GTEx Robokop-interfaces discovery, design and implementation

Description of the GTEx project (<https://gtexportal.org/home/>)

The Genotype-Tissue Expression (GTEx) project is an ongoing effort to build a comprehensive public resource to study tissue-specific gene expression and regulation. Samples were collected from 53 non-diseased tissue sites across nearly 1000 individuals, primarily for molecular assays including WGS, WES, and RNA-Seq. Remaining samples are available from the GTEx Biobank. The GTEx Portal provides open access to data including gene expression, QTLs, and histology images.

GTEx – The Genotype-Tissue Expression project

<https://www.nature.com/articles/ng.2653>

The Common Fund's**Genotype-Tissue Expression (GTEx) Program**established a data resource and tissue bank to study the relationship between genetic variation and gene expression in multiple human tissues. GTEx is also examining sex-based differences in how genes are turned on and off and how they are regulated. GTEx resources are valuable tools for exploring the genetic basis of complex human diseases.

GTEx data is:

* Released in versions
* Current version is v7.
* 48 non-diseased tissue sites across 1000 individuals.
* Data is using reference genome hg37.

Available GTEx data sets:

* dbGaP data used, BAM, VCF, etc. available direct.
* Annotations in the form of Excel files.
* RNA-Seq data
* Single-tissue cis-eQTL data – The target of this effort
* Multi-tissue cis-eQTL data
* Reference data (used for analysis)

Ensembl hosted feeds:

* Ensembl has exposed 3 rest eQTL APIs that interface with the GTEx data.
  + Returns the p-value for each SNP in a given gene (e.g. ENSG00000227232)
  + Returns the p-values for a SNP (e.g. rs123)
  + Returns all tissues currently available in the DB
* Ensembl may not provide the correct information for our purposes as they have loaded an older data version (6 vs. 7).

GTEx data details:

GTEx Datasets and details: <https://gtexportal.org/home/datasets>

No Y chromosomes were found in the data.

GTEx hosted data feeds:

* The website allows for browsing by gene ID, variant, tissue, histology images, expression and QTL.
* Data files can be in the format of Excel spreadsheets, text files, compressed GCT and BED files.
* These feeds includes mitochondrial DNA reads. Mitochondrial DNA reads are not included in the eGenes and significant gene/allele data sets.
* REST based API calls for GTEx data located at: <https://gtexportal.org/home/api-docs/#/>
  + - * + [Find gene expression PCA data.](https://gtexportal.org/home/api-docs/#!/expression/expressionPca)
        + [Find normalized gene expression data.](https://gtexportal.org/home/api-docs/#!/expression/geneExpression)
        + [Find median exon expression data.](https://gtexportal.org/home/api-docs/#!/expression/medianExonExpression)
        + [Find median gene expression data.](https://gtexportal.org/home/api-docs/#!/expression/medianGeneExpression)
        + [Find median gene expression data.](https://gtexportal.org/home/api-docs/#!/expression/medianJunctionExpression)
        + [Find median transcript expression data of all known transcripts of a gene.](https://gtexportal.org/home/api-docs/#!/expression/medianTranscriptExpression)
        + [Find top expressed genes for a specified tissue.](https://gtexportal.org/home/api-docs/#!/expression/topExpressedGene)
        + [Find transcript expression data at the sample level.](https://gtexportal.org/home/api-docs/#!/expression/transcriptExpression)

GTEx single-tissue flat files (GTEx\_Analysis\_v7\_eQTL archive, tab delimited):

* eGene and significant variant-gene associations based on permutations. The archive contains a <tissue>.egenes.txt.gz and <tissue>.signif\_variant\_gene\_pairs.txt.gz file for each tissue.
* Note that the \*.egenes.txt.gz files contain data for all genes tested; to obtain the list of eGenes, select the rows with 'qval' ≤ 0.05.

Tissue name and uberon data elements are added to each record for the 2 datasets noted above. Uberons were collected manually using EMBL-EBI (<https://www.ebi.ac.uk/ols/index>).

eGene and significant variant-gene association data elements

* eGene data elements: gene\_id, gene\_name, gene\_chr, gene\_start, gene\_end, strand, num\_var, beta\_shape1, beta\_shape2, true\_df, pval\_true\_df, variant\_id, tss\_distance, chr, pos, ref, alt, num\_alt\_per\_site, rs\_id\_dbSNP147\_GRCh37p13, minor\_allele\_samples, minor\_allele\_count, maf, ref\_factor, pval\_nominal, slope, slope\_se, pval\_perm, pval\_beta, qval, pval\_nominal\_threshold, log2\_aFC, log2\_aFC\_lower, log2\_aFC\_upper
* significant variant-gene association data elements: variant\_id, gene\_id, tss\_distance, ma\_samples, ma\_count, maf, pval\_nominal, slope, slope\_se, pval\_nominal\_threshold, min\_pval\_nominal, pval\_beta

GTEx SmartBag - Development and deployment

GTEx Smartbag - Development settings – smartBag project:

* IDE: PyCharm.
* Github: https://github.com/NCATS-Tangerine/smartBag.git
* Branch: phil\_smartbag

Deployment steps for GTEx SmaptAPI:

* The “deploy” bash script runs all steps to create smartBag API, as well as create and launch docker container image.
* Individual steps
  + Create SmartAPI definition of endpoints
  + Add SmartAPI definition to github at: <https://github.com/NCATS-Gamma/translator-api-registry.git>
  + Determine container host and deploy API website there.
  + Register Smart API prod instance at: <https://smart-api.info/add_api>
    - FQDN of target server has to be known first

GTEx Robocop - Development and deployment

GTEx robokop services and graph database pre-loading

Robokop-interface: Development environment

* IDE: PyCharm.
* Github: https://github.com/NCATS-Gamma/robokop-interfaces.git
* Branch: GTEx\_inclusion
* greent.conf file modifications to connect to development servers
* Element for redis cache
* Redis cache database
  + host: robokopdb2.renci.org  
    port: 6380  
    db: 0  
    password: "stayoutoftrouble"
* Element rosetta-graph
* rosetta-graph:
* url: "bolt://robokopdev.renci.org:7687"
* neo4j\_password: "ncatsgamma”

Robokop-interface development approach:

* Information should be represented by 3 relationships (variant –> gene, variant –> anatomy, gene –> anatomy).
* Important GTEx data elements are associated to edges as properties.
* There are no P-value threshold cutoffs. The GTEx significant gene/variant association data (noted above) should be used in its entirety.

Potential target code files (create or update) for GTEx inclusion:

|  |  |  |
| --- | --- | --- |
| **Operation** | **Location (robokop-interfaces/)** | **Current status** |
| Service configurations (file updates) | greent/greent-\*.conf | Updated |
| Source map (file update) | greent/conf/source\_map.json | Updated. |
| Core (file update) | greent/core.py | Updated. |
| Service | greent/services/gtexcatalog.py | File created. Needs further investigation. |
| Data concepts and operators (file update) | greent/rosetta.yml | Predicated on GTEx service definition. Anatomical “tissue to gene expression” needed? |
| Builder (leverage/addition to OBH?) | builder/gtex\_builder.py | Created. |
| Test | greent/test/test\_gtex.py | Created. |
| Biolink model overlay (file update) | greent/conf/biolink-model\_overlay.yaml | Needs further investigation. Add to “affects expression of”? |
| Predicates (file update) | greent/conf/predicates.json | Needs further investigation. Add tissue anatomy to “sequence\_variant” predicate? |
| Source map | greent/conf/source\_map.json | Needs further investigation. |
| Uber context | greent/conf/uber\_context.jsonId | Needs further investigation. |
| Node type | greent/node\_types.py | use ANATOMICAL\_ENTITY, SEQUENCE\_VARIANT, GENE |

GTEx node types (if not present)

* Sequence variant
* Gene
* Anatomical entity

GTEx graph edges (add if not present?):

* Variant to Gene
* Variant to Anatomy
* Gene to Anatomy

GTEx graph edge attributes:

* Uberon code, Tissue name , p/beta vals, MAF?

GTEx graph Other relationship entities?:

* Concept
* Genetic condition
* Biological process or activity
* Disease
* Cell