# Human Chromosome Variation

### Isaac McPadden

## Introduction

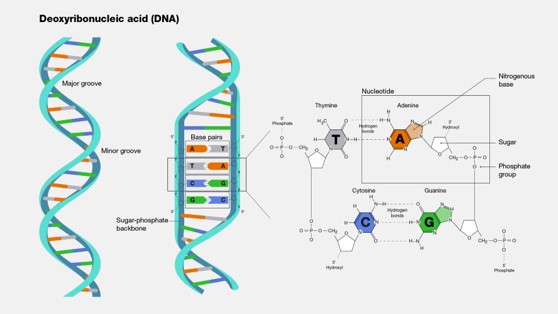
### Project Description

For my project I used 1000 Genomes (<https://www.internationalgenome.org/>) genetics data to measure genetic variation in each of the 24 (01-22, X, and Y) human chromosomes. The two goals for this project were to determine which chromosome experienced the highest amount of genetic variation and to determine how many mutations in each chromosome had high impact.

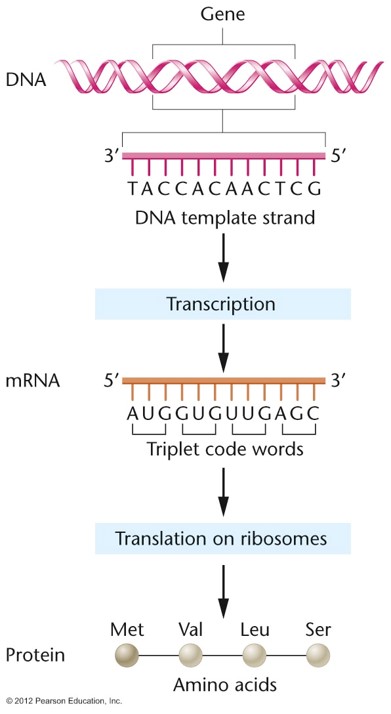
### Biochemistry and Genetics Primer

To understand the genetics data and results, a short primer on biochemistry is needed. I have a Bachelor of Arts in Biochemistry and all of the following information is common knowledge in the field. Nonetheless, any information included in this section can be found in Biochemistry, Fourth Edition, by Voet and Voet.

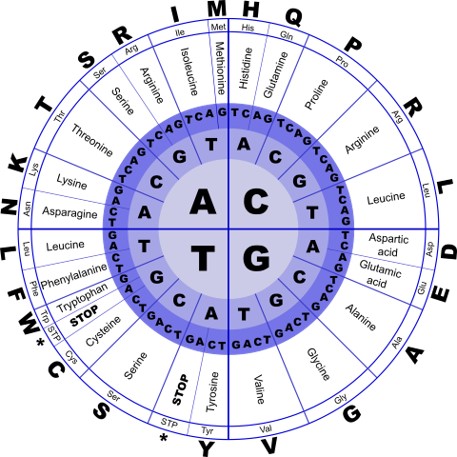
DNA (deoxyribonucleic acid) is the molecule that holds the heritable information passed from person to person. A DNA molecule is a pair of strands of polymers of nucleic acids. Nucleic acids have three main building blocks, an acidic phosphate group, a deoxyribose (pentagonal carbohydrate monomer with a missing oxygen atom), and a nitrogenous base. The sequence of nitrogenous bases are what hold the information and they come in pairs. A's on one strand correspond to T's on the other and vice-versa. C's on one strand correspond to G's on the other and vice-versa. The letters are abbreviations of the names of nitrogenous bases. The pairs together are called *base pairs* and are the genetic equivalent of a bit.



DNA is "read" by transcribing it to single stranded RNA which are fed to cell machinery called ribosomes. RNA trades out T for U and the ribose is not missing an oxygen atom. Ribosomes read the RNA 3 bases at a time and construct proteins by matching the trio with amino acids tied to their own RNA sequences.

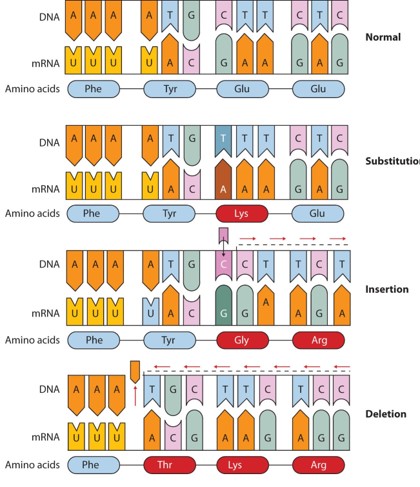


DNA base trios to their corresponding amino acids:



There are 64 possible base pair trios but only 20 amino acids used by humans so there is some redundancy built into the genetics system. Base pair trios are called *codons*, both in RNA and DNA.

Damage to DNA can cause mutations which can cause changes to what proteins are produced by a gene. Changes to proteins can cause changes to traits of a person but can also be the cause of genetic diseases. Substitutions are mutations where a single base pair is switched to another. That includes inversions where an A-T becomes a T-A pair. Substitutions can result in a single amino acid change in a protein, but because of the redundancy in the genetics system, often the substitution will do nothing. Insertions are where one or more base pairs get inserted into the sequence and deletions are where one or more base pairs are removed from a sequence. The consequences of insertions and deletions (indels) are reading frame shifts. This can completely change the amino acid sequence and protein length, it can introduce a new protein by creating a start codon, or it can disable a protein by deleting its start codon.



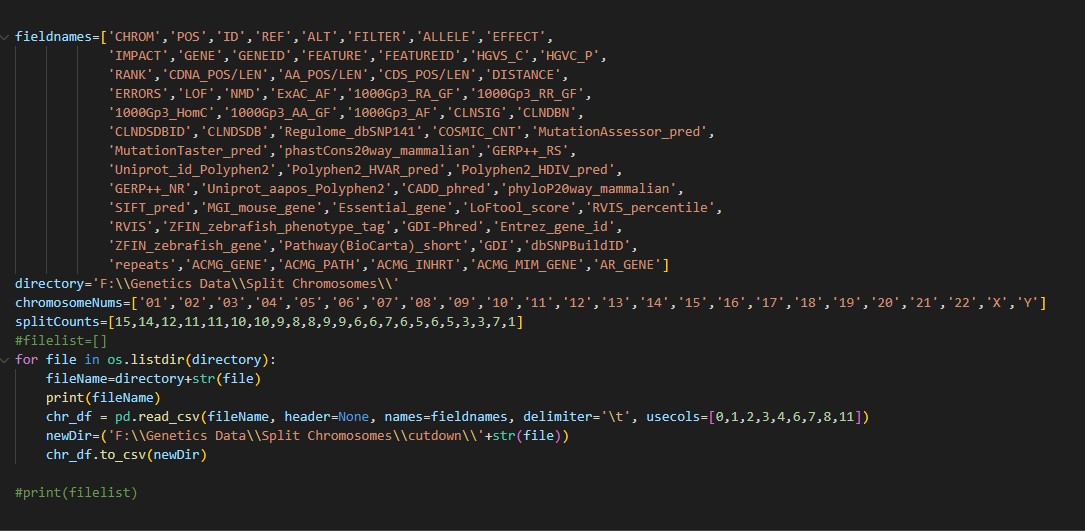
Much of a person's DNA is non-coding so the vast majority of mutations have little to no effect on a person.

## The Data

The 1000 Genomes project (<https://www.internationalgenome.org/>) has many genetics data sets. I used the "1000 Genomes 30x on GRCh38" dataset which is based on the Human Genome Project's (HGP) GRCh38 build of the human genome. GRCh38 is short for Genome Reference Consortium Human Build 38. The 1000 Genomes 30x on GRCh38 is an annotated dataset that contains a list of variants one each chromosome. A variant is any deviation from the sequence of a reference genome, in this case the Human Genome Project's GRCh38 build. The reference genome sequence is 3.1 GB (gigabytes, gigabases, either one works here) of G’s, A’s, T’s, and C’s. Annotating a whole genome would be many times more data so sequencing companies track the differences from the reference genomes and only annotate that. Even so, the annotated variants of the 1000 Genomes 30x on GRCh38 dataset consist of 24 text files (one for each human chromosome; 01-22, X, and Y) adding up to almost 60 GB (only gigabytes in this case) of data. In addition to the 1000 genomes data, I made a CSV from the Human Genome Project site listing how many base pairs each chromosome has for doing relative size calculations on each chromosome. (<https://www.ncbi.nlm.nih.gov/assembly/GCF_000001405.40/#/st_Primary-Assembly>).

## Summary of Initial Data Cleaning and Processing

Because the initial dataset was 60 GB, I had to do a considerable amount of data cleaning to make it usable. I opted to do this initial cleaning outside of a Jupyter notebook, mainly because I don't have the Jupyter and Python Venv set up on my desktop (I do all my school work on my laptop usually because I need it to be portable) and because my desktop hardware is heavy duty for gaming and photo processing.  
To start, I used "Text File Split", a free program on the Microsoft Store, to split each chromosome's .txt file into many smaller text files with 1,000,000 lines of data each. This allowed me to open the text files and see which columns were there and which columns I could remove. From there, I removed the columns I did not need (see screenshot below or gene\_data\_cleaning.py in the github repository). This reduced the 60 GB of data to 12 GB of data.



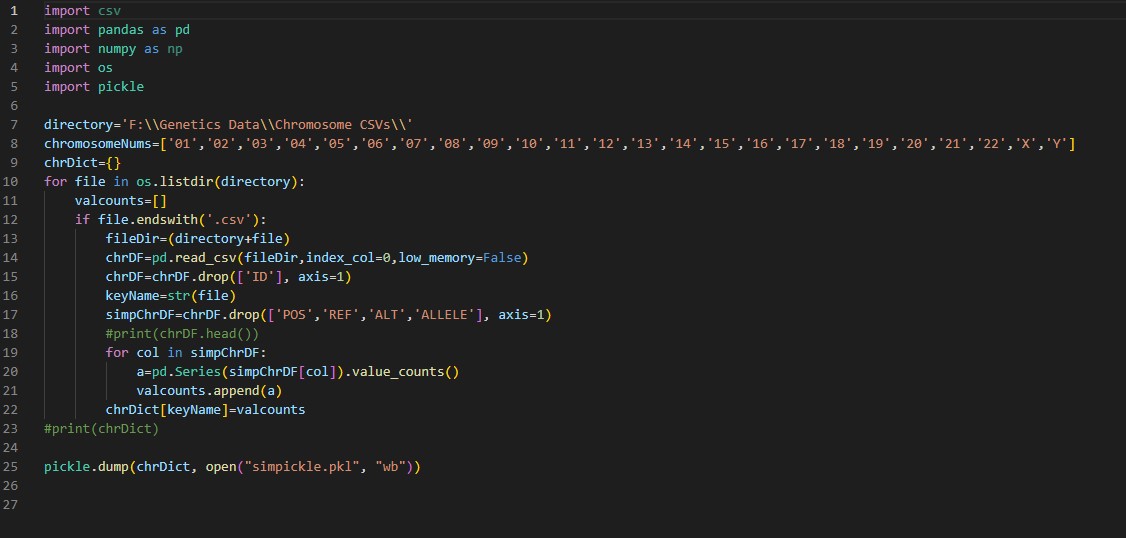
After that, I merged the separated text files back into 24 CSV files using the command line.

File merge commands in command line (NN is the chromosome number, 01-22 or X or Y):

cd <chromosomeNN text files directory>

copy \* chrNN.csv

From there, I loaded each chromosome csv into a dataframe and ran pd.value\_counts() on each column. I combined the value counts into a nested dictionary and saved that to a pickle file. Unfortunately the pickle file was still 3 GB so I had to drop all columns containing any GATC seqences (see screenshot below or gene\_data\_cleaning\_2.py in the github repository). I finally had a pickle file that was small enough (111 MB) for use in a Jupyter notebook.



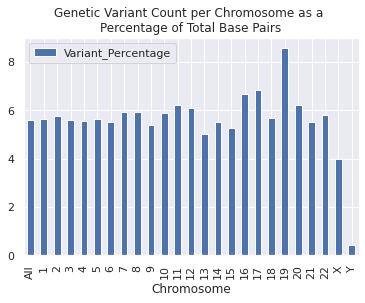
## Data Analysis

To start with, import packages and load in the pickle file, named simpickle.pkl because it is the simplified pickle file and I like portmanteaus.

!pip install matplotlib  
!pip install wordcloud  
import pandas as pd  
import numpy as np  
import os  
import pickle  
import matplotlib as mpl  
import matplotlib.pyplot as plt  
from wordcloud import WordCloud  
import seaborn as sns  
sns.set()

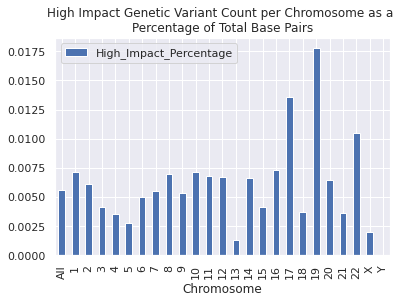
Code Output is in Appendix I

ax = totalCountsDF.plot.bar(x='Chromosome',y='Variant\_Percentage', rot=90, title="Genetic Variant Count per Chromosome as a \nPercentage of Total Base Pairs")



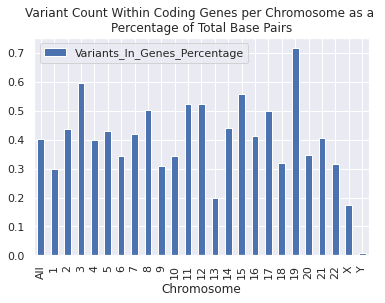
Chromosome 19 has the highest percentage of variants at 8 percent while Chromosome Y has the lowest at less than 1 percent. Most chromosomes have a variant count of about 6 percent. Chromosome X has a noticeably low variant percentage of 4 percent.

ax2 = totalCountsDF.plot.bar(x='Chromosome',y='High\_Impact\_Percentage', rot=90, title="High Impact Genetic Variant Count per Chromosome as a \nPercentage of Total Base Pairs")



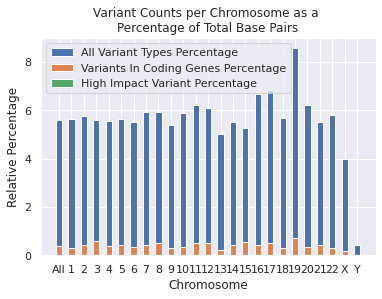
Chromosome 19 has the highest proportion of high impact variants at a little over 0.0175 percent. Most are close to either 0.0050 percent or 0.0030 percent. Chromosome 13 has noticeably low high impact variant percent at about 0.0012 percent, lower than even the X chromosome at around 0.0020 percent. Chromosomes 17 and 22 also have a greater proportion of high impact variants although not a great a proportion as Chromosome 19. Chromosome Y has a nearly 0 percentage of high impact variants.

ax3 = totalCountsDF.plot.bar(x='Chromosome',y='Variants\_In\_Genes\_Percentage', rot=90, title="Variant Count Within Coding Genes per Chromosome as a \nPercentage of Total Base Pairs")



Variant count within coding genes resembles the distribution of variants overall, with Chromosome 19 being the highest at about 0.7 percent and Chromosome Y being the lowest at about 0.01 percent. Chromosome 13 has a very small proportion of variants within coding genes at 0.2 percent, just above chromosome X this time which has a proportion of variants in coding genes of about 0.18 percent.

plt.bar(totalCountsDF['Chromosome'], totalCountsDF['Variant\_Percentage'], 0.5, label = 'All Variant Types Percentage')  
plt.bar(totalCountsDF['Chromosome'], totalCountsDF['Variants\_In\_Genes\_Percentage'], 0.5, label = 'Variants In Coding Genes Percentage')  
plt.bar(totalCountsDF['Chromosome'], totalCountsDF['High\_Impact\_Percentage'], 0.5, label = 'High Impact Variant Percentage')   
  
plt.xlabel("Chromosome")  
plt.ylabel("Relative Percentage")  
plt.title("Variant Counts per Chromosome as a \nPercentage of Total Base Pairs")  
plt.legend()  
plt.show()



The bar graph shows relative percentage of variant types and impact levels. It sums them so actual numbers on the Y-axis are not valuable, but the size of the bars are. As can be seen in the bar graph, a majority of variants occur outside of coding genes and the percentage of high impact variants is imperceptible relative to all variants and variants within coding genes.

## Conclusion

Chromosome 19 experiences the most variance of all the chromosomes, suggesting it is a good target for finding possible causes of genetic disease or predisposition to other types of diseases. Chromosome 13 has a low proportion of high impact variants suggesting it may contain genes that if mutated could result in lethal consequences for humans or human cells. The X and Y chromosomes experience a low proportion of variants of each type as well. This is likely because they encode the differences between female and male humans and errors there would result in infertile humans so those variants can't be passed on to the next generation. The Y Chromosome has an especially small proportion of variants of all types. This is likely because the Y Chromosome is so small, there is no room for non-lethal and defertilizing variants to exist.

According to the human genome project, 41% of human DNA is GC pairs and 59% is AT pairs. So I made that into a word cloud.

nbases={'A':59,'C':41,'G':41,'T':59}  
  
wordcloud = WordCloud().generate\_from\_frequencies(nbases)  
  
# Display the generated image:  
# the matplotlib way:  
import matplotlib.pyplot as plt  
plt.imshow(wordcloud, interpolation='bilinear')  
plt.axis("off")  
plt.show()



Appendix I: Code output

Requirement already satisfied: matplotlib in /mnt/c/users/imcpa/OneDrive/Documents/Data Science at DU/DSTools1/lib/python3.8/site-packages (3.5.2)  
Requirement already satisfied: pillow>=6.2.0 in /mnt/c/users/imcpa/OneDrive/Documents/Data Science at DU/DSTools1/lib/python3.8/site-packages (from matplotlib) (9.1.1)  
Requirement already satisfied: packaging>=20.0 in /mnt/c/users/imcpa/OneDrive/Documents/Data Science at DU/DSTools1/lib/python3.8/site-packages (from matplotlib) (21.3)  
Requirement already satisfied: python-dateutil>=2.7 in /mnt/c/users/imcpa/OneDrive/Documents/Data Science at DU/DSTools1/lib/python3.8/site-packages (from matplotlib) (2.8.2)  
Requirement already satisfied: numpy>=1.17 in /mnt/c/users/imcpa/OneDrive/Documents/Data Science at DU/DSTools1/lib/python3.8/site-packages (from matplotlib) (1.22.3)  
Requirement already satisfied: pyparsing>=2.2.1 in /mnt/c/users/imcpa/OneDrive/Documents/Data Science at DU/DSTools1/lib/python3.8/site-packages (from matplotlib) (3.0.8)  
Requirement already satisfied: fonttools>=4.22.0 in /mnt/c/users/imcpa/OneDrive/Documents/Data Science at DU/DSTools1/lib/python3.8/site-packages (from matplotlib) (4.33.3)  
Requirement already satisfied: cycler>=0.10 in /mnt/c/users/imcpa/OneDrive/Documents/Data Science at DU/DSTools1/lib/python3.8/site-packages (from matplotlib) (0.11.0)  
Requirement already satisfied: kiwisolver>=1.0.1 in /mnt/c/users/imcpa/OneDrive/Documents/Data Science at DU/DSTools1/lib/python3.8/site-packages (from matplotlib) (1.4.2)  
Requirement already satisfied: six>=1.5 in /mnt/c/users/imcpa/OneDrive/Documents/Data Science at DU/DSTools1/lib/python3.8/site-packages (from python-dateutil>=2.7->matplotlib) (1.16.0)  
WARNING: You are using pip version 22.0.4; however, version 22.1.2 is available.  
You should consider upgrading via the '/mnt/c/users/imcpa/OneDrive/Documents/Data Science at DU/DSTools1/bin/python3 -m pip install --upgrade pip' command.  
Requirement already satisfied: wordcloud in /mnt/c/users/imcpa/OneDrive/Documents/Data Science at DU/DSTools1/lib/python3.8/site-packages (1.8.1)  
Requirement already satisfied: numpy>=1.6.1 in /mnt/c/users/imcpa/OneDrive/Documents/Data Science at DU/DSTools1/lib/python3.8/site-packages (from wordcloud) (1.22.3)  
Requirement already satisfied: pillow in /mnt/c/users/imcpa/OneDrive/Documents/Data Science at DU/DSTools1/lib/python3.8/site-packages (from wordcloud) (9.1.1)  
Requirement already satisfied: matplotlib in /mnt/c/users/imcpa/OneDrive/Documents/Data Science at DU/DSTools1/lib/python3.8/site-packages (from wordcloud) (3.5.2)  
Requirement already satisfied: cycler>=0.10 in /mnt/c/users/imcpa/OneDrive/Documents/Data Science at DU/DSTools1/lib/python3.8/site-packages (from matplotlib->wordcloud) (0.11.0)  
Requirement already satisfied: kiwisolver>=1.0.1 in /mnt/c/users/imcpa/OneDrive/Documents/Data Science at DU/DSTools1/lib/python3.8/site-packages (from matplotlib->wordcloud) (1.4.2)  
Requirement already satisfied: packaging>=20.0 in /mnt/c/users/imcpa/OneDrive/Documents/Data Science at DU/DSTools1/lib/python3.8/site-packages (from matplotlib->wordcloud) (21.3)  
Requirement already satisfied: pyparsing>=2.2.1 in /mnt/c/users/imcpa/OneDrive/Documents/Data Science at DU/DSTools1/lib/python3.8/site-packages (from matplotlib->wordcloud) (3.0.8)  
Requirement already satisfied: fonttools>=4.22.0 in /mnt/c/users/imcpa/OneDrive/Documents/Data Science at DU/DSTools1/lib/python3.8/site-packages (from matplotlib->wordcloud) (4.33.3)  
Requirement already satisfied: python-dateutil>=2.7 in /mnt/c/users/imcpa/OneDrive/Documents/Data Science at DU/DSTools1/lib/python3.8/site-packages (from matplotlib->wordcloud) (2.8.2)  
Requirement already satisfied: six>=1.5 in /mnt/c/users/imcpa/OneDrive/Documents/Data Science at DU/DSTools1/lib/python3.8/site-packages (from python-dateutil>=2.7->matplotlib->wordcloud) (1.16.0)  
WARNING: You are using pip version 22.0.4; however, version 22.1.2 is available.  
You should consider upgrading via the '/mnt/c/users/imcpa/OneDrive/Documents/Data Science at DU/DSTools1/bin/python3 -m pip install --upgrade pip' command.

pickleData = pickle.load( open( "simpickle.pkl", "rb" ) )

print(pickleData)

{'chr01.csv': [chr1 14053427  
Name: CHROM, dtype: int64, intergenic\_region 5349797  
intron\_variant 5163578  
downstream\_gene\_variant 1202395  
upstream\_gene\_variant 1193612  
intragenic\_variant 749399  
 ...   
stop\_lost&disruptive\_inframe\_deletion 1  
start\_lost&conservative\_inframe\_insertion 1  
frameshift\_variant&start\_lost&splice\_region\_variant 1  
frameshift\_variant&stop\_gained&splice\_region\_variant 1  
stop\_lost&splice\_region\_variant&conservative\_inframe\_deletion 1  
Name: EFFECT, Length: 75, dtype: int64, MODIFIER 13903103  
MODERATE 74668  
LOW 57828  
HIGH 17828  
Name: IMPACT, dtype: int64, transcript 7942462  
intergenic\_region 5349797  
gene\_variant 749428  
interaction 11740  
Name: FEATURE, dtype: int64], 'chr02.csv': [chr2 13961577  
Name: CHROM, dtype: int64, intron\_variant 5432413  
intergenic\_region 5259484  
intragenic\_variant 1062786  
upstream\_gene\_variant 962662  
downstream\_gene\_variant 947611  
 ...   
splice\_donor\_variant&splice\_region\_variant&conservative\_inframe\_deletion&intron\_variant 1  
splice\_region\_variant&downstream\_gene\_variant 1  
splice\_acceptor\_variant&splice\_region\_variant&conservative\_inframe\_deletion&intron\_variant 1  
initiator\_codon\_variant&splice\_region\_variant 1  
start\_lost&conservative\_inframe\_insertion 1  
Name: EFFECT, Length: 69, dtype: int64, MODIFIER 13849129  
MODERATE 55176  
LOW 42398  
HIGH 14874  
Name: IMPACT, dtype: int64, transcript 7628815  
intergenic\_region 5259484  
gene\_variant 1062824  
interaction 10453  
transcript﻿ 1  
Name: FEATURE, dtype: int64], 'chr03.csv': [chr3 11149042  
Name: CHROM, dtype: int64, intron\_variant 4503237  
intergenic\_region 3861920  
intragenic\_variant 1184206  
downstream\_gene\_variant 691091  
upstream\_gene\_variant 682915  
 ...   
stop\_gained&splice\_region\_variant&disruptive\_inframe\_insertion 1  
splice\_acceptor\_variant&splice\_region\_variant&conservative\_inframe\_deletion&intron\_variant 1  
splice\_donor\_variant&5\_prime\_UTR\_variant&intron\_variant 1  
frameshift\_variant&splice\_acceptor\_variant&splice\_donor\_variant&splice\_region\_variant&intron\_variant 1  
stop\_lost&splice\_region\_variant 1  
Name: EFFECT, Length: 68, dtype: int64, MODIFIER 11069445  
MODERATE 40043  
LOW 31303  
HIGH 8251  
Name: IMPACT, dtype: int64, transcript 6097987  
intergenic\_region 3861920  
gene\_variant 1184227  
interaction 4908  
Name: FEATURE, dtype: int64], 'chr04.csv': [chr4 10591957  
Name: CHROM, dtype: int64, intergenic\_region 4737833  
intron\_variant 3687485  
intragenic\_variant 757502  
upstream\_gene\_variant 616037  
downstream\_gene\_variant 614436  
non\_coding\_transcript\_exon\_variant 68600  
3\_prime\_UTR\_variant 45328  
missense\_variant 25263  
synonymous\_variant 14997  
5\_prime\_UTR\_variant 9368  
structural\_interaction\_variant 4425  
splice\_region\_variant&intron\_variant 4151  
5\_prime\_UTR\_premature\_start\_codon\_gain\_variant 1265  
frameshift\_variant 921  
splice\_region\_variant&non\_coding\_transcript\_exon\_variant 751  
missense\_variant&splice\_region\_variant 669  
stop\_gained 586  
disruptive\_inframe\_deletion 396  
splice\_region\_variant&synonymous\_variant 382  
splice\_donor\_variant&intron\_variant 348  
splice\_acceptor\_variant&intron\_variant 330  
splice\_region\_variant 203  
disruptive\_inframe\_insertion 190  
conservative\_inframe\_deletion 153  
conservative\_inframe\_insertion 74  
start\_lost 47  
stop\_lost 25  
stop\_retained\_variant 22  
frameshift\_variant&splice\_region\_variant 21  
splice\_acceptor\_variant&splice\_region\_variant&intron\_variant&non\_coding\_transcript\_exon\_variant 17  
non\_coding\_transcript\_variant 17  
frameshift\_variant&stop\_gained 16  
stop\_gained&splice\_region\_variant 16  
splice\_donor\_variant&splice\_region\_variant&intron\_variant 11  
splice\_acceptor\_variant&splice\_region\_variant&intron\_variant 10  
frameshift\_variant&start\_lost 9  
splice\_region\_variant&disruptive\_inframe\_deletion 7  
bidirectional\_gene\_fusion 6  
splice\_donor\_variant&splice\_region\_variant&intron\_variant&non\_coding\_transcript\_exon\_variant 6  
frameshift\_variant&splice\_acceptor\_variant&splice\_region\_variant&intron\_variant 4  
frameshift\_variant&stop\_lost 3  
frameshift\_variant&splice\_donor\_variant&splice\_region\_variant&intron\_variant 3  
initiator\_codon\_variant 3  
protein\_protein\_contact 3  
stop\_lost&conservative\_inframe\_deletion 3  
start\_lost&splice\_region\_variant 2  
stop\_lost&splice\_region\_variant 2  
stop\_gained&disruptive\_inframe\_insertion 2  
stop\_gained&conservative\_inframe\_insertion 2  
initiator\_codon\_variant&non\_canonical\_start\_codon 1  
splice\_acceptor\_variant&splice\_region\_variant&disruptive\_inframe\_deletion&intron\_variant 1  
frameshift\_variant&stop\_lost&splice\_region\_variant 1  
frameshift\_variant&start\_lost&splice\_region\_variant 1  
gene\_fusion 1  
splice\_acceptor\_variant&splice\_region\_variant&conservative\_inframe\_deletion&intron\_variant 1  
splice\_region\_variant&conservative\_inframe\_deletion 1  
Name: EFFECT, dtype: int64, MODIFIER 10536606  
MODERATE 26753  
LOW 21775  
HIGH 6823  
Name: IMPACT, dtype: int64, transcript 5092187  
intergenic\_region 4737833  
gene\_variant 757509  
interaction 4428  
Name: FEATURE, dtype: int64], 'chr05.csv': [chr5 10283606  
Name: CHROM, dtype: int64, intergenic\_region 4173842  
intron\_variant 3803902  
intragenic\_variant 783880  
upstream\_gene\_variant 661815  
downstream\_gene\_variant 660644  
 ...   
stop\_lost&splice\_region\_variant 1  
splice\_acceptor\_variant&5\_prime\_UTR\_variant&intron\_variant 1  
splice\_acceptor\_variant&splice\_region\_variant&disruptive\_inframe\_deletion&intron\_variant 1  
stop\_gained&conservative\_inframe\_insertion 1  
start\_lost&conservative\_inframe\_deletion 1  
Name: EFFECT, Length: 62, dtype: int64, MODIFIER 10221746  
MODERATE 30884  
LOW 25914  
HIGH 5062  
Name: IMPACT, dtype: int64, transcript 5323097  
intergenic\_region 4173842  
gene\_variant 783940  
interaction 2727  
Name: FEATURE, dtype: int64], 'chr06.csv': [chr6 9430386  
Name: CHROM, dtype: int64, intergenic\_region 4034763  
intron\_variant 3277205  
downstream\_gene\_variant 662829  
upstream\_gene\_variant 661989  
intragenic\_variant 590030  
 ...   
splice\_donor\_variant&splice\_region\_variant&conservative\_inframe\_deletion&intron\_variant 1  
splice\_acceptor\_variant&splice\_region\_variant&5\_prime\_UTR\_variant&intron\_variant 1  
gene\_fusion 1  
splice\_donor\_variant&splice\_region\_variant&disruptive\_inframe\_deletion&intron\_variant 1  
stop\_lost&splice\_region\_variant&conservative\_inframe\_deletion 1  
Name: EFFECT, Length: 62, dtype: int64, MODIFIER 9358625  
MODERATE 35471  
LOW 27715  
HIGH 8575  
Name: IMPACT, dtype: int64, transcript 4799901  
intergenic\_region 4034763  
gene\_variant 590047  
interaction 5675  
Name: FEATURE, dtype: int64], 'chr07.csv': [chr7 9468709  
Name: CHROM, dtype: int64, intron\_variant 3781258  
intergenic\_region 3391514  
downstream\_gene\_variant 705639  
upstream\_gene\_variant 697360  
intragenic\_variant 670119  
 ...   
frameshift\_variant&splice\_acceptor\_variant&splice\_donor\_variant&splice\_region\_variant&intron\_variant 1  
start\_lost&disruptive\_inframe\_insertion 1  
splice\_region\_variant&conservative\_inframe\_insertion 1  
stop\_gained&disruptive\_inframe\_deletion 1  
splice\_region\_variant&conservative\_inframe\_deletion 1  
Name: EFFECT, Length: 67, dtype: int64, MODIFIER 9391860  
MODERATE 37836  
LOW 30283  
HIGH 8730  
Name: IMPACT, dtype: int64, transcript 5401588  
intergenic\_region 3391514  
gene\_variant 670150  
interaction 5457  
Name: FEATURE, dtype: int64], 'chr08.csv': [chr8 8614681  
Name: CHROM, dtype: int64, intron\_variant 3376717  
intergenic\_region 3150101  
intragenic\_variant 727946  
downstream\_gene\_variant 596351  
upstream\_gene\_variant 588630  
 ...   
5\_prime\_UTR\_truncation&exon\_loss\_variant 1  
splice\_acceptor\_variant&5\_prime\_UTR\_variant&intron\_variant 1  
start\_lost&conservative\_inframe\_insertion 1  
frameshift\_variant&stop\_lost&splice\_region\_variant 1  
splice\_acceptor\_variant&splice\_region\_variant&conservative\_inframe\_deletion&intron\_variant 1  
Name: EFFECT, Length: 63, dtype: int64, MODIFIER 8557239  
MODERATE 25325  
LOW 21965  
HIGH 10152  
Name: IMPACT, dtype: int64, transcript 4728458  
intergenic\_region 3150101  
gene\_variant 727985  
interaction 8137  
Name: FEATURE, dtype: int64], 'chr09.csv': [chr9 7462036  
Name: CHROM, dtype: int64, intergenic\_region 3183293  
intron\_variant 2534950  
downstream\_gene\_variant 570459  
upstream\_gene\_variant 566665  
intragenic\_variant 430236  
 ...   
splice\_acceptor\_variant&splice\_region\_variant&conservative\_inframe\_deletion&intron\_variant 1  
splice\_donor\_variant&splice\_region\_variant&disruptive\_inframe\_deletion&intron\_variant 1  
splice\_donor\_variant&splice\_region\_variant&5\_prime\_UTR\_variant&intron\_variant 1  
stop\_lost&splice\_region\_variant 1  
stop\_lost&splice\_region\_variant&conservative\_inframe\_deletion 1  
Name: EFFECT, Length: 64, dtype: int64, MODIFIER 7398898  
MODERATE 30347  
LOW 25319  
HIGH 7472  
Name: IMPACT, dtype: int64, transcript 3843322  
intergenic\_region 3183293  
gene\_variant 430246  
interaction 5175  
Name: FEATURE, dtype: int64], 'chr10.csv': [chr10 7881973  
Name: CHROM, dtype: int64, intergenic\_region 3195460  
intron\_variant 3000050  
downstream\_gene\_variant 528778  
upstream\_gene\_variant 527086  
intragenic\_variant 459411  
 ...   
splice\_acceptor\_variant&splice\_region\_variant&disruptive\_inframe\_deletion&intron\_variant 1  
stop\_lost&splice\_region\_variant 1  
start\_lost&splice\_region\_variant&conservative\_inframe\_deletion 1  
splice\_donor\_variant&splice\_region\_variant&disruptive\_inframe\_deletion&intron\_variant 1  
splice\_donor\_variant&splice\_region\_variant&3\_prime\_UTR\_variant&intron\_variant 1  
Name: EFFECT, Length: 66, dtype: int64, MODIFIER 7822173  
MODERATE 27443  
LOW 22848  
HIGH 9509  
Name: IMPACT, dtype: int64, transcript 4219694  
intergenic\_region 3195460  
gene\_variant 459432  
interaction 7387  
Name: FEATURE, dtype: int64], 'chr11.csv': [chr11 8401457  
Name: CHROM, dtype: int64, intergenic\_region 3015909  
intron\_variant 2906571  
downstream\_gene\_variant 770821  
upstream\_gene\_variant 769794  
intragenic\_variant 709370  
 ...   
start\_lost&conservative\_inframe\_deletion 1  
5\_prime\_UTR\_truncation&exon\_loss\_variant 1  
splice\_donor\_variant&3\_prime\_UTR\_variant&intron\_variant 1  
start\_lost&conservative\_inframe\_insertion 1  
splice\_region\_variant&downstream\_gene\_variant 1  
Name: EFFECT, Length: 68, dtype: int64, MODIFIER 8311419  
MODERATE 45641  
LOW 35151  
HIGH 9246  
Name: IMPACT, dtype: int64, transcript 4670462  
intergenic\_region 3015909  
gene\_variant 709409  
interaction 5677  
Name: FEATURE, dtype: int64], 'chr12.csv': [chr12 8130250  
Name: CHROM, dtype: int64, intron\_variant 3079886  
intergenic\_region 2756441  
intragenic\_variant 700417  
downstream\_gene\_variant 683192  
upstream\_gene\_variant 680918  
 ...   
splice\_donor\_variant&splice\_region\_variant&conservative\_inframe\_deletion&intron\_variant 1  
transcript\_ablation 1  
splice\_acceptor\_variant&splice\_donor\_variant&splice\_region\_variant&conservative\_inframe\_deletion&intron\_variant 1  
splice\_acceptor\_variant&splice\_donor\_variant&intron\_variant 1  
start\_lost&conservative\_inframe\_deletion 1  
Name: EFFECT, Length: 66, dtype: int64, MODIFIER 8058186  
MODERATE 33759  
LOW 29322  
HIGH 8983  
Name: IMPACT, dtype: int64, transcript 4667358  
intergenic\_region 2756441  
gene\_variant 700442  
interaction 6009  
Name: FEATURE, dtype: int64], 'chr13.csv': [chr13 5756600  
Name: CHROM, dtype: int64, intergenic\_region 3089603  
intron\_variant 1731958  
upstream\_gene\_variant 310203  
downstream\_gene\_variant 307622  
intragenic\_variant 227650  
non\_coding\_transcript\_exon\_variant 39441  
3\_prime\_UTR\_variant 22383  
missense\_variant 11007  
synonymous\_variant 6954  
5\_prime\_UTR\_variant 4160  
splice\_region\_variant&intron\_variant 2118  
5\_prime\_UTR\_premature\_start\_codon\_gain\_variant 619  
splice\_region\_variant&non\_coding\_transcript\_exon\_variant 421  
structural\_interaction\_variant 402  
frameshift\_variant 318  
missense\_variant&splice\_region\_variant 275  
stop\_gained 242  
splice\_donor\_variant&intron\_variant 206  
splice\_acceptor\_variant&intron\_variant 204  
disruptive\_inframe\_deletion 185  
splice\_region\_variant&synonymous\_variant 155  
splice\_region\_variant 90  
conservative\_inframe\_deletion 86  
disruptive\_inframe\_insertion 76  
conservative\_inframe\_insertion 63  
protein\_protein\_contact 24  
start\_lost 23  
frameshift\_variant&splice\_region\_variant 18  
non\_coding\_transcript\_variant 16  
splice\_acceptor\_variant&splice\_region\_variant&intron\_variant 11  
stop\_gained&splice\_region\_variant 8  
splice\_acceptor\_variant&splice\_region\_variant&intron\_variant&non\_coding\_transcript\_exon\_variant 8  
stop\_lost 7  
splice\_donor\_variant&splice\_region\_variant&intron\_variant&non\_coding\_transcript\_exon\_variant 6  
bidirectional\_gene\_fusion 5  
frameshift\_variant&stop\_gained 5  
frameshift\_variant&splice\_donor\_variant&splice\_region\_variant&intron\_variant 4  
splice\_donor\_variant&splice\_region\_variant&intron\_variant 4  
gene\_fusion 3  
frameshift\_variant&splice\_acceptor\_variant&splice\_region\_variant&intron\_variant 3  
stop\_retained\_variant 3  
initiator\_codon\_variant 2  
frameshift\_variant&start\_lost 2  
splice\_donor\_variant&5\_prime\_UTR\_variant&intron\_variant 1  
splice\_region\_variant&conservative\_inframe\_deletion 1  
splice\_region\_variant&disruptive\_inframe\_insertion 1  
start\_lost&conservative\_inframe\_deletion 1  
start\_lost&splice\_region\_variant 1  
splice\_donor\_variant&splice\_region\_variant&disruptive\_inframe\_deletion&intron\_variant 1  
splice\_acceptor\_variant&splice\_region\_variant&5\_prime\_UTR\_variant&intron\_variant 1  
Name: EFFECT, dtype: int64, MODIFIER 5733036  
MODERATE 11694  
LOW 10362  
HIGH 1508  
Name: IMPACT, dtype: int64, intergenic\_region 3089603  
transcript 2438913  
gene\_variant 227658  
interaction 426  
Name: FEATURE, dtype: int64], 'chr14.csv': [chr14 5948454  
Name: CHROM, dtype: int64, intron\_variant 2073750  
intergenic\_region 1807833  
downstream\_gene\_variant 731348  
upstream\_gene\_variant 694346  
intragenic\_variant 477909  
 ...   
splice\_region\_variant&stop\_retained\_variant 1  
splice\_donor\_variant&splice\_region\_variant&5\_prime\_UTR\_variant&intron\_variant 1  
exon\_loss\_variant&splice\_region\_variant 1  
stop\_lost&conservative\_inframe\_insertion 1  
splice\_donor\_variant&splice\_region\_variant&conservative\_inframe\_deletion&intron\_variant 1  
Name: EFFECT, Length: 73, dtype: int64, MODIFIER 5883699  
MODERATE 33778  
LOW 23844  
HIGH 7133  
Name: IMPACT, dtype: int64, transcript 3659050  
intergenic\_region 1807833  
gene\_variant 477927  
interaction 3644  
Name: FEATURE, dtype: int64], 'chr15.csv': [chr15 5379964  
Name: CHROM, dtype: int64, intron\_variant 2272152  
intergenic\_region 1311179  
intragenic\_variant 573350  
downstream\_gene\_variant 527253  
upstream\_gene\_variant 522776  
non\_coding\_transcript\_exon\_variant 73026  
3\_prime\_UTR\_variant 42899  
missense\_variant 23729  
synonymous\_variant 14208  
5\_prime\_UTR\_variant 7131  
splice\_region\_variant&intron\_variant 4122  
structural\_interaction\_variant 2141  
5\_prime\_UTR\_premature\_start\_codon\_gain\_variant 1001  
splice\_region\_variant&non\_coding\_transcript\_exon\_variant 791  
frameshift\_variant 759  
missense\_variant&splice\_region\_variant 665  
stop\_gained 481  
splice\_donor\_variant&intron\_variant 372  
splice\_region\_variant&synonymous\_variant 371  
disruptive\_inframe\_deletion 363  
splice\_acceptor\_variant&intron\_variant 295  
conservative\_inframe\_deletion 218  
splice\_region\_variant 143  
disruptive\_inframe\_insertion 130  
conservative\_inframe\_insertion 120  
start\_lost 35  
frameshift\_variant&stop\_gained 23  
stop\_lost 23  
non\_coding\_transcript\_variant 22  
frameshift\_variant&splice\_region\_variant 18  
protein\_protein\_contact 18  
stop\_gained&splice\_region\_variant 17  
frameshift\_variant&splice\_donor\_variant&splice\_region\_variant&intron\_variant 15  
splice\_donor\_variant&splice\_region\_variant&intron\_variant&non\_coding\_transcript\_exon\_variant 15  
bidirectional\_gene\_fusion 14  
stop\_retained\_variant 12  
splice\_acceptor\_variant&splice\_region\_variant&intron\_variant 10  
splice\_donor\_variant&splice\_region\_variant&intron\_variant 8  
gene\_fusion 8  
splice\_acceptor\_variant&splice\_region\_variant&intron\_variant&non\_coding\_transcript\_exon\_variant 7  
splice\_region\_variant&conservative\_inframe\_deletion 5  
frameshift\_variant&start\_lost 5  
splice\_region\_variant&disruptive\_inframe\_deletion 4  
stop\_lost&splice\_region\_variant 4  
initiator\_codon\_variant 4  
frameshift\_variant&splice\_acceptor\_variant&splice\_region\_variant&intron\_variant 4  
start\_lost&conservative\_inframe\_deletion 3  
splice\_acceptor\_variant&splice\_region\_variant&conservative\_inframe\_deletion&intron\_variant 3  
frameshift\_variant&stop\_lost&splice\_region\_variant 2  
splice\_acceptor\_variant&splice\_region\_variant&5\_prime\_UTR\_variant&intron\_variant 1  
splice\_region\_variant&stop\_retained\_variant 1  
stop\_gained&disruptive\_inframe\_deletion 1  
splice\_acceptor\_variant&splice\_donor\_variant&splice\_region\_variant&intron\_variant 1  
frameshift\_variant&splice\_acceptor\_variant&splice\_donor\_variant&splice\_region\_variant&intron\_variant 1  
splice\_acceptor\_variant&3\_prime\_UTR\_variant&intron\_variant 1  
start\_lost&splice\_region\_variant 1  
splice\_donor\_variant&splice\_region\_variant&conservative\_inframe\_deletion&intron\_variant 1  
initiator\_codon\_variant&splice\_region\_variant 1  
frameshift\_variant&stop\_lost 1  
Name: EFFECT, dtype: int64, MODIFIER 5329788  
MODERATE 25234  
LOW 20654  
HIGH 4288  
Name: IMPACT, dtype: int64, transcript 3493254  
intergenic\_region 1311179  
gene\_variant 573372  
interaction 2159  
Name: FEATURE, dtype: int64], 'chr16.csv': [chr16 6140757  
Name: CHROM, dtype: int64, intron\_variant 2484202  
intergenic\_region 1714259  
downstream\_gene\_variant 673882  
upstream\_gene\_variant 671319  
intragenic\_variant 381783  
 ...   
stop\_gained&disruptive\_inframe\_insertion 1  
stop\_gained&splice\_donor\_variant&splice\_region\_variant&disruptive\_inframe\_deletion&intron\_variant 1  
stop\_lost&disruptive\_inframe\_insertion 1  
splice\_region\_variant&stop\_retained\_variant 1  
splice\_acceptor\_variant&splice\_donor\_variant&splice\_region\_variant&intron\_variant&non\_coding\_transcript\_exon\_variant 1  
Name: EFFECT, Length: 75, dtype: int64, MODIFIER 6067639  
MODERATE 35695  
LOW 30641  
HIGH 6782  
Name: IMPACT, dtype: int64, transcript 4040642  
intergenic\_region 1714259  
gene\_variant 381801  
interaction 4055  
Name: FEATURE, dtype: int64], 'chr17.csv': [chr17 5738905  
Name: CHROM, dtype: int64, intron\_variant 2180550  
intergenic\_region 1453082  
downstream\_gene\_variant 728046  
upstream\_gene\_variant 724892  
intragenic\_variant 418494  
 ...   
splice\_donor\_variant&splice\_region\_variant&3\_prime\_UTR\_variant&intron\_variant 1  
splice\_region\_variant&downstream\_gene\_variant 1  
splice\_region\_variant&stop\_retained\_variant 1  
splice\_acceptor\_variant&splice\_region\_variant&conservative\_inframe\_deletion&intron\_variant 1  
splice\_acceptor\_variant&splice\_donor\_variant&splice\_region\_variant&intron\_variant 1  
Name: EFFECT, Length: 68, dtype: int64, MODIFIER 5651017  
MODERATE 41099  
LOW 35384  
HIGH 11405  
Name: IMPACT, dtype: int64, transcript 3859221  
intergenic\_region 1453082  
gene\_variant 418544  
interaction 8058  
Name: FEATURE, dtype: int64], 'chr18.csv': [chr18 4572260  
Name: CHROM, dtype: int64, intergenic\_region 2065539  
intron\_variant 1595926  
downstream\_gene\_variant 284864  
upstream\_gene\_variant 282536  
intragenic\_variant 256058  
non\_coding\_transcript\_exon\_variant 36183  
3\_prime\_UTR\_variant 24597  
missense\_variant 10005  
synonymous\_variant 6372  
5\_prime\_UTR\_variant 3823  
structural\_interaction\_variant 2058  
splice\_region\_variant&intron\_variant 1702  
5\_prime\_UTR\_premature\_start\_codon\_gain\_variant 506  
splice\_region\_variant&non\_coding\_transcript\_exon\_variant 389  
frameshift\_variant 299  
missense\_variant&splice\_region\_variant 244  
stop\_gained 193  
splice\_donor\_variant&intron\_variant 164  
splice\_region\_variant&synonymous\_variant 155  
splice\_acceptor\_variant&intron\_variant 131  
disruptive\_inframe\_deletion 129  
splice\_region\_variant 79  
conservative\_inframe\_deletion 64  
disruptive\_inframe\_insertion 61  
conservative\_inframe\_insertion 44  
non\_coding\_transcript\_variant 22  
start\_lost 22  
protein\_protein\_contact 14  
frameshift\_variant&splice\_region\_variant 9  
splice\_donor\_variant&splice\_region\_variant&intron\_variant&non\_coding\_transcript\_exon\_variant 9  
frameshift\_variant&stop\_gained 6  
splice\_acceptor\_variant&splice\_region\_variant&intron\_variant&non\_coding\_transcript\_exon\_variant 6  
frameshift\_variant&splice\_donor\_variant&splice\_region\_variant&intron\_variant 6  
stop\_gained&splice\_region\_variant 6  
splice\_donor\_variant&splice\_region\_variant&intron\_variant 5  
bidirectional\_gene\_fusion 5  
stop\_lost 4  
gene\_fusion 4  
initiator\_codon\_variant 4  
splice\_acceptor\_variant&splice\_region\_variant&disruptive\_inframe\_deletion&intron\_variant 2  
stop\_retained\_variant 2  
frameshift\_variant&start\_lost 2  
frameshift\_variant&splice\_acceptor\_variant&splice\_region\_variant&intron\_variant 2  
splice\_acceptor\_variant&splice\_region\_variant&intron\_variant 2  
stop\_gained&disruptive\_inframe\_insertion 1  
splice\_donor\_variant&splice\_region\_variant&conservative\_inframe\_deletion&intron\_variant 1  
splice\_region\_variant&conservative\_inframe\_deletion 1  
stop\_lost&splice\_region\_variant 1  
start\_lost&conservative\_inframe\_insertion 1  
splice\_region\_variant&downstream\_gene\_variant 1  
splice\_acceptor\_variant&splice\_donor\_variant&splice\_region\_variant&intron\_variant 1  
Name: EFFECT, dtype: int64, MODIFIER 4549548  
MODERATE 10548  
LOW 9210  
HIGH 2954  
Name: IMPACT, dtype: int64, transcript 2248582  
intergenic\_region 2065539  
gene\_variant 256067  
interaction 2072  
Name: FEATURE, dtype: int64], 'chr19.csv': [chr19 5025452  
Name: CHROM, dtype: int64, intron\_variant 1739413  
intergenic\_region 1018240  
downstream\_gene\_variant 806542  
upstream\_gene\_variant 793569  
intragenic\_variant 420258  
 ...   
splice\_acceptor\_variant&5\_prime\_UTR\_variant&intron\_variant 1  
start\_lost&splice\_region\_variant&conservative\_inframe\_deletion 1  
start\_lost&conservative\_inframe\_insertion 1  
start\_lost&conservative\_inframe\_deletion 1  
splice\_donor\_variant&5\_prime\_UTR\_variant&intron\_variant 1  
Name: EFFECT, Length: 74, dtype: int64, MODIFIER 4912823  
MODERATE 56200  
LOW 46014  
HIGH 10415  
Name: IMPACT, dtype: int64, transcript 3581346  
intergenic\_region 1018240  
gene\_variant 420303  
interaction 5563  
Name: FEATURE, dtype: int64], 'chr20.csv': [chr20 4021075  
Name: CHROM, dtype: int64, intergenic\_region 1686894  
intron\_variant 1319291  
downstream\_gene\_variant 345250  
upstream\_gene\_variant 341023  
intragenic\_variant 223942  
non\_coding\_transcript\_exon\_variant 31041  
3\_prime\_UTR\_variant 30776  
missense\_variant 16325  
synonymous\_variant 11090  
5\_prime\_UTR\_variant 5980  
structural\_interaction\_variant 2844  
splice\_region\_variant&intron\_variant 2651  
5\_prime\_UTR\_premature\_start\_codon\_gain\_variant 842  
frameshift\_variant 446  
missense\_variant&splice\_region\_variant 439  
splice\_region\_variant&non\_coding\_transcript\_exon\_variant 390  
splice\_region\_variant&synonymous\_variant 303  
stop\_gained 282  
disruptive\_inframe\_deletion 246  
splice\_donor\_variant&intron\_variant 213  
splice\_acceptor\_variant&intron\_variant 192  
splice\_region\_variant 138  
disruptive\_inframe\_insertion 107  
conservative\_inframe\_deletion 104  
conservative\_inframe\_insertion 72  
start\_lost 34  
protein\_protein\_contact 29  
splice\_acceptor\_variant&splice\_region\_variant&intron\_variant 15  
non\_coding\_transcript\_variant 11  
stop\_lost 11  
gene\_fusion 11  
splice\_donor\_variant&splice\_region\_variant&intron\_variant&non\_coding\_transcript\_exon\_variant 8  
frameshift\_variant&stop\_gained 8  
stop\_gained&splice\_region\_variant 7  
initiator\_codon\_variant 6  
frameshift\_variant&start\_lost 5  
frameshift\_variant&splice\_region\_variant 5  
splice\_acceptor\_variant&splice\_region\_variant&intron\_variant&non\_coding\_transcript\_exon\_variant 5  
stop\_retained\_variant 4  
splice\_region\_variant&disruptive\_inframe\_deletion 4  
frameshift\_variant&splice\_acceptor\_variant&splice\_region\_variant&intron\_variant 4  
splice\_donor\_variant&splice\_region\_variant&intron\_variant 3  
start\_lost&conservative\_inframe\_deletion 3  
frameshift\_variant&stop\_lost 3  
stop\_lost&splice\_region\_variant 3  
stop\_gained&conservative\_inframe\_insertion 2  
frameshift\_variant&splice\_donor\_variant&splice\_region\_variant&intron\_variant 2  
frameshift\_variant&stop\_gained&splice\_region\_variant 2  
splice\_acceptor\_variant&5\_prime\_UTR\_truncation&exon\_loss\_variant&splice\_region\_variant&intron\_variant 1  
start\_lost&splice\_region\_variant 1  
feature\_ablation 1  
bidirectional\_gene\_fusion 1  
splice\_region\_variant&conservative\_inframe\_deletion 1  
splice\_acceptor\_variant&splice\_region\_variant&disruptive\_inframe\_deletion&intron\_variant 1  
frameshift\_variant&stop\_lost&splice\_region\_variant 1  
frameshift\_variant&start\_lost&splice\_region\_variant 1  
splice\_donor\_variant&splice\_region\_variant&disruptive\_inframe\_deletion&intron\_variant 1  
Name: EFFECT, dtype: int64, MODIFIER 3984208  
MODERATE 17298  
LOW 15424  
HIGH 4145  
Name: IMPACT, dtype: int64, transcript 2107353  
intergenic\_region 1686894  
gene\_variant 223955  
interaction 2873  
Name: FEATURE, dtype: int64], 'chr21.csv': [chr21 2586151  
Name: CHROM, dtype: int64, intergenic\_region 1037051  
intron\_variant 867433  
downstream\_gene\_variant 221920  
upstream\_gene\_variant 213186  
intragenic\_variant 189527  
non\_coding\_transcript\_exon\_variant 25845  
3\_prime\_UTR\_variant 12250  
missense\_variant 7189  
synonymous\_variant 4884  
5\_prime\_UTR\_variant 2431  
splice\_region\_variant&intron\_variant 1400  
structural\_interaction\_variant 957  
splice\_region\_variant&non\_coding\_transcript\_exon\_variant 357  
5\_prime\_UTR\_premature\_start\_codon\_gain\_variant 336  
frameshift\_variant 281  
missense\_variant&splice\_region\_variant 197  
stop\_gained 138  
splice\_donor\_variant&intron\_variant 135  
splice\_region\_variant&synonymous\_variant 117  
disruptive\_inframe\_deletion 117  
splice\_acceptor\_variant&intron\_variant 101  
splice\_region\_variant 65  
conservative\_inframe\_deletion 48  
disruptive\_inframe\_insertion 46  
conservative\_inframe\_insertion 42  
start\_lost 16  
splice\_donor\_variant&splice\_region\_variant&intron\_variant&non\_coding\_transcript\_exon\_variant 10  
non\_coding\_transcript\_variant 8  
splice\_acceptor\_variant&splice\_region\_variant&intron\_variant 7  
frameshift\_variant&splice\_region\_variant 7  
frameshift\_variant&stop\_gained 6  
stop\_retained\_variant 5  
gene\_fusion 5  
splice\_acceptor\_variant&splice\_region\_variant&intron\_variant&non\_coding\_transcript\_exon\_variant 4  
frameshift\_variant&start\_lost 3  
splice\_donor\_variant&splice\_region\_variant&intron\_variant 3  
stop\_lost&conservative\_inframe\_deletion 2  
stop\_gained&splice\_region\_variant 2  
splice\_donor\_variant&splice\_region\_variant&5\_prime\_UTR\_variant&intron\_variant 2  
splice\_acceptor\_variant&splice\_region\_variant&conservative\_inframe\_deletion&intron\_variant 2  
stop\_lost 2  
bidirectional\_gene\_fusion 2  
initiator\_codon\_variant 1  
frameshift\_variant&splice\_donor\_variant&splice\_region\_variant&intron\_variant 1  
splice\_region\_variant&disruptive\_inframe\_deletion 1  
splice\_donor\_variant&splice\_region\_variant&conservative\_inframe\_deletion&intron\_variant 1  
frameshift\_variant&splice\_acceptor\_variant&splice\_region\_variant&intron\_variant 1  
splice\_region\_variant&conservative\_inframe\_deletion 1  
splice\_region\_variant&disruptive\_inframe\_insertion 1  
protein\_protein\_contact 1  
stop\_gained&disruptive\_inframe\_insertion 1  
stop\_gained&conservative\_inframe\_insertion 1  
frameshift\_variant&stop\_lost 1  
start\_lost&conservative\_inframe\_deletion 1  
Name: EFFECT, dtype: int64, MODIFIER 2569651  
MODERATE 7642  
LOW 7165  
HIGH 1693  
Name: IMPACT, dtype: int64, transcript 1358608  
intergenic\_region 1037051  
gene\_variant 189534  
interaction 958  
Name: FEATURE, dtype: int64], 'chr22.csv': [chr22 2993743  
Name: CHROM, dtype: int64, intron\_variant 1001421  
intergenic\_region 928212  
downstream\_gene\_variant 377965  
upstream\_gene\_variant 376993  
intragenic\_variant 164170  
 ...   
splice\_donor\_variant&splice\_region\_variant&conservative\_inframe\_deletion&intron\_variant 1  
stop\_lost&splice\_region\_variant 1  
splice\_donor\_variant&splice\_region\_variant&5\_prime\_UTR\_variant&intron\_variant 1  
splice\_donor\_variant&splice\_region\_variant&disruptive\_inframe\_deletion&intron\_variant 1  
splice\_acceptor\_variant&splice\_donor\_variant&splice\_region\_variant&intron\_variant 1  
Name: EFFECT, Length: 63, dtype: int64, MODIFIER 2948947  
MODERATE 22073  
LOW 17288  
HIGH 5435  
Name: IMPACT, dtype: int64, transcript 1898070  
intergenic\_region 928212  
gene\_variant 164182  
interaction 3279  
Name: FEATURE, dtype: int64], 'chrX.csv': [chrX 6214235  
Name: CHROM, dtype: int64, intergenic\_region 3393167  
intron\_variant 1660210  
downstream\_gene\_variant 394677  
upstream\_gene\_variant 390720  
intragenic\_variant 270389  
non\_coding\_transcript\_exon\_variant 36330  
3\_prime\_UTR\_variant 27767  
missense\_variant 15483  
synonymous\_variant 10332  
5\_prime\_UTR\_variant 6954  
splice\_region\_variant&intron\_variant 2219  
structural\_interaction\_variant 2016  
5\_prime\_UTR\_premature\_start\_codon\_gain\_variant 975  
frameshift\_variant 384  
splice\_region\_variant&non\_coding\_transcript\_exon\_variant 369  
disruptive\_inframe\_deletion 351  
missense\_variant&splice\_region\_variant 322  
stop\_gained 230  
splice\_region\_variant&synonymous\_variant 204  
disruptive\_inframe\_insertion 182  
splice\_region\_variant 178  
splice\_acceptor\_variant&intron\_variant 176  
splice\_donor\_variant&intron\_variant 169  
conservative\_inframe\_deletion 166  
conservative\_inframe\_insertion 102  
stop\_lost 26  
start\_lost 22  
non\_coding\_transcript\_variant 20  
protein\_protein\_contact 12  
splice\_acceptor\_variant&splice\_region\_variant&intron\_variant&non\_coding\_transcript\_exon\_variant 9  
stop\_retained\_variant 7  
frameshift\_variant&splice\_region\_variant 7  
splice\_acceptor\_variant&splice\_region\_variant&intron\_variant 7  
frameshift\_variant&stop\_gained 6  
gene\_fusion 5  
splice\_donor\_variant&splice\_region\_variant&intron\_variant 4  
transcript\_ablation 3  
splice\_donor\_variant&splice\_region\_variant&intron\_variant&non\_coding\_transcript\_exon\_variant 3  
frameshift\_variant&start\_lost 3  
frameshift\_variant&splice\_donor\_variant&splice\_region\_variant&intron\_variant 3  
splice\_donor\_variant&splice\_region\_variant&5\_prime\_UTR\_variant&intron\_variant 3  
bidirectional\_gene\_fusion 2  
stop\_gained&splice\_region\_variant 2  
stop\_gained&disruptive\_inframe\_insertion 2  
frameshift\_variant&splice\_acceptor\_variant&splice\_region\_variant&intron\_variant 2  
initiator\_codon\_variant 2  
splice\_acceptor\_variant&splice\_region\_variant&5\_prime\_UTR\_variant&intron\_variant 1  
splice\_acceptor\_variant&splice\_region\_variant&conservative\_inframe\_deletion&intron\_variant 1  
splice\_donor\_variant&splice\_region\_variant&disruptive\_inframe\_deletion&intron\_variant 1  
splice\_region\_variant&conservative\_inframe\_deletion 1  
start\_lost&conservative\_inframe\_deletion 1  
splice\_region\_variant&disruptive\_inframe\_deletion 1  
splice\_donor\_variant&splice\_region\_variant&conservative\_inframe\_deletion&intron\_variant 1  
splice\_acceptor\_variant&splice\_donor\_variant&intron\_variant 1  
frameshift\_variant&stop\_lost 1  
start\_lost&splice\_region\_variant 1  
stop\_lost&splice\_region\_variant 1  
initiator\_codon\_variant&splice\_region\_variant 1  
splice\_acceptor\_variant&5\_prime\_UTR\_variant&intron\_variant 1  
Name: EFFECT, dtype: int64, MODIFIER 6180234  
MODERATE 16608  
LOW 14287  
HIGH 3106  
Name: IMPACT, dtype: int64, intergenic\_region 3393167  
transcript 2548644  
gene\_variant 270396  
interaction 2028  
Name: FEATURE, dtype: int64], 'chrY.csv': [chrY 251542  
Name: CHROM, dtype: int64, intergenic\_region 172543  
intron\_variant 36066  
upstream\_gene\_variant 17703  
downstream\_gene\_variant 17621  
intragenic\_variant 4337  
non\_coding\_transcript\_exon\_variant 2628  
missense\_variant 171  
3\_prime\_UTR\_variant 167  
splice\_region\_variant&intron\_variant 90  
synonymous\_variant 79  
5\_prime\_UTR\_variant 51  
splice\_region\_variant&non\_coding\_transcript\_exon\_variant 42  
splice\_donor\_variant&intron\_variant 11  
splice\_acceptor\_variant&intron\_variant 11  
5\_prime\_UTR\_premature\_start\_codon\_gain\_variant 6  
splice\_region\_variant&synonymous\_variant 4  
stop\_gained 3  
missense\_variant&splice\_region\_variant 2  
frameshift\_variant 2  
frameshift\_variant&splice\_region\_variant 1  
splice\_region\_variant 1  
non\_coding\_transcript\_variant 1  
disruptive\_inframe\_deletion 1  
splice\_acceptor\_variant&splice\_region\_variant&intron\_variant&non\_coding\_transcript\_exon\_variant 1  
Name: EFFECT, dtype: int64, MODIFIER 251117  
LOW 222  
MODERATE 174  
HIGH 29  
Name: IMPACT, dtype: int64, intergenic\_region 171696  
transcript 74662  
gene\_variant 4337  
Name: FEATURE, dtype: int64]}

Now we pull information from the pickle file into a dataframe.

#count variants, high impact variants, and variants within coding genes  
varcounts=[]  
highcounts=[]  
geneVarCounts=[]  
for i in pickleData:  
 a=pickleData[i][0][0]  
 b=pickleData[i][2][3]  
 c=pickleData[i][3][2]  
 varcounts.append(a)  
 highcounts.append(b)  
 geneVarCounts.append(c)  
  
#calculate variants, high impact variants, and variants within coding genes for all chromosomes together   
d=sum(varcounts)  
e=sum(highcounts)  
f=sum(geneVarCounts)  
varcounts2=[d]  
highcounts2=[e]  
geneVarCounts2=[f]  
  
for j in varcounts:  
 varcounts2.append(j)  
for j in highcounts:  
 highcounts2.append(j)  
for j in geneVarCounts:  
 geneVarCounts2.append(j)

totalCountsDF=pd.read\_csv('Chromosome Counts.csv')  
#totalCountsDF=totalCountsDF.drop(['Unnamed: 2'], axis=1)  
totalCountsDF['Variant\_Count']=varcounts2  
totalCountsDF['Variant\_Percentage']=totalCountsDF['Variant\_Count']/totalCountsDF['BP\_length']\*100  
totalCountsDF['High\_Impact\_Variant\_Count']=highcounts2  
totalCountsDF['High\_Impact\_Percentage']=totalCountsDF['High\_Impact\_Variant\_Count']/totalCountsDF['BP\_length']\*100  
totalCountsDF['Variants\_In\_Genes\_Count']=geneVarCounts2  
totalCountsDF['Variants\_In\_Genes\_Percentage']=totalCountsDF['Variants\_In\_Genes\_Count']/totalCountsDF['BP\_length']\*100

totalCountsDF

Chromosome BP\_length Variant\_Count Variant\_Percentage \  
0 All 3099441038 174058239 5.615794   
1 1 249698942 14053427 5.628148   
2 2 242508799 13961577 5.757142   
3 3 198450956 11149042 5.618034   
4 4 190424264 10591957 5.562294   
5 5 181630948 10283606 5.661814   
6 6 170805979 9430386 5.521110   
7 7 159345973 9468709 5.942233   
8 8 145138636 8614681 5.935484   
9 9 138688728 7462036 5.380420   
10 10 133797422 7881973 5.890975   
11 11 135086622 8401457 6.219311   
12 12 133275309 8130250 6.100342   
13 13 114364328 5756600 5.033563   
14 14 108136338 5948454 5.500884   
15 15 102439437 5379964 5.251848   
16 16 92211104 6140757 6.659455   
17 17 83836422 5738905 6.845360   
18 18 80373285 4572260 5.688781   
19 19 58617616 5025452 8.573279   
20 20 64444167 4021075 6.239626   
21 21 46709983 2586151 5.536613   
22 22 51692466 2993743 5.791449   
23 X 156040895 6214235 3.982440   
24 Y 57264655 251542 0.439262   
  
 High\_Impact\_Variant\_Count High\_Impact\_Percentage \  
0 174398 0.005627   
1 17828 0.007140   
2 14874 0.006133   
3 8251 0.004158   
4 6823 0.003583   
5 5062 0.002787   
6 8575 0.005020   
7 8730 0.005479   
8 10152 0.006995   
9 7472 0.005388   
10 9509 0.007107   
11 9246 0.006844   
12 8983 0.006740   
13 1508 0.001319   
14 7133 0.006596   
15 4288 0.004186   
16 6782 0.007355   
17 11405 0.013604   
18 2954 0.003675   
19 10415 0.017768   
20 4145 0.006432   
21 1693 0.003624   
22 5435 0.010514   
23 3106 0.001991   
24 29 0.000051   
  
 Variants\_In\_Genes\_Count Variants\_In\_Genes\_Percentage   
0 12433715 0.401160   
1 749428 0.300133   
2 1062824 0.438262   
3 1184227 0.596735   
4 757509 0.397801   
5 783940 0.431611   
6 590047 0.345449   
7 670150 0.420563   
8 727985 0.501579   
9 430246 0.310224   
10 459432 0.343379   
11 709409 0.525151   
12 700442 0.525560   
13 227658 0.199064   
14 477927 0.441967   
15 573372 0.559718   
16 381801 0.414051   
17 418544 0.499239   
18 256067 0.318597   
19 420303 0.717025   
20 223955 0.347518   
21 189534 0.405768   
22 164182 0.317613   
23 270396 0.173285   
24 4337 0.007574

ax = totalCountsDF.plot.bar(x='Chromosome',y='Variant\_Percentage', rot=90, title="Genetic Variant Count per Chromosome as a \nPercentage of Total Base Pairs")