases, with the B domain showing a particularly large proportion of severe cases relative to non-severe. In contrast, the C1, C2, A1, A2, and A3 domains all exhibit a mix of severity groups, but severe cases consistently outnumber non-severe cases across these domains.

Non-coding regions like the 3'UTR and Promoter are sparsely represented, with the 3'UTR having only a single non-severe case and the Promoter associated with one severe case, suggesting these regions contribute minimally to clinical outcomes. The a1, a2, and a3 domains also have relatively smaller contributions to overall severity compared to their uppercase counterparts.

This analysis highlights the concentration of severe cases in coding regions, particularly in the B domain, while non-coding regions such as the 3'UTR and Promoter domains and regulatory regions like the Signal domain play a less prominent role, with fewer cases overall. However, the Signal domain shows a slightly higher proportion of severe cases, suggesting its potential relevance to severe presentations.

```
# Fisher's test for subtype
contingency_subtype <- table(variants_clean$subtype, variants_clean$severity_group)
fisher_subtype_result <- fisher.test(contingency_subtype, simulate.p.value = TRUE)
print(fisher_subtype_result)
```

```
    Fisher's Exact Test for Count Data with simulated p-value (based on
    2000 replicates)

data:  contingency_subtype
p-value = 0.0004998
alternative hypothesis: two.sided
```

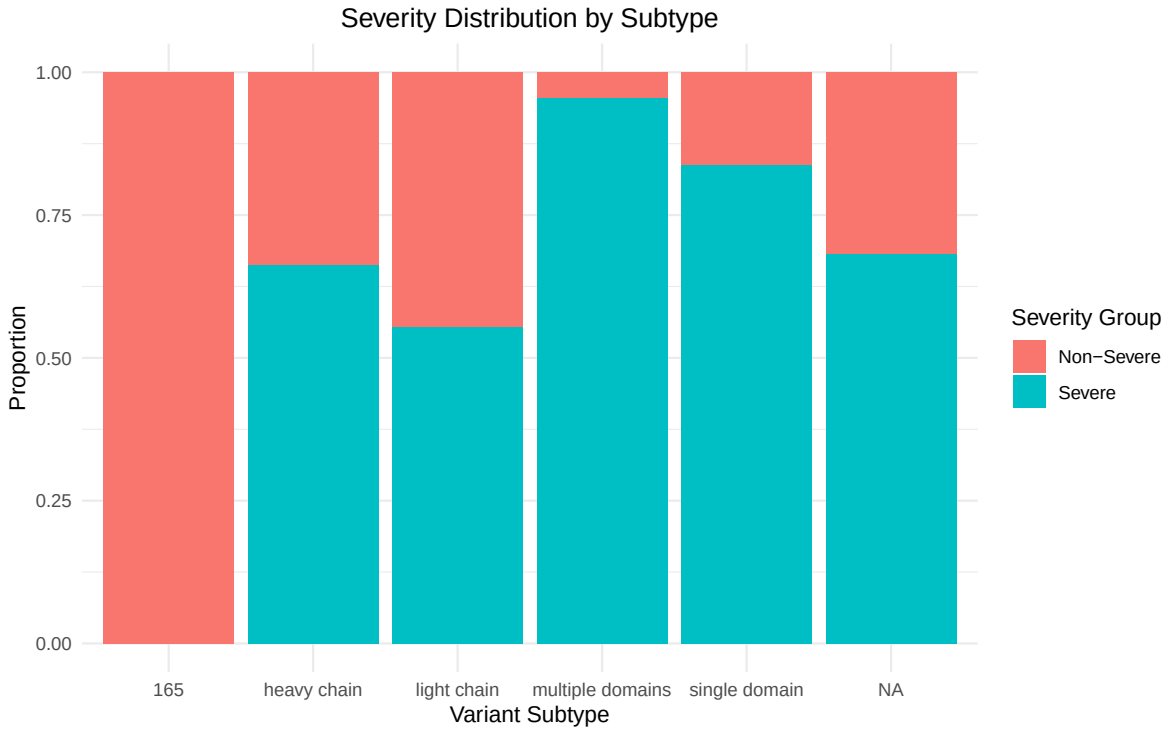
```
# Fisher's test for codon
contingency_codons <- table(variants_clean$codon, variants_clean$severity_group)
fisher_codon_result <- fisher.test(contingency_codons, simulate.p.value = TRUE)
print(fisher_codon_result)
```

```
    Fisher's Exact Test for Count Data with simulated p-value (based on
    2000 replicates)

data:  contingency_codons
p-value = 0.0004998
alternative hypothesis: two.sided
```

Both Fisher's Exact Tests suggest highly significant associations (p-value = 0.0004998) between clinical severity and both the subtype and codon of genetic variants. This indicates that the distribution of specific subtypes and codons is not random and may play an important role in distinguishing severe from non-severe cases.

```
contingency_subtype
```

```
# Plot the distribution of severity across subtypes
ggplot(variants_clean, aes(x = subtype, fill = severity_group)) +
  geom_bar(position = "fill") +
  labs(title = "Severity Distribution by Subtype",
       x = "Variant Subtype",
       y = "Proportion",
       fill = "Severity Group") +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5))
```



Interpreting the contingency table and distribution graph, I noticed that heavy chain and light chain subtypes dominate both severe and non-severe cases, but heavy chain has a higher proportion of severe cases compared to non-severe cases. On the other hand, the light chain has a relatively balanced distribution but leans slightly toward severe cases. The multiple domains subtype is strongly associated with severe cases, while single domain and 165 subtypes show a much smaller presence, with 165 only appearing in non-severe cases. This suggests that subtypes like heavy chain and multiple domains may be strong influencers of clinical severity.

```
# Fisher's test for Poly-A in regulatory region
contingency_poly_a <- table(variants_clean$in_poly_a, variants_clean$severity_group)

fisher_poly_a_result <- fisher.test(contingency_poly_a, simulate.p.value = TRUE)
print(fisher_poly_a_result)
```

```
    Fisher's Exact Test for Count Data

data:  contingency_poly_a
p-value = 6.915e-11
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
```

```
 2.460158 6.739861
sample estimates:
odds ratio
  3.972602
```

The Fisher's Exact Test shows a significant relationship between poly-A regions and severity group (p-value = 6.915e-11, with an odds ratio of 3.97. This suggests that variants with poly-A regions are nearly four times more likely to be severe, suggesting a potential link between poly-A regions and increased severity.

```
contingency_poly_a
```

```
  Non-Severe Severe
n       1267   2200
y         20    138
```

Discussion

The primary goal of this analysis was to identify the most frequently reported genetic variants and their associated genomic locations linked to severe Hemophilia A. Through a combination of exploratory data analysis and statistical hypothesis testing, we uncovered significant relationships between variant characteristics and clinical severity, giving us deeper insights on the underlying genetic mechanisms of the condition.

Key Findings:

1. **Variant Type and Mechanism Association:**
    - Fisher's exact test (p-value < 0.001) confirmed a statistically significant association between variant_type and mechanism. The most common combinations included Frameshift/Deletion, Missense/Substitution, and Nonsense/Substitution, indicating that structural changes in the F8 gene are critical drivers of severe clinical outcomes.
2. **Variant Location Type by Clinical Severity:**
    - A significant relationship was observed between variant location type (coding, non-coding, mixed, and regulatory) and severity (p-value = 0.006). Severe cases were predominantly associated with coding regions, while mixed exon-intron boundaries were exclusively linked to severe cases, suggesting a distinct functional role.

3. **Genomic Coordinates and Severity:**

    - Wilcoxon rank-sum testing (p-value $<$ 2.2e-16) revealed significant differences in the genomic coordinate ranges between severe and non-severe cases. Severe cases spanned broader genomic regions, suggesting that large-scale mutations contribute more frequently to severe phenotypes.

4. **Domain, Subtype, and Codon Relevance:**

    - Significant associations exist between severity and the F8 gene domains (p-value = 0.0005), subtypes (p-value = 0.0005), and codons (p-value = 0.0005). The B domain and heavy chain subtypes were highly represented within severe cases, suggesting their importance in determining clinical severity.

5. **Poly-A Regions and Severity:**

    - Fisher's test showed a strong association between poly-A region presence and severity (p-value = 6.915e-11, odds ratio = 3.97), suggesting that poly-A regions might play a regulatory role influencing severe outcomes.

Interpretation and Implications:

The findings confirmed expected associations, such as the prominence of Frameshift and Missense variants causing protein dysfunction, while also highlighting lesser-studied areas like exon-intron boundaries and poly-A regulatory regions. These findings align with existing literature but also suggest underexplored factors may call for further study.

Limitations:

- **Dataset Representation:** Due to the rarity of Hemophilia A, larger datasets may not exist, limiting the ability to generalize or conduct analysis on larger samples.
- **Data Quality:** Missing entries for variables like codons and domains, likely due to data collection partially from literature, complicated the analysis.
- **Domain Knowledge:** My lack of genetic expertise limited deeper biological interpretation.

Future Research Directions:

1. Conduct year-reported and time-series analyses.
2. Investigate variants with atypical patterns using specialized genetic sequence databases.
3. Further investigate synonymous, boundary, and regulatory variants, which showed unique connections to severe Hemophilia, and seem under-researched within existing literature.
4. Collaborate with geneticists to refine biological interpretations and identify new research questions.

Future research, guided by domain expertise could create further insights into the genetic basis of severe Hemophilia A.

Conclusion

In summary, this study found that severe Hemophilia A is influenced, in part, by five genetic factors studied: structural changes in variant types linked to specific mechanisms, coding region mutations, broad genomic coordinate involvement, domain, subtype, and codon-specific associations, and poly-A region disruptions. These findings highlight a complex genetic landscape where specific variants influence clinical severity. Future studies integrating genetic, clinical, and temporal data could clarify how these factors interact to affect clinical severity.

References

Centers for Disease Control and Prevention. (n.d.). *Hemophilia Mutation Project*. Retrieved from https://www.cdc.gov/hemophilia/mutation-project/index.html

Inaba, H., Shinozawa, K., Fukutake, K. and Amano, K. (2018), A novel synonymous variant in the F8 gene, p.(Leu40=)/c.120C>A, likely causes mild haemophilia A. Haemophilia, 24: e289-e292. https://doi.org/10.1111/hae.13568

Schipper, Marijn, Posthuma, Danielle. (2022) Demystifying non-coding GWAS variants: an overview of computational tools and methods, *Human Molecular Genetics*, Volume 31, Issue R1, Pages R73–R83, https://doi.org/10.1093/hmg/ddac198