Genomic analysis of GI-related variants for John M

Dima Lituiev

Reducing the search space of genetic variants

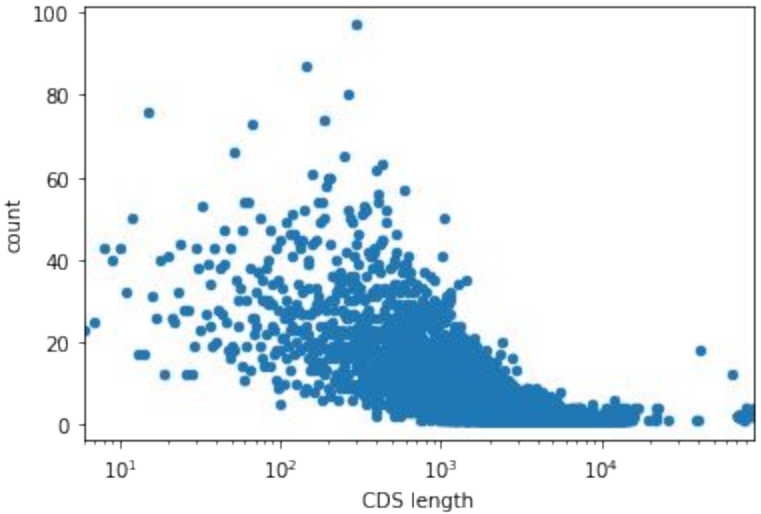
- Remove low quality variants and heterozygous loci

  ~ 2M variants (1995031)
- Annotate using SnpEff
- Filter for genes highly expressed in GI tissues (714 genes)

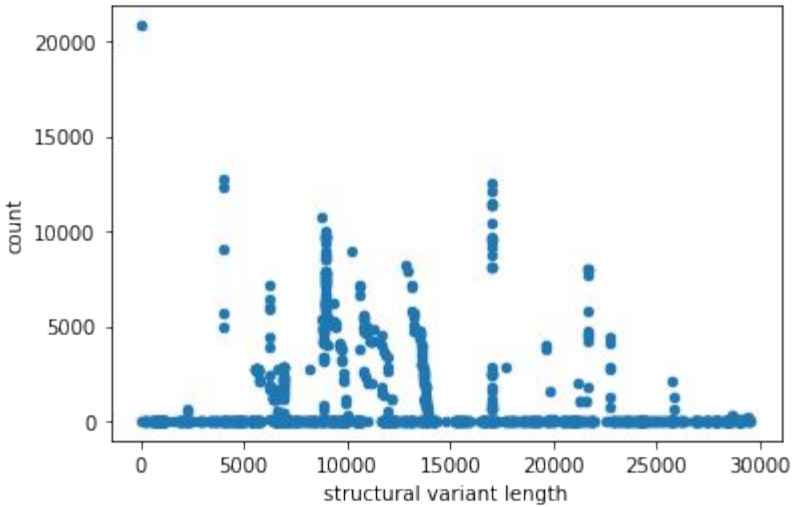
  ~ 35K variants (34650)

Distribution of deletion lengths

protein coding sequences



all deletions



Selecting GI-specific genes

THE HUMAN PROTEIN ATLAS

Filter genes highly expressed in stomach, duodenum, and small intestine using reference RNA expression dataset

https://www.proteinatlas.org/about/download

```
positive = {'stomach', 'duodenum', 'small intestine'}
negative = {'lung', 'skeletal muscle', 'prostate', 'skin',
            'breast', 'ovary', 'bone marrow', 'adipose tissue'}
```

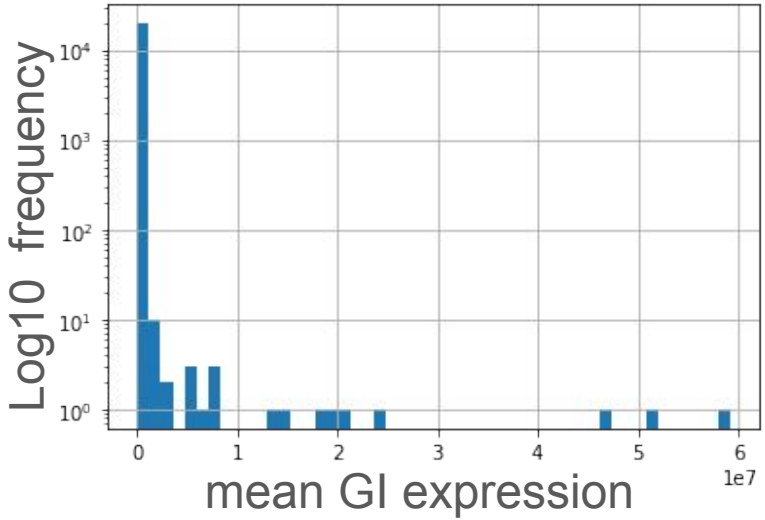
```
neg_mask = protein_atlas.Sample.map(lambda x: x in negative)
pos_mask = protein_atlas.Sample.map(lambda x: x in positive)
```

```
mean_neg_expression = protein_atlas[neg_mask].groupby(['Gene name'])['Value'].mean()
var_neg_expression = protein_atlas[neg_mask].groupby(['Gene name'])['Value'].var()

var_total_expression = protein_atlas.groupby(['Gene name'])['Value'].var()
```

```
pos_expression = protein_atlas[pos_mask].groupby(['Sample','Gene name'])['Value'].mean()
```

```
# mean_neg_expression
# protein_atlas
eps = var_total_expression[var_neg_expression>0].min()/100
diff_expression = (pos_expression - mean_neg_expression)/(var_neg_expression+eps)
diff_expression = diff_expression.swaplevel(0,1).unstack()
```



Analysis of GI-specific genes

- Intersect with dbSNP:

  24 variants (non-benign: 3 variants, all non-remarkable)

- Intersect with OMIM database pathogenic genes:

  58 variants:

  - HNF4A: Maturity onset diabetes of the young (MODY)
  - PAX4: MODY type IX, ketosis-prone diabetes
  - ALG14: Myasthenic syndrome, congenital
  - IL23R: IBD

- MTTP deletion/nonsense-mediated decay:

  Abetalipoproteinemia, a fat absorption disorder

  The first symptoms of abetalipoproteinemia appear in infancy, including failure to gain weight and grow at the expected rate (failure to thrive)
- SLC10A2 (3’UTR variant): Bile acid malabsorption
- TREH (17033 nt deletion): Trehalase deficiency: may lead to vomiting, abdominal discomfort and diarrhea after eating foods containing trehalose. *Trehalose occurs naturally in mushrooms, honey, lobsters, shrimps, certain seaweeds (algae), wine, beer, bread and other foods produced by using baker's or brewer's yeast.*

- A4GNT (370 aa deletion, chr3):
    - Necessary for the synthesis of type III mucin which is specifically produced in the stomach, duodenum, and pancreatic duct (PubMed:10430883). May protect against inflammation-associated gastric adenocarcinomas.
- MUC13 (1240 aa deletion, chr3): IBD risk
- INSM1 (25810 nt deletion, chr29)
    - Promotes the generation and expansion of neuronal basal progenitor cells in the developing neocortex. Involved in the differentiation of endocrine cells of the developing anterior pituitary gland, of the pancreas and intestine, and of sympatho-adrenal cells in the peripheral nervous system. Promotes cell cycle signaling arrest and inhibition of cellular proliferation.

Future directions

Confirm deficiencies in proteomics data for:

- mucin 3 and 13,
- IL23R (IBD autoimmunity risk)
- MTTP (fat malabsorption)
- SLC10A2 (bile)

Confirm trehalase activity, take supplement enzyme(?), exclude trehalose

Monitor MODY symptoms, confirm gene/isoform expression levels for:

- HNF4A
- PAX4