main

August 11, 2021

1 TWAS feature summary

[1]:
```
import pandas as pd
```

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 PGC2+COLUZK GWAS

[3]:
```
pgc2_file = '/ceph/projects/v4_phase3_paper/inputs/sz_gwas/'+\
            'pgc2_clozuk/map_phase3/_m/libd_hg38_pgc2sz_snps.tsv'
pgc2_df = pd.read_csv(pgc2_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 MHC

Genes

[4]:
```
genes = pd.read_csv('/ceph/projects/v4_phase3_paper/analysis/twas/'+\
                    'gene_weights/fusion_pgc2/summary_stats/_m/
 ↪fusion_associations.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  \
6805  ENSG00000158691  ENSG00000158691    ZSCAN12  0.070262
7214  ENSG00000219891  ENSG00000219891  ZSCAN12P1  0.266109

           BEST.GWAS.ID            EQTL.ID     TWAS.Z        TWAS.P  \
6805  chr6:28744470:A:G  chr6:28744886:A:G -12.627320  1.492752e-36
7214  chr6:28426903:C:T  chr6:27883095:G:A  12.353178  4.682431e-35

                FDR    Bonferroni  Type
6805  1.225699e-32  1.225699e-32  Gene
7214  1.922372e-31  3.844744e-31  Gene
```

Transcripts

[5]:
```
trans = pd.read_csv('/ceph/projects/v4_phase3_paper/analysis/twas/'+\
                    'transcript_weights/fusion_pgc2/summary_stats/_m/
 ↪fusion_associations.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       BEST.GWAS.ID  \
12743  ENST00000421553  ENSG00000197062  ZSCAN26  0.040779  chr6:28744470:A:G
14531  ENST00000508906  ENSG00000186470   BTN3A2  0.187261  chr6:26463346:G:T

                 EQTL.ID     TWAS.Z        TWAS.P           FDR    Bonferroni  \
12743  chr6:28650974:A:G  12.745212  3.314893e-37  4.880849e-33  4.880849e-33
14531  chr6:26354866:G:A  11.909938  1.050557e-32  7.734201e-29  1.546840e-28

            Type
12743  Transcript
14531  Transcript
```

Exons

[6]:
```
exons = pd.read_csv('/ceph/projects/v4_phase3_paper/analysis/twas/'+\
                    'exon_weights/fusion_pgc2/summary_stats/_m/
 ↪fusion_associations.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	EQTL.ID	TWAS.Z	TWAS.P	FDR	Bonferroni	Type
62254	e385121	ENSG00000168477	TNXB	0.043518	chr6:31793436:G:A	chr6:32253775:G:A	12.941234	2.633644e-38	1.783056e-33	1.783056e-33	Exon
62253	e385001	ENSG00000168477	TNXB	0.044636	chr6:31793436:G:A	chr6:32253775:G:A	12.728702	4.095902e-37	1.386524e-32	2.773049e-32	Exon

Junctions

[7]:
```
dj_file = '../../../differential_expression/_m/junctions/diffExpr_szVctl_full.
 ↪txt'
dj = pd.read_csv(dj_file, sep='\t', index_col=0)
dj = dj[['Symbol', 'ensemblID']]

jannot_file = '/ceph/projects/v4_phase3_paper/analysis/twas/_m/junctions/
 ↪jxn_annotation.tsv'
jannot = pd.read_csv(jannot_file, sep='\t', index_col=1)

jannot = jannot[['JxnID']]
annot = pd.merge(jannot, dj, left_index=True, right_index=True)

juncs = pd.read_csv('/ceph/projects/v4_phase3_paper/analysis/twas/'+\
                    'junction_weights/fusion_pgc2/summary_stats/_m/
 ↪fusion_associations.txt', sep='\t')
juncs = pd.merge(annot, juncs, left_on='JxnID', right_on='FILE')
juncs = juncs[['FILE', 'ensemblID', 'Symbol', 'HSQ', 'BEST.GWAS.ID', 'EQTL.ID',
               'TWAS.Z', 'TWAS.P', 'FDR', 'Bonferroni']]
juncs['Type'] = 'Junction'
juncs.rename(columns={'Symbol': 'ID', 'FILE': 'Feature'}, inplace=True)
juncs.sort_values('TWAS.P').head(2)
```

```
/usr/lib/python3.9/site-packages/IPython/core/interactiveshell.py:3146:
DtypeWarning: Columns (2) have mixed types.Specify dtype option on import or set
low_memory=False.
  has_raised = await self.run_ast_nodes(code_ast.body, cell_name,
```

[7]:

	Feature	ensemblID	ID	HSQ	BEST.GWAS.ID
19664	j125659	NaN	NaN	0.148096	chr6:31204374:T:C

```
18979  j122115  ENSG00000137411  VARS2  0.118039  chr6:31348749:T:C

                 EQTL.ID     TWAS.Z        TWAS.P           FDR    Bonferroni  \
19664  chr6:31229085:G:A -12.920964  3.428198e-38  8.003127e-34  8.003127e-34
18979  chr6:30951614:G:A  12.375775  3.534662e-35  3.201745e-31  8.251668e-31

          Type
19664  Junction
18979  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:           8211
Transcript:     14724
Exon:           67703
Junction:       23345

===Unique Ensembl Gene===
Gene:           8211
Transcript:     8996
Exon:           10656
```

```
Junction:       5908

===Unique Gene Name===
Gene:           8210
Transcript:     8989
Exon:           12543
Junction:       5906
```

1.2.2 Overlap

[9]:
```
features = {
    'Genes': set(genes['ensemblID']),
    'Transcripts': set(trans['ensemblID']),
    'Exons': set(exons['ensemblID']),
    'Junctions': set(juncs['ensemblID']),
}

limiting_features(features, 'Genes', 'Transcripts')
limiting_features(features, 'Genes', 'Junctions')
limiting_features(features, 'Exons', 'Genes')
print("\n")
limiting_features(features, 'Transcripts', 'Junctions')
limiting_features(features, 'Exons', 'Transcripts')
limiting_features(features, 'Exons', 'Junctions')
```

```
Comparing Genes with Transcripts: 66.08%
Features in common: 5945
Comparing Genes with Junctions: 65.06%
Features in common: 3844
Comparing Exons with Genes: 87.87%
Features in common: 7215


Comparing Transcripts with Junctions: 72.07%
Features in common: 4258
Comparing Exons with Transcripts: 78.57%
Features in common: 7068
Comparing Exons with Junctions: 88.86%
Features in common: 5250
```

[10]:
```
len(features['Genes'] & features['Transcripts'] & features['Exons'] &␣
↪features['Junctions'])
```

[10]: 3257

[11]:
```
len(features['Genes'] | features['Transcripts'] | features['Exons'] |␣
↪features['Junctions'])
```

[11]: 13693

1.2.3 SNPs not in significant PGC2+COLUZK GWAS

[12]:
```
new_genes = pd.merge(genes, pgc2_df, left_on='BEST.GWAS.ID',␣
 ↪right_on='our_snp_id', suffixes=['_TWAS', '_PGC2'])
new_trans = pd.merge(trans, pgc2_df, left_on='BEST.GWAS.ID',␣
 ↪right_on='our_snp_id', suffixes=['_TWAS', '_PGC2'])
new_exons = pd.merge(exons, pgc2_df, left_on='BEST.GWAS.ID',␣
 ↪right_on='our_snp_id', suffixes=['_TWAS', '_PGC2'])
new_juncs = pd.merge(juncs, pgc2_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()

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:           2800
Transcript:     3005
Exon:           3727
Junction:       2994
```

[13]:
```
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']))
```

[13]: 4345

1.3 TWAS P-value < 0.05

1.3.1 Feature summary

[14]:
```
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:           1660
Transcript:     3115
Exon:           14327
Junction:       4890

===Unique Ensembl Gene===
Gene:           1660
Transcript:     2301
Exon:           3547
Junction:       1925

===Unique Gene Names===
Gene:           1660
Transcript:     2299
Exon:           3827
Junction:       1925
```

1.3.2 Overlap

[15]:
```
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']),
}
```

```
limiting_features(features, 'Genes', 'Transcripts')
limiting_features(features, 'Genes', 'Junctions')
limiting_features(features, 'Exons', 'Genes')
print("\n")
limiting_features(features, 'Transcripts', 'Junctions')
limiting_features(features, 'Exons', 'Transcripts')
limiting_features(features, 'Exons', 'Junctions')
```

```
Comparing Genes with Transcripts: 40.81%
Features in common: 939
Comparing Genes with Junctions: 33.61%
Features in common: 647
Comparing Exons with Genes: 79.82%
Features in common: 1325


Comparing Transcripts with Junctions: 45.19%
Features in common: 870
Comparing Exons with Transcripts: 63.71%
Features in common: 1466
Comparing Exons with Junctions: 73.04%
Features in common: 1406
```

[16]:
```
len(features['Genes'] & features['Transcripts'] & features['Exons'] &␣
↪features['Junctions'])
```

[16]:
```
469
```

[17]:
```
len(features['Genes'] | features['Transcripts'] | features['Exons'] |␣
↪features['Junctions'])
```

[17]:
```
5051
```

1.3.3 SNPs not in significant PGC2+COLUZK GWAS

[18]:
```
new_genes = pd.merge(genes[(genes['TWAS.P'] <= 0.05)], pgc2_df, left_on='BEST.
↪GWAS.ID',
                     right_on='our_snp_id', suffixes=['_TWAS', '_PGC2'])
new_trans = pd.merge(trans[(trans['TWAS.P'] <= 0.05)], pgc2_df, left_on='BEST.
↪GWAS.ID',
                     right_on='our_snp_id', suffixes=['_TWAS', '_PGC2'])
new_exons = pd.merge(exons[(exons['TWAS.P'] <= 0.05)], pgc2_df, left_on='BEST.
↪GWAS.ID',
                     right_on='our_snp_id', suffixes=['_TWAS', '_PGC2'])
new_juncs = pd.merge(juncs[(juncs['TWAS.P'] <= 0.05)], pgc2_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()

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:           811
Transcript:     1059
Exon:           1600
Junction:       1127
```

[19]:
```
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']))
```

[19]: 2045

1.4 TWAS FDR < 0.05

1.4.1 Feature summary

[20]:
```
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:           684
Transcript:     1321
Exon:           5776
Junction:       2057

===Unique Ensembl Gene===
Gene:           684
Transcript:     990
Exon:           1511
Junction:       854

===Unique Gene Name===
Gene:           684
Transcript:     990
Exon:           1605
Junction:       854
```

1.4.2 Overlap

[21]:
```
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']),
}

limiting_features(features, 'Genes', 'Transcripts')
limiting_features(features, 'Genes', 'Junctions')
limiting_features(features, 'Exons', 'Genes')
print("\n")
limiting_features(features, 'Transcripts', 'Junctions')
limiting_features(features, 'Exons', 'Transcripts')
limiting_features(features, 'Exons', 'Junctions')
```

```
Comparing Genes with Transcripts: 40.10%
Features in common: 397
Comparing Genes with Junctions: 30.33%
Features in common: 259
Comparing Exons with Genes: 78.80%
Features in common: 539
```

```
Comparing Transcripts with Junctions: 41.92%
Features in common: 358
Comparing Exons with Transcripts: 61.31%
Features in common: 607
Comparing Exons with Junctions: 68.03%
Features in common: 581
```

[22]:
```
len(features['Genes'] & features['Transcripts'] & features['Exons'] &␣
↪features['Junctions'])
```

[22]:
```
195
```

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

[23]:
```
2237
```

1.4.3 SNPs not in significant PGC2+CLOZUK GWAS

[24]:
```
new_genes = pd.merge(genes[(genes['FDR'] <= 0.05)], pgc2_df, left_on='BEST.GWAS.
↪ID',
                     right_on='our_snp_id', suffixes=['_TWAS', '_PGC2'])
new_trans = pd.merge(trans[(trans['FDR'] <= 0.05)], pgc2_df, left_on='BEST.GWAS.
↪ID',
                     right_on='our_snp_id', suffixes=['_TWAS', '_PGC2'])
new_exons = pd.merge(exons[(exons['FDR'] <= 0.05)], pgc2_df, left_on='BEST.GWAS.
↪ID',
                     right_on='our_snp_id', suffixes=['_TWAS', '_PGC2'])
new_juncs = pd.merge(juncs[(juncs['FDR'] <= 0.05)], pgc2_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()

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:           274
```

```
Transcript:     403
Exon:           632
Junction:       447
```

[25]:
```
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']))
```

[25]: 854

1.5 TWAS Bonferroni < 0.05

1.5.1 Feature summary

[26]:
```
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:           143
Transcript:     264
Exon:           854
Junction:       417

===Unique Ensembl Gene===
Gene:           143
Transcript:     212
Exon:           227
```

```
Junction:       151

===Unique Gene Name===
Gene:           143
Transcript:     212
Exon:           237
Junction:       151
```

1.5.2 Overlap

[27]:
```
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']),
}

limiting_features(features, 'Genes', 'Transcripts')
limiting_features(features, 'Genes', 'Junctions')
limiting_features(features, 'Exons', 'Genes')
print("\n")
limiting_features(features, 'Transcripts', 'Junctions')
limiting_features(features, 'Exons', 'Transcripts')
limiting_features(features, 'Exons', 'Junctions')
```

```
Comparing Genes with Transcripts: 38.21%
Features in common: 81
Comparing Genes with Junctions: 29.80%
Features in common: 45
Comparing Exons with Genes: 69.93%
Features in common: 100


Comparing Transcripts with Junctions: 45.70%
Features in common: 69
Comparing Exons with Transcripts: 53.30%
Features in common: 113
Comparing Exons with Junctions: 59.60%
Features in common: 90
```

[28]:
```
len(features['Genes'] & features['Transcripts'] & features['Exons'] &␣
 ↪features['Junctions'])
```

[28]: 32

[29]:
```
len(features['Genes'] | features['Transcripts'] | features['Exons'] |␣
↪features['Junctions'])
```

[29]: 399

1.5.3 SNPs not in significant PGC2+CLOZUK GWAS

[30]:
```
new_genes = pd.merge(genes[(genes['Bonferroni'] <= 0.05)], pgc2_df,␣
↪left_on='BEST.GWAS.ID',
                     right_on='our_snp_id', suffixes=['_TWAS', '_PGC2'])
new_trans = pd.merge(trans[(trans['Bonferroni'] <= 0.05)], pgc2_df,␣
↪left_on='BEST.GWAS.ID',
                     right_on='our_snp_id', suffixes=['_TWAS', '_PGC2'])
new_exons = pd.merge(exons[(exons['Bonferroni'] <= 0.05)], pgc2_df,␣
↪left_on='BEST.GWAS.ID',
                     right_on='our_snp_id', suffixes=['_TWAS', '_PGC2'])
new_juncs = pd.merge(juncs[(juncs['Bonferroni'] <= 0.05)], pgc2_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()

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:           14
Transcript:     26
Exon:           19
Junction:       22
```

[31]:
```
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']))
```

[31]: 52

1.6 Joint analysis

1.6.1 Prepare data

Genes

[32]:
```
genes = pd.read_csv('/ceph/projects/v4_phase3_paper/analysis/twas/'+\
                    'gene_weights/fusion_pgc2/summary_stats/_m/
 ↪fusion_twas_joint_assoc.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', 'TWAS.Z', 'TWAS.P', "JOINT.Z", "JOINT.
 ↪P"]]
genes['Type'] = 'Gene'
genes.rename(columns={'FILE': 'Feature'}, inplace=True)
genes.sort_values('JOINT.P').head(2)
```

[32]:

	Feature	ensemblID	ID	TWAS.Z	\
112	ENSG00000137411	ENSG00000137411	VARS2	9.558187	
29	ENSG00000261353	ENSG00000261353	ENSG00000261353	9.381070	

	TWAS.P	JOINT.Z	JOINT.P	Type
112	1.198372e-21	22.633593	2.024222e-113	Gene
29	6.530637e-21	18.809150	6.354831e-79	Gene

Transcripts

[33]:
```
trans = pd.read_csv('/ceph/projects/v4_phase3_paper/analysis/twas/'+\
                    'transcript_weights/fusion_pgc2/summary_stats/_m/
 ↪fusion_twas_joint_assoc.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', 'TWAS.Z', 'TWAS.P', "JOINT.Z", "JOINT.
 ↪P"]]
trans['Type'] = 'Transcript'
trans.rename(columns={'FILE': 'Feature'}, inplace=True)
trans.sort_values('JOINT.P').head(2)
```

[33]:

	Feature	ensemblID	ID	TWAS.Z	TWAS.P	\
53	ENST00000433076	ENSG00000241370	RPP21	9.807137	1.049029e-22	
79	ENST00000426643	ENSG00000228962	HCG23	-10.807684	3.165626e-27	

	JOINT.Z	JOINT.P	Type
53	26.378509	2.417846e-153	Transcript
79	-25.798882	9.127039e-147	Transcript

Exons

[34]:
```
exons = pd.read_csv('/ceph/projects/v4_phase3_paper/analysis/twas/'+\
                    'exon_weights/fusion_pgc2/summary_stats/_m/
 ↪fusion_twas_joint_assoc.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', 'TWAS.Z', 'TWAS.P', "JOINT.Z", "JOINT.
 ↪P"]]
exons['Type'] = 'Exon'
exons.rename(columns={'FILE': 'Feature'}, inplace=True)
exons.sort_values('JOINT.P').head(2)
```

[34]:

	Feature	ensemblID	ID	TWAS.Z	TWAS.P	JOINT.Z	\
147	e384607	ENSG00000244731	C4A	11.164919	6.054653e-29	-28.484721	
195	e805810	ENSG00000156414	TDRD9	-4.531112	5.867408e-06	-18.385187	

	JOINT.P	Type
147	1.811364e-178	Exon
195	1.726436e-75	Exon

Junctions

[35]:
```
dj_file = '../../../differential_expression/_m/junctions/diffExpr_szVctl_full.
 ↪txt'
dj = pd.read_csv(dj_file, sep='\t', index_col=0)
dj = dj[['Symbol', 'ensemblID']]

jannot_file = '/ceph/projects/v4_phase3_paper/analysis/twas/_m/junctions/
 ↪jxn_annotation.tsv'
jannot = pd.read_csv(jannot_file, sep='\t', index_col=1)

jannot = jannot[['JxnID']]
annot = pd.merge(jannot, dj, left_index=True, right_index=True)

juncs = pd.read_csv('/ceph/projects/v4_phase3_paper/analysis/twas/'+\
                    'junction_weights/fusion_pgc2/summary_stats/_m/
 ↪fusion_twas_joint_assoc.txt', sep='\t')
juncs = pd.merge(annot, juncs, left_on='JxnID', right_on='FILE')
juncs = juncs[['FILE', 'ensemblID', 'Symbol', 'TWAS.Z', 'TWAS.P', "JOINT.Z",␣
 ↪"JOINT.P"]]
juncs['Type'] = 'Junction'
juncs.rename(columns={'Symbol': 'ID', 'FILE': 'Feature'}, inplace=True)
juncs.sort_values('JOINT.P').head(2)
```

```
/usr/lib/python3.9/site-packages/IPython/core/interactiveshell.py:3146:
DtypeWarning: Columns (2) have mixed types.Specify dtype option on import or set
low_memory=False.
```

```
    has_raised = await self.run_ast_nodes(code_ast.body, cell_name,
```

[35]:

	Feature	ensemblID	ID	TWAS.Z	TWAS.P	JOINT.Z	\
158	j121894	ENSG00000186470	BTN3A2	11.715968	1.055787e-31	36.408550	
157	j121892	ENSG00000186470	BTN3A2	9.535201	1.495956e-21	34.610465	

	JOINT.P	Type
158	3.117598e-290	Junction
157	1.758359e-262	Junction

1.6.2 Feature summary

[36]:
```
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:           161
Transcript:     214
Exon:           247
Junction:       213

===Unique Ensembl Gene===
Gene:           161
Transcript:     210
Exon:           222
Junction:       183
```

```
===Unique Gene Name===
Gene:           161
Transcript:     210
Exon:           222
Junction:       183
```

1.6.3 Overlap

[37]:
```
features = {
    'Genes': set(genes['ensemblID']),
    'Transcripts': set(trans['ensemblID']),
    'Exons': set(exons['ensemblID']),
    'Junctions': set(juncs['ensemblID']),
}

limiting_features(features, 'Genes', 'Transcripts')
limiting_features(features, 'Genes', 'Junctions')
limiting_features(features, 'Exons', 'Genes')
print("\n")
limiting_features(features, 'Transcripts', 'Junctions')
limiting_features(features, 'Exons', 'Transcripts')
limiting_features(features, 'Exons', 'Junctions')
```

```
Comparing Genes with Transcripts: 28.10%
Features in common: 59
Comparing Genes with Junctions: 25.68%
Features in common: 47
Comparing Exons with Genes: 47.83%
Features in common: 77


Comparing Transcripts with Junctions: 27.87%
Features in common: 51
Comparing Exons with Transcripts: 32.38%
Features in common: 68
Comparing Exons with Junctions: 41.53%
Features in common: 76
```

[38]:
```
len(features['Genes'] & features['Transcripts'] & features['Exons'] &␣
↪features['Junctions'])
```

[38]: 23

[39]:
```
len(features['Genes'] | features['Transcripts'] | features['Exons'] |␣
↪features['Junctions'])
```

[39]: 513

1.7 Session Information

[40]:
```
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
```

[41]:
```
print(sys_info())
#print the names and versions of the imported modules
print("\nImported Modules:")
for module in pip_modules[2:]:
    name, version = module.split('==')
    if name in imported_modules:
        print(name + ':\t' + version)
```

```
{'commit_hash': '<not found>',
 'commit_source': '(none found)',
 'default_encoding': 'utf-8',
 'ipython_path': '/usr/lib/python3.9/site-packages/IPython',
 'ipython_version': '7.19.0',
 'os_name': 'posix',
 'platform': 'Linux-5.10.14-arch1-1-x86_64-with-glibc2.33',
 'sys_executable': '/usr/bin/python',
 'sys_platform': 'linux',
 'sys_version': '3.9.1 (default, Feb  6 2021, 06:49:13) \n[GCC 10.2.0]'}

Imported Modules:
pandas: 1.1.5
```