main_noMHC

March 17, 2022

1 Venn diagram and summary

[1]:
```
import numpy as np
import pandas as pd
from venn import venn
from matplotlib import pyplot as plt
```

1.1 Prepare data

[2]:
```
def limiting_features(set_dict, f1, f2):
    xx = len(set_dict[f1] & set_dict[f2]) / len(set_dict[f2]) * 100
    print("Comparing %s with %s: %0.2f%%" % (f1, f2, xx))
    print("Features in common: %d" % len(set_dict[f1] & set_dict[f2]))
```

1.1.1 Load PGC3 GWAS

[3]:
```
pgc3_file = '/ceph/projects/v4_phase3_paper/inputs/sz_gwas/'+\
            'pgc2_clozuk/map_phase3/_m/libd_hg38_pgc2sz_snps.tsv'
pgc3_df = pd.read_csv(pgc3_file, sep='\t', low_memory=False, index_col=0)
```

```
/home/jbenja13/.local/lib/python3.9/site-packages/numpy/lib/arraysetops.py:583:
FutureWarning: elementwise comparison failed; returning scalar instead, but in
the future will perform elementwise comparison
  mask |= (ar1 == a)
```

1.1.2 With no MHC

Genes

[4]:
```
genes = pd.read_csv('/ceph/projects/v4_phase3_paper/analysis/twas_ea/'+\
                    'gene_weights/fusion/summary_stats/_m/
 ↪fusion_associations_noMHC.txt', sep='\t')
annot = pd.read_csv('../../../differential_expression/_m/genes/
 ↪diffExpr_szVctl_full.txt', sep='\t')
genes = annot[['ensemblID']].merge(genes, left_on='ensemblID', right_on='FILE')
genes = genes[['FILE', 'ensemblID', 'ID', 'HSQ', 'BEST.GWAS.ID', 'EQTL.ID',
               'TWAS.Z', 'TWAS.P', 'FDR', 'Bonferroni']]
genes['Type'] = 'Gene'
genes.rename(columns={'FILE': 'Feature'}, inplace=True)
genes.sort_values('TWAS.P').head(2)
```

```
[4]:
              Feature        ensemblID     ID       HSQ       BEST.GWAS.ID  \
4665  ENSG00000163938  ENSG00000163938   GNL3  0.410009  chr3:52781889:T:C
3609  ENSG00000166159  ENSG00000166159  LRTM2  0.239365  chr12:2221292:C:T

                EQTL.ID    TWAS.Z        TWAS.P           FDR    Bonferroni  \
4665  chr3:52588070:G:A  9.415273  4.718536e-21  2.348887e-17  2.348887e-17
3609  chr12:2224318:C:T -9.064394  1.252984e-19  3.118677e-16  6.237353e-16

      Type
4665  Gene
3609  Gene
```

Transcripts

```
[5]: trans = pd.read_csv('/ceph/projects/v4_phase3_paper/analysis/twas_ea/'+\
                         'transcript_weights/fusion/summary_stats/_m/
      ↪fusion_associations_noMHC.txt', sep='\t')
     annot = pd.read_csv('../../../differential_expression/_m/transcripts/
      ↪diffExpr_szVctl_full.txt', sep='\t')
     annot['ensemblID'] = annot.gene_id.str.replace('\\..*', '', regex=True)
     annot['FILE'] = annot.transcript_id.str.replace('\\..*', '', regex=True)
     trans = annot[['ensemblID', 'FILE']].merge(trans, on='FILE')
     trans = trans[['FILE', 'ensemblID', 'ID', 'HSQ', 'BEST.GWAS.ID', 'EQTL.ID',
                    'TWAS.Z', 'TWAS.P', 'FDR', 'Bonferroni']]
     trans['Type'] = 'Transcript'
     trans.rename(columns={'FILE': 'Feature'}, inplace=True)
     trans.sort_values('TWAS.P').head(2)
```

```
[5]:
              Feature        ensemblID       ID       HSQ  \
4596  ENST00000394799  ENSG00000163938     GNL3  0.122586
1396  ENST00000315580  ENSG00000182196  ARL6IP4  0.482633

             BEST.GWAS.ID              EQTL.ID    TWAS.Z        TWAS.P  \
4596    chr3:52781889:T:C    chr3:52799789:C:A  8.983223  2.629480e-19
1396  chr12:123148383:G:A  chr12:122973072:C:T -8.699604  3.330436e-18

               FDR    Bonferroni        Type
4596  2.283966e-15  2.283966e-15  Transcript
1396  1.446408e-14  2.892817e-14  Transcript
```

Exons

```
[6]: exons = pd.read_csv('/ceph/projects/v4_phase3_paper/analysis/twas_ea/'+\
                         'exon_weights/fusion/summary_stats/_m/
      ↪fusion_associations_noMHC.txt', sep='\t')
     annot = pd.read_csv('../../../differential_expression/_m/exons/
      ↪diffExpr_szVctl_full.txt', sep='\t', index_col=0)
```

```
exons = annot[['ensemblID']].merge(exons, left_index=True, right_on='FILE')
exons = exons[['FILE', 'ensemblID', 'ID', 'HSQ', 'BEST.GWAS.ID', 'EQTL.ID',
               'TWAS.Z', 'TWAS.P', 'FDR', 'Bonferroni']]
exons['Type'] = 'Exon'
exons.rename(columns={'FILE': 'Feature'}, inplace=True)
exons.sort_values('TWAS.P').head(2)
```

```
[6]:       Feature        ensemblID    ID       HSQ       BEST.GWAS.ID  \
     32333  e228054  ENSG00000163938  GNL3  0.080457  chr3:52781889:T:C
     32332  e228120  ENSG00000163938  GNL3  0.078287  chr3:52781889:T:C

                      EQTL.ID    TWAS.Z        TWAS.P           FDR    Bonferroni  \
     32333  chr3:52588070:G:A  9.615423  6.882578e-22  1.111886e-17  2.607946e-17
     32332  chr3:52588070:G:A  9.597698  8.174979e-22  1.111886e-17  3.097663e-17

            Type
     32333  Exon
     32332  Exon
```

1.1.3 Junctions

```
[7]: annot = pd.read_csv('jxn_annotation.tsv', sep='\t', index_col=1)
     annot["gene_id"] = annot.index
     juncs = pd.read_csv('/ceph/projects/v4_phase3_paper/analysis/twas_ea/'+\
                         'junction_weights/fusion/summary_stats/_m/
      ↪fusion_associations_noMHC.txt', sep='\t')
     juncs = pd.merge(annot, juncs, left_on='JxnID', right_on='FILE')
     juncs = juncs[['gene_id', 'ensemblID', 'ID', 'HSQ', 'BEST.GWAS.ID', 'EQTL.ID',
                    'TWAS.Z', 'TWAS.P', 'FDR', 'Bonferroni']]
     juncs['Type'] = 'Junction'
     juncs.rename(columns={'Symbol': 'ID', 'gene_id': 'Feature'}, inplace=True)
     juncs.sort_values('TWAS.P').head(2)
```

```
[7]:                        Feature        ensemblID    ID       HSQ  \
     8293  chr3:52690705-52690944(+)  ENSG00000163938  GNL3  0.062694
     8295  chr3:52693808-52694036(+)  ENSG00000163938  GNL3  0.114983

                BEST.GWAS.ID            EQTL.ID    TWAS.Z        TWAS.P  \
     8293  chr3:52781889:T:C  chr3:52507237:A:G  9.632857  5.809151e-22
     8295  chr3:52781889:T:C  chr3:52594040:C:A  9.401695  5.369125e-21

                    FDR    Bonferroni      Type
     8293  7.466501e-18  7.466501e-18  Junction
     8295  3.450468e-17  6.900936e-17  Junction
```

1.2 Heritable features

1.2.1 Feature summary

[8]:
```
gg = len(set(genes['Feature']))
tt = len(set(trans['Feature']))
ee = len(set(exons['Feature']))
jj = len(set(juncs['Feature']))

print("===Unique Features===\nGene:\t\t%d\nTranscript:\t%d\nExon:
 ↪\t\t%d\nJunction:\t%d\n" % (gg, tt, ee, jj))

gg = len(set(genes['ensemblID']))
tt = len(set(trans['ensemblID']))
ee = len(set(exons['ensemblID']))
jj = len(set(juncs['ensemblID']))

print("===Unique Ensembl Gene===\nGene:\t\t%d\nTranscript:\t%d\nExon:
 ↪\t\t%d\nJunction:\t%d\n" % (gg, tt, ee, jj))

gg = len(set(genes['ID']))
tt = len(set(trans['ID']))
ee = len(set(exons['ID']))
jj = len(set(juncs['ID']))

print("===Unique Gene Name===\nGene:\t\t%d\nTranscript:\t%d\nExon:
 ↪\t\t%d\nJunction:\t%d\n" % (gg, tt, ee, jj))
```

```
===Unique Features===
Gene:           4978
Transcript:     8686
Exon:           37892
Junction:       12853

===Unique Ensembl Gene===
Gene:           4978
Transcript:     5857
Exon:           7587
Junction:       4282

===Unique Gene Name===
Gene:           4978
Transcript:     5854
Exon:           8660
Junction:       4876
```

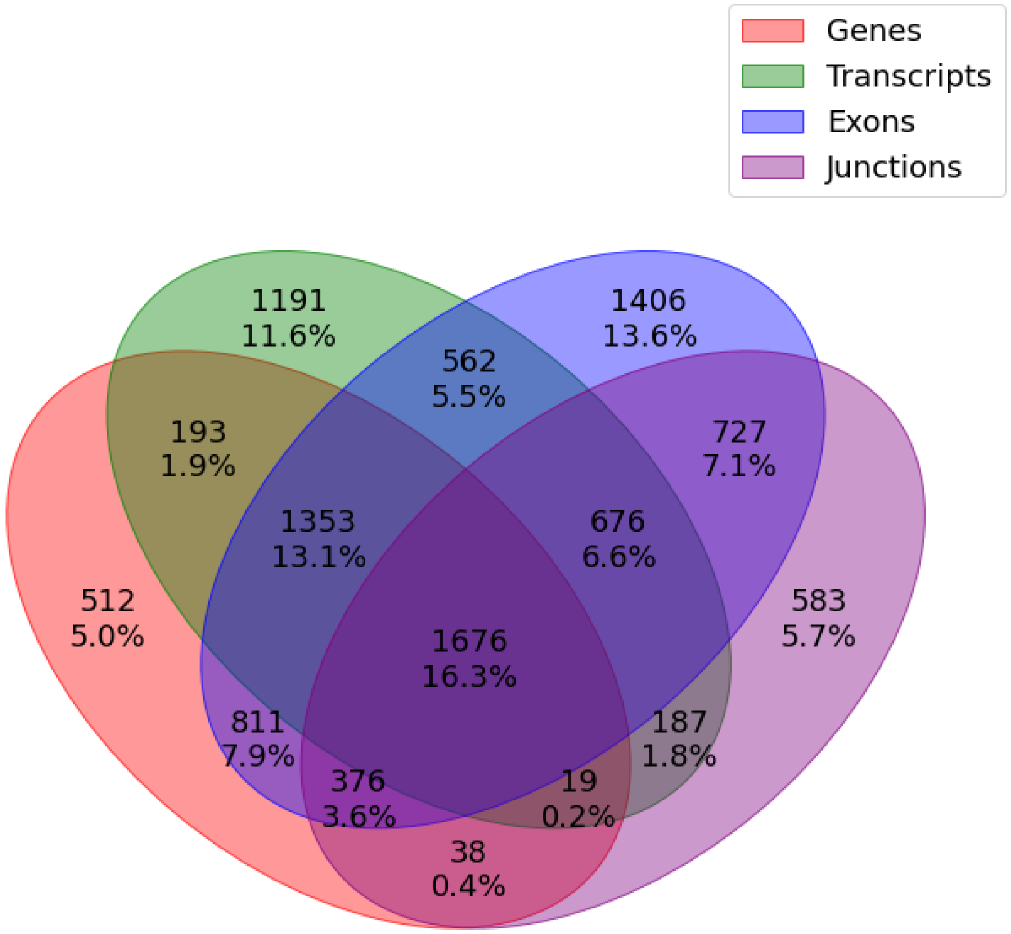
1.2.2 Plot venn

[9]:

```
features = {
    'Genes': set(genes['ensemblID']),
    'Transcripts': set(trans['ensemblID']),
    'Exons': set(exons['ensemblID']),
    'Junctions': set(juncs['ensemblID']),
}
```

[10]:

```
venn(features, fmt="{size}\n{percentage:0.1f}%", fontsize=18, legend_loc="best",
     figsize=(12, 12), cmap=['red', 'green', 'blue', 'purple'])
plt.savefig('heritable_allFeatures_venn_diagram_percentage.png')
plt.savefig('heritable_allFeatures_venn_diagram_percentage.pdf')
plt.savefig('heritable_allFeatures_venn_diagram_percentage.svg')
plt.show()
```



[11]:

```
limiting_features(features, 'Genes', 'Transcripts')
limiting_features(features, 'Genes', 'Junctions')
limiting_features(features, 'Exons', 'Genes')
```

```
Comparing Genes with Transcripts: 55.34%
Features in common: 3241
Comparing Genes with Junctions: 49.25%
Features in common: 2109
Comparing Exons with Genes: 84.69%
Features in common: 4216
```

[12]:

```
limiting_features(features, 'Transcripts', 'Junctions')
limiting_features(features, 'Exons', 'Transcripts')
limiting_features(features, 'Exons', 'Junctions')
```

```
Comparing Transcripts with Junctions: 59.74%
Features in common: 2558
Comparing Exons with Transcripts: 72.85%
Features in common: 4267
Comparing Exons with Junctions: 80.69%
Features in common: 3455
```

[13]:

```
len(features['Genes'] & features['Transcripts'] & features['Exons'] &␣
 ↪features['Junctions'])
```

[13]:

```
1676
```

[14]:

```
len(features['Genes'] | features['Transcripts'] | features['Exons'] |␣
 ↪features['Junctions'])
```

[14]:

```
10310
```

1.2.3 SNPs not in significant PGC2+CLOZUK GWAS

[15]:

```
new_genes = pd.merge(genes, pgc3_df, left_on='BEST.GWAS.ID',␣
 ↪right_on='our_snp_id', suffixes=['_TWAS', '_PGC2'])
new_trans = pd.merge(trans, pgc3_df, left_on='BEST.GWAS.ID',␣
 ↪right_on='our_snp_id', suffixes=['_TWAS', '_PGC2'])
new_exons = pd.merge(exons, pgc3_df, left_on='BEST.GWAS.ID',␣
 ↪right_on='our_snp_id', suffixes=['_TWAS', '_PGC2'])
new_juncs = pd.merge(juncs, pgc3_df, left_on='BEST.GWAS.ID',␣
 ↪right_on='our_snp_id', suffixes=['_TWAS', '_PGC2'])

new_genes = new_genes[(new_genes['P'] > 5e-8)].copy()
new_trans = new_trans[(new_trans['P'] > 5e-8)].copy()
new_exons = new_exons[(new_exons['P'] > 5e-8)].copy()
```

```
new_juncs = new_juncs[(new_juncs['P'] > 5e-8)].copy()
```

[16]:

```
gg = len(set(new_genes['BEST.GWAS.ID']))
tt = len(set(new_trans['BEST.GWAS.ID']))
ee = len(set(new_exons['BEST.GWAS.ID']))
jj = len(set(new_juncs['BEST.GWAS.ID']))

print("===Unique novel SNPs===\nGene:\t\t%d\nTranscript:\t%d\nExon:
 ↪\t\t%d\nJunction:\t%d\n" % (gg, tt, ee, jj))
```

```
===Unique novel SNPs===
Gene:           2115
Transcript:     2341
Exon:           2997
Junction:       2323
```

[17]:

```
len(set(new_genes['BEST.GWAS.ID']) | set(new_trans['BEST.GWAS.ID']) |
    set(new_exons['BEST.GWAS.ID']) | set(new_juncs['BEST.GWAS.ID']))
```

[17]: 3617

1.3 TWAS P-value $<$ 0.05

1.3.1 Feature summary

[18]:

```
gg = len(set(genes[(genes['TWAS.P'] <= 0.05)].loc[:, 'Feature']))
tt = len(set(trans[(trans['TWAS.P'] <= 0.05)].loc[:, 'Feature']))
ee = len(set(exons[(exons['TWAS.P'] <= 0.05)].loc[:, 'Feature']))
jj = len(set(juncs[(juncs['TWAS.P'] <= 0.05)].loc[:, 'Feature']))

print("===Unique Features===\nGene:\t\t%d\nTranscript:\t%d\nExon:
 ↪\t\t%d\nJunction:\t%d\n" % (gg, tt, ee, jj))

gg = len(set(genes[(genes['TWAS.P'] <= 0.05)].loc[:, 'ensemblID']))
tt = len(set(trans[(trans['TWAS.P'] <= 0.05)].loc[:, 'ensemblID']))
ee = len(set(exons[(exons['TWAS.P'] <= 0.05)].loc[:, 'ensemblID']))
jj = len(set(juncs[(juncs['TWAS.P'] <= 0.05)].loc[:, 'ensemblID']))

print("===Unique Ensembl Gene===\nGene:\t\t%d\nTranscript:\t%d\nExon:
 ↪\t\t%d\nJunction:\t%d\n" % (gg, tt, ee, jj))

gg = len(set(genes[(genes['TWAS.P'] <= 0.05)].loc[:, 'ID']))
tt = len(set(trans[(trans['TWAS.P'] <= 0.05)].loc[:, 'ID']))
ee = len(set(exons[(exons['TWAS.P'] <= 0.05)].loc[:, 'ID']))
jj = len(set(juncs[(juncs['TWAS.P'] <= 0.05)].loc[:, 'ID']))
```

```
print("===Unique Gene Names===\nGene:\t\t%d\nTranscript:\t%d\nExon:
 ↪\t\t%d\nJunction:\t%d\n" % (gg, tt, ee, jj))
```

```
===Unique Features===
Gene:           1109
Transcript:     2024
Exon:           9100
Junction:       3027

===Unique Ensembl Gene===
Gene:           1109
Transcript:     1576
Exon:           2584
Junction:       1403

===Unique Gene Names===
Gene:           1109
Transcript:     1575
Exon:           2753
Junction:       1492
```

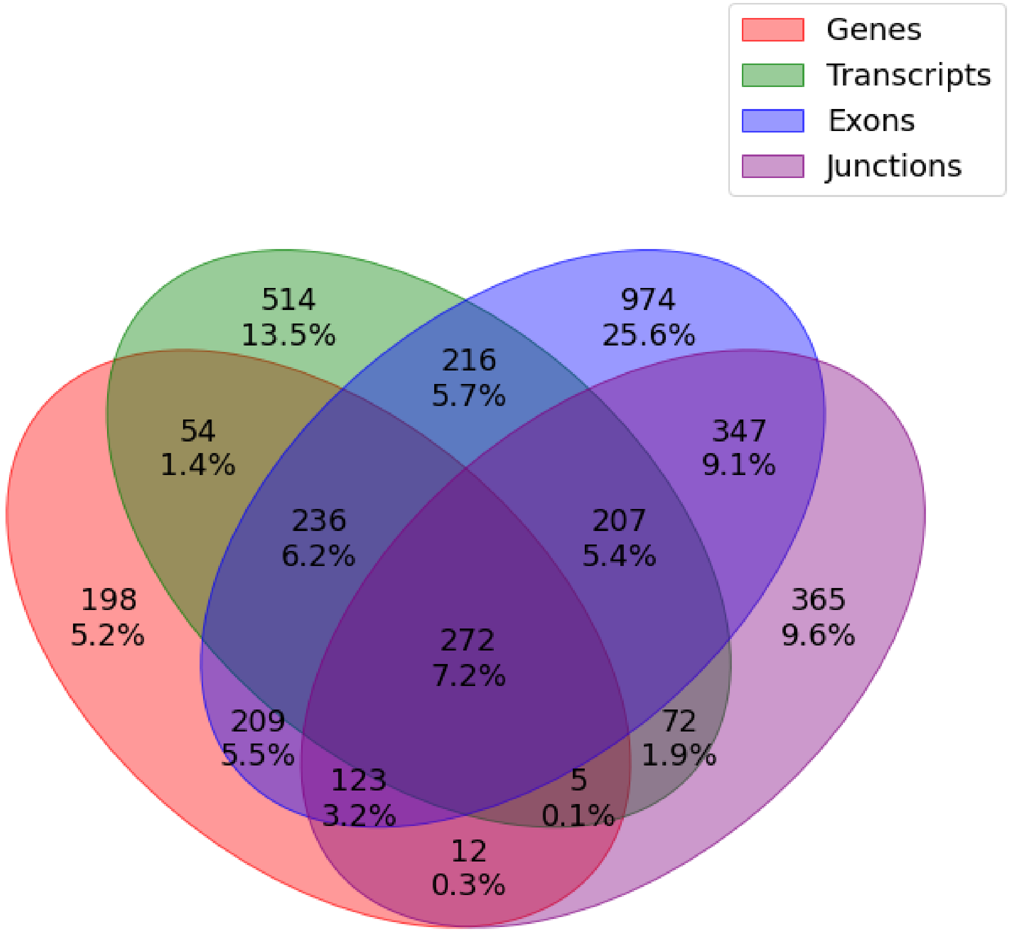
1.3.2 Plot venn

[19]:

```
features = {
    'Genes': set(genes[(genes['TWAS.P'] <= 0.05)].loc[:, 'ensemblID']),
    'Transcripts': set(trans[(trans['TWAS.P'] <= 0.05)].loc[:, 'ensemblID']),
    'Exons': set(exons[(exons['TWAS.P'] <= 0.05)].loc[:, 'ensemblID']),
    'Junctions': set(juncs[(juncs['TWAS.P'] <= 0.05)].loc[:, 'ensemblID']),
}
```

[20]:

```
venn(features, fmt="{size}\n{percentage:0.1f}%", fontsize=18, legend_loc="best",
     figsize=(12, 12), cmap=['red', 'green', 'blue', 'purple'])
plt.savefig('sigPval_allFeatures_venn_diagram_percentage.png')
plt.savefig('sigPval_allFeatures_venn_diagram_percentage.pdf')
plt.savefig('sigPval_allFeatures_venn_diagram_percentage.svg')
plt.show()
```



[21]:

```
limiting_features(features, 'Genes', 'Transcripts')
limiting_features(features, 'Genes', 'Junctions')
limiting_features(features, 'Exons', 'Genes')
```

```
Comparing Genes with Transcripts: 35.98%
Features in common: 567
Comparing Genes with Junctions: 29.37%
Features in common: 412
Comparing Exons with Genes: 75.74%
Features in common: 840
```

[22]:

```
limiting_features(features, 'Transcripts', 'Junctions')
limiting_features(features, 'Exons', 'Transcripts')
limiting_features(features, 'Exons', 'Junctions')
```

```
Comparing Transcripts with Junctions: 39.63%
Features in common: 556
Comparing Exons with Transcripts: 59.07%
Features in common: 931
Comparing Exons with Junctions: 67.64%
Features in common: 949
```

[23]:
```
len(features['Genes'] & features['Transcripts'] & features['Exons'] &␣
 ↪features['Junctions'])
```

[23]:
```
272
```

[24]:
```
len(features['Genes'] | features['Transcripts'] | features['Exons'] |␣
 ↪features['Junctions'])
```

[24]:
```
3804
```

1.3.3 SNPs not in significant PGC2+CLOZUK GWAS

[25]:
```
new_genes = pd.merge(genes[(genes['TWAS.P'] <= 0.05)], pgc3_df, left_on='BEST.
 ↪GWAS.ID',
                     right_on='our_snp_id', suffixes=['_TWAS', '_PGC2'])
new_trans = pd.merge(trans[(trans['TWAS.P'] <= 0.05)], pgc3_df, left_on='BEST.
 ↪GWAS.ID',
                     right_on='our_snp_id', suffixes=['_TWAS', '_PGC2'])
new_exons = pd.merge(exons[(exons['TWAS.P'] <= 0.05)], pgc3_df, left_on='BEST.
 ↪GWAS.ID',
                     right_on='our_snp_id', suffixes=['_TWAS', '_PGC2'])
new_juncs = pd.merge(juncs[(juncs['TWAS.P'] <= 0.05)], pgc3_df, left_on='BEST.
 ↪GWAS.ID',
                     right_on='our_snp_id', suffixes=['_TWAS', '_PGC2'])

new_genes = new_genes[(new_genes['P'] > 5e-8)].copy()
new_trans = new_trans[(new_trans['P'] > 5e-8)].copy()
new_exons = new_exons[(new_exons['P'] > 5e-8)].copy()
new_juncs = new_juncs[(new_juncs['P'] > 5e-8)].copy()
```

[26]:
```
gg = len(set(new_genes['BEST.GWAS.ID']))
tt = len(set(new_trans['BEST.GWAS.ID']))
ee = len(set(new_exons['BEST.GWAS.ID']))
jj = len(set(new_juncs['BEST.GWAS.ID']))

print("===Unique novel SNPs===\nGene:\t\t%d\nTranscript:\t%d\nExon:
 ↪\t\t%d\nJunction:\t%d\n" % (gg, tt, ee, jj))
```

```
===Unique novel SNPs===
Gene:           622
Transcript:     822
Exon:           1271
Junction:       861
```

[27]:
```
len(set(new_genes['BEST.GWAS.ID']) | set(new_trans['BEST.GWAS.ID']) |
    set(new_exons['BEST.GWAS.ID']) | set(new_juncs['BEST.GWAS.ID']))
```

[27]:
```
1658
```

1.4 TWAS FDR < 0.05

1.4.1 Feature summary

[28]:
```
gg = len(set(genes[(genes['FDR'] <= 0.05)].loc[:, 'Feature']))
tt = len(set(trans[(trans['FDR'] <= 0.05)].loc[:, 'Feature']))
ee = len(set(exons[(exons['FDR'] <= 0.05)].loc[:, 'Feature']))
jj = len(set(juncs[(juncs['FDR'] <= 0.05)].loc[:, 'Feature']))

print("===Unique Features===\nGene:\t\t%d\nTranscript:\t%d\nExon:
 ↪\t\t%d\nJunction:\t%d\n" % (gg, tt, ee, jj))

gg = len(set(genes[(genes['FDR'] <= 0.05)].loc[:, 'ensemblID']))
tt = len(set(trans[(trans['FDR'] <= 0.05)].loc[:, 'ensemblID']))
ee = len(set(exons[(exons['FDR'] <= 0.05)].loc[:, 'ensemblID']))
jj = len(set(juncs[(juncs['FDR'] <= 0.05)].loc[:, 'ensemblID']))

print("===Unique Ensembl Gene===\nGene:\t\t%d\nTranscript:\t%d\nExon:
 ↪\t\t%d\nJunction:\t%d\n" % (gg, tt, ee, jj))

gg = len(set(genes[(genes['FDR'] <= 0.05)].loc[:, 'ID']))
tt = len(set(trans[(trans['FDR'] <= 0.05)].loc[:, 'ID']))
ee = len(set(exons[(exons['FDR'] <= 0.05)].loc[:, 'ID']))
jj = len(set(juncs[(juncs['FDR'] <= 0.05)].loc[:, 'ID']))

print("===Unique Gene Name===\nGene:\t\t%d\nTranscript:\t%d\nExon:
 ↪\t\t%d\nJunction:\t%d\n" % (gg, tt, ee, jj))
```

```
===Unique Features===
Gene:           466
Transcript:     972
Exon:           4240
Junction:       1295

===Unique Ensembl Gene===
Gene:           466
Transcript:     747
```

```
Exon:           1282
Junction:       644

===Unique Gene Name===
Gene:           466
Transcript:     747
Exon:           1344
Junction:       688
```

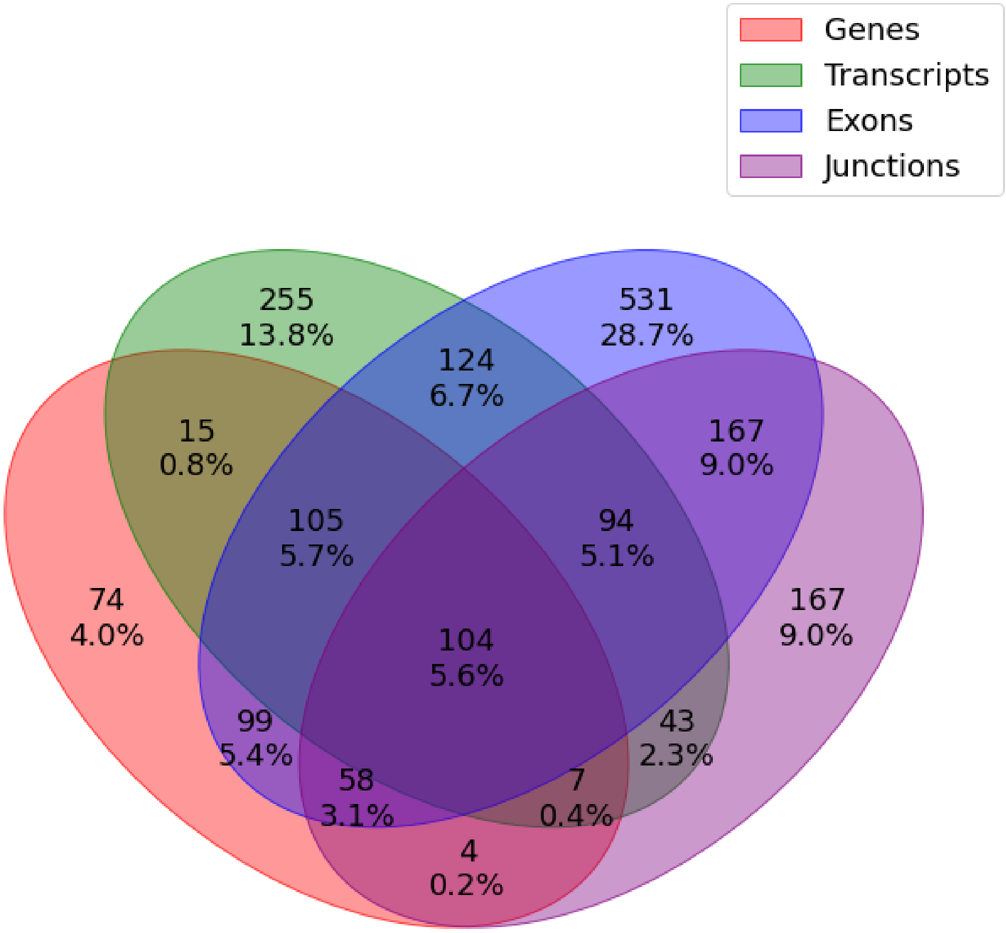
1.4.2 Plot venn

[29]:

```
features = {
    'Genes': set(genes[(genes['FDR'] <= 0.05)].loc[:, 'ensemblID']),
    'Transcripts': set(trans[(trans['FDR'] <= 0.05)].loc[:, 'ensemblID']),
    'Exons': set(exons[(exons['FDR'] <= 0.05)].loc[:, 'ensemblID']),
    'Junctions': set(juncs[(juncs['FDR'] <= 0.05)].loc[:, 'ensemblID']),
}
```

[30]:

```
venn(features, fmt="{size}\n{percentage:0.1f}%", fontsize=18, legend_loc="best",
     figsize=(12, 12), cmap=['red', 'green', 'blue', 'purple'])
plt.savefig('fdr_allFeatures_venn_diagram_percentage.png')
plt.savefig('fdr_allFeatures_venn_diagram_percentage.pdf')
plt.savefig('fdr_allFeatures_venn_diagram_percentage.svg')
plt.show()
```



```
[31]: limiting_features(features, 'Genes', 'Transcripts')
      limiting_features(features, 'Genes', 'Junctions')
      limiting_features(features, 'Exons', 'Genes')
```

```
Comparing Genes with Transcripts: 30.92%
Features in common: 231
Comparing Genes with Junctions: 26.86%
Features in common: 173
Comparing Exons with Genes: 78.54%
Features in common: 366
```

```
[32]: limiting_features(features, 'Transcripts', 'Junctions')
      limiting_features(features, 'Exons', 'Transcripts')
limiting_features(features, 'Exons', 'Junctions')
```

```
Comparing Transcripts with Junctions: 38.51%
Features in common: 248
Comparing Exons with Transcripts: 57.16%
Features in common: 427
Comparing Exons with Junctions: 65.68%
Features in common: 423
```

```
[33]: len(features['Genes'] & features['Transcripts'] & features['Exons'] &␣
       ↪features['Junctions'])
```

```
[33]: 104
```

```
[34]: len(features['Genes'] | features['Transcripts'] | features['Exons'] |␣
       ↪features['Junctions'])
```

```
[34]: 1847
```

1.4.3 SNPs not in significant PGC2+CLOZUK GWAS

```
[35]: new_genes = pd.merge(genes[(genes['FDR'] <= 0.05)], pgc3_df, left_on='BEST.GWAS.
       ↪ID',
                             right_on='our_snp_id', suffixes=['_TWAS', '_PGC2'])
      new_trans = pd.merge(trans[(trans['FDR'] <= 0.05)], pgc3_df, left_on='BEST.GWAS.
       ↪ID',
                             right_on='our_snp_id', suffixes=['_TWAS', '_PGC2'])
      new_exons = pd.merge(exons[(exons['FDR'] <= 0.05)], pgc3_df, left_on='BEST.GWAS.
       ↪ID',
                             right_on='our_snp_id', suffixes=['_TWAS', '_PGC2'])
      new_juncs = pd.merge(juncs[(juncs['FDR'] <= 0.05)], pgc3_df, left_on='BEST.GWAS.
       ↪ID',
                             right_on='our_snp_id', suffixes=['_TWAS', '_PGC2'])

      new_genes = new_genes[(new_genes['P'] > 5e-8)].copy()
      new_trans = new_trans[(new_trans['P'] > 5e-8)].copy()
      new_exons = new_exons[(new_exons['P'] > 5e-8)].copy()
      new_juncs = new_juncs[(new_juncs['P'] > 5e-8)].copy()
```

```
[36]: gg = len(set(new_genes['BEST.GWAS.ID']))
      tt = len(set(new_trans['BEST.GWAS.ID']))
      ee = len(set(new_exons['BEST.GWAS.ID']))
      jj = len(set(new_juncs['BEST.GWAS.ID']))

      print("===Unique novel SNPs===\nGene:\t\t%d\nTranscript:\t%d\nExon:
       ↪\t\t%d\nJunction:\t%d\n" % (gg, tt, ee, jj))
```

```
===Unique novel SNPs===
Gene:           234
Transcript:     388
Exon:           620
Junction:       398
```

[37]:

```
len(set(new_genes['BEST.GWAS.ID']) | set(new_trans['BEST.GWAS.ID']) |
    set(new_exons['BEST.GWAS.ID']) | set(new_juncs['BEST.GWAS.ID']))
```

[37]:

```
830
```

1.5 TWAS Bonferroni < 0.05

1.5.1 Feature summary

[38]:

```
gg = len(set(genes[(genes['Bonferroni'] <= 0.05)].loc[:, 'Feature']))
tt = len(set(trans[(trans['Bonferroni'] <= 0.05)].loc[:, 'Feature']))
ee = len(set(exons[(exons['Bonferroni'] <= 0.05)].loc[:, 'Feature']))
jj = len(set(juncs[(juncs['Bonferroni'] <= 0.05)].loc[:, 'Feature']))

print("===Unique Features===\nGene:\t\t%d\nTranscript:\t%d\nExon:
 ↪\t\t%d\nJunction:\t%d\n" % (gg, tt, ee, jj))

gg = len(set(genes[(genes['Bonferroni'] <= 0.05)].loc[:, 'ensemblID']))
tt = len(set(trans[(trans['Bonferroni'] <= 0.05)].loc[:, 'ensemblID']))
ee = len(set(exons[(exons['Bonferroni'] <= 0.05)].loc[:, 'ensemblID']))
jj = len(set(juncs[(juncs['Bonferroni'] <= 0.05)].loc[:, 'ensemblID']))

print("===Unique Ensembl Gene===\nGene:\t\t%d\nTranscript:\t%d\nExon:
 ↪\t\t%d\nJunction:\t%d\n" % (gg, tt, ee, jj))

gg = len(set(genes[(genes['Bonferroni'] <= 0.05)].loc[:, 'ID']))
tt = len(set(trans[(trans['Bonferroni'] <= 0.05)].loc[:, 'ID']))
ee = len(set(exons[(exons['Bonferroni'] <= 0.05)].loc[:, 'ID']))
jj = len(set(juncs[(juncs['Bonferroni'] <= 0.05)].loc[:, 'ID']))

print("===Unique Gene Name===\nGene:\t\t%d\nTranscript:\t%d\nExon:
 ↪\t\t%d\nJunction:\t%d\n" % (gg, tt, ee, jj))
```

```
===Unique Features===
Gene:           120
Transcript:     209
Exon:           589
Junction:       264

===Unique Ensembl Gene===
Gene:           120
Transcript:     164
```

```
Exon:           213
Junction:       129

===Unique Gene Name===
Gene:           120
Transcript:     164
Exon:           219
Junction:       142
```

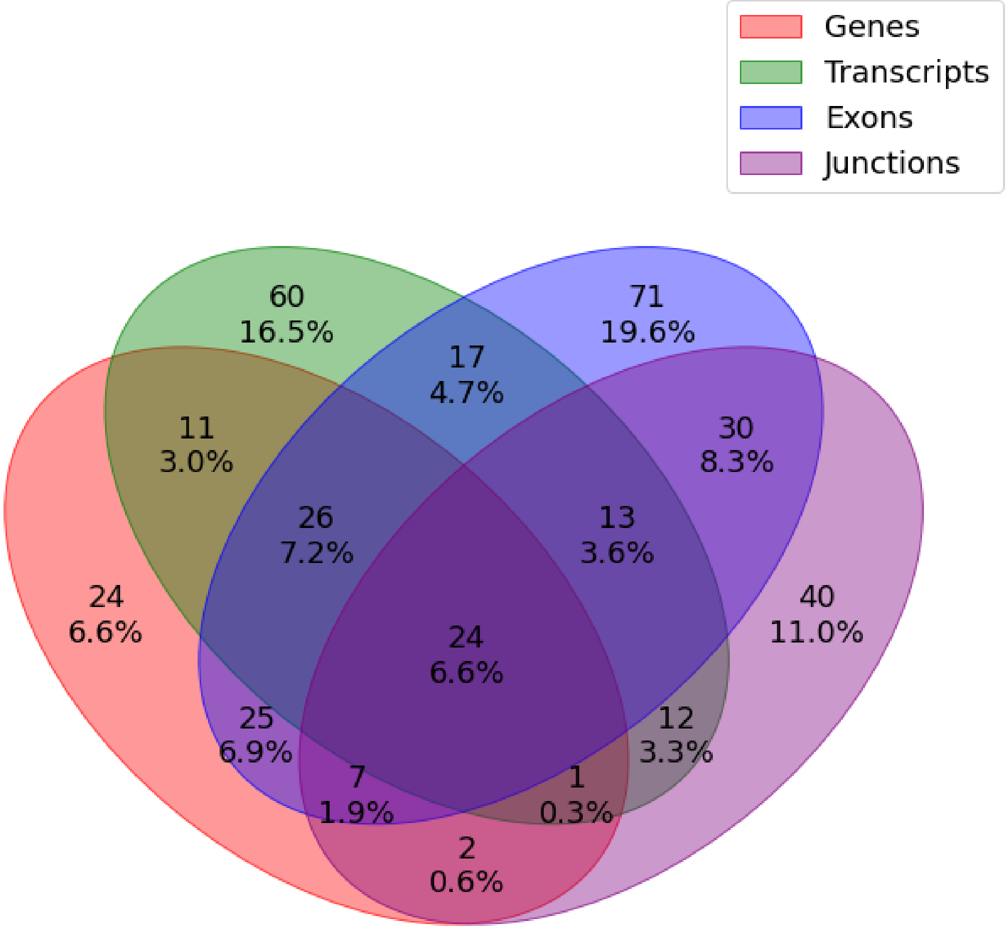
1.5.2 Plot venn

[39]:

```
features = {
    'Genes': set(genes[(genes['Bonferroni'] <= 0.05)].loc[:, 'ensemblID']),
    'Transcripts': set(trans[(trans['Bonferroni'] <= 0.05)].loc[:,␣
 ↪'ensemblID']),
    'Exons': set(exons[(exons['Bonferroni'] <= 0.05)].loc[:, 'ensemblID']),
    'Junctions': set(juncs[(juncs['Bonferroni'] <= 0.05)].loc[:, 'ensemblID']),
}
```

[40]:

```
venn(features, fmt="{size}\n{percentage:0.1f}%", fontsize=18, legend_loc="best",
     figsize=(12, 12), cmap=['red', 'green', 'blue', 'purple'])
plt.savefig('bonferroni_allFeatures_venn_diagram_percentage.png')
plt.savefig('bonferroni_allFeatures_venn_diagram_percentage.pdf')
plt.savefig('bonferroni_allFeatures_venn_diagram_percentage.svg')
plt.show()
```



[41]:
```
limiting_features(features, 'Genes', 'Transcripts')
limiting_features(features, 'Genes', 'Junctions')
limiting_features(features, 'Exons', 'Genes')
```

```
Comparing Genes with Transcripts: 37.80%
Features in common: 62
Comparing Genes with Junctions: 26.36%
Features in common: 34
Comparing Exons with Genes: 68.33%
Features in common: 82
```

[42]:
```
limiting_features(features, 'Transcripts', 'Junctions')
limiting_features(features, 'Exons', 'Transcripts')
limiting_features(features, 'Exons', 'Junctions')
```

```
Comparing Transcripts with Junctions: 38.76%
Features in common: 50
Comparing Exons with Transcripts: 48.78%
Features in common: 80
Comparing Exons with Junctions: 57.36%
Features in common: 74
```

[43]:
```
len(features['Genes'] & features['Transcripts'] & features['Exons'] &␣
 ↪features['Junctions'])
```

[43]:
```
24
```

[44]:
```
len(features['Genes'] | features['Transcripts'] | features['Exons'] |␣
 ↪features['Junctions'])
```

[44]:
```
363
```

1.5.3 SNPs not in significant PGC2+CLOZUK GWAS

[45]:
```
new_genes = pd.merge(genes[(genes['Bonferroni'] <= 0.05)], pgc3_df,␣
 ↪left_on='BEST.GWAS.ID',
                     right_on='our_snp_id', suffixes=['_TWAS', '_PGC2'])
new_trans = pd.merge(trans[(trans['Bonferroni'] <= 0.05)], pgc3_df,␣
 ↪left_on='BEST.GWAS.ID',
                     right_on='our_snp_id', suffixes=['_TWAS', '_PGC2'])
new_exons = pd.merge(exons[(exons['Bonferroni'] <= 0.05)], pgc3_df,␣
 ↪left_on='BEST.GWAS.ID',
                     right_on='our_snp_id', suffixes=['_TWAS', '_PGC2'])
new_juncs = pd.merge(juncs[(juncs['Bonferroni'] <= 0.05)], pgc3_df,␣
 ↪left_on='BEST.GWAS.ID',
                     right_on='our_snp_id', suffixes=['_TWAS', '_PGC2'])

new_genes = new_genes[(new_genes['P'] > 5e-8)].copy()
new_trans = new_trans[(new_trans['P'] > 5e-8)].copy()
new_exons = new_exons[(new_exons['P'] > 5e-8)].copy()
new_juncs = new_juncs[(new_juncs['P'] > 5e-8)].copy()
```

[46]:
```
gg = len(set(new_genes['BEST.GWAS.ID']))
tt = len(set(new_trans['BEST.GWAS.ID']))
ee = len(set(new_exons['BEST.GWAS.ID']))
jj = len(set(new_juncs['BEST.GWAS.ID']))

print("===Unique novel SNPs===\nGene:\t\t%d\nTranscript:\t%d\nExon:
 ↪\t\t%d\nJunction:\t%d\n" % (gg, tt, ee, jj))
```

```
===Unique novel SNPs===
Gene:           37
Transcript:     50
Exon:           57
Junction:       48
```

[47]:

```
len(set(new_genes['BEST.GWAS.ID']) | set(new_trans['BEST.GWAS.ID']) |
    set(new_exons['BEST.GWAS.ID']) | set(new_juncs['BEST.GWAS.ID']))
```

[47]:

```
97
```

1.6 Session Information

[48]:

```
import types
from IPython import sys_info

def imports():
    for name, val in globals().items():
        if isinstance(val, types.ModuleType):
            yield val.__name__

#exclude all modules not listed by `!pip freeze`
excludes = ['__builtin__', 'types', 'IPython.core.shadowns', 'sys', 'os']
function_modules = []
imported_modules = [module for module in imports() if module not in excludes] +␣
 ↪function_modules
pip_modules = !pip freeze #you could also use `!conda list` with anaconda
```

[49]:

```
print(sys_info())
```

```
{'commit_hash': '3813660de',
 'commit_source': 'installation',
 'default_encoding': 'utf-8',
 'ipython_path': '/usr/lib/python3.9/site-packages/IPython',
 'ipython_version': '7.29.0',
 'os_name': 'posix',
 'platform': 'Linux-5.15.5-arch1-1-x86_64-with-glibc2.33',
 'sys_executable': '/usr/bin/python3',
 'sys_platform': 'linux',
 'sys_version': '3.9.7 (default, Oct 10 2021, 15:13:22) \n[GCC 11.1.0]'}
```