Variant and Orthology Service

The goal of this document is to develop a schema design for the service in a graph database structure. A rough specification for the variant+ortholog service is also presented.

- Rough Specifications
 - Additional Notes
- Data Model and Index Choices
 - Database Indexes
 - Indexed Fields
 - Variant_transcript_rel.clinical_significance
 - Variants.variant_effect
 - · Genes.biotype
 - Transcripts.biotypes
 - Testing Database

Rough Specifications

This page just lists some of our rough thoughts on what the service should do. Tim has also provided access to a set of scripts for converting the Ensembl data to a database: https://github.com/treynr/gripv

The key endpoint has the following inputs:

source species, source genome build, rs ID, target species, target genome build, filter on variant effects

Phase I

Return a list of genes associated with the variant (within, upstream cis, upstream trans, downstream cis, downstream trans)

There is an ISMB meeting. submission for the tech track is due by May 9th, 2019.

Genes

- · make sure they are all the species that we want for GeneWeaver
- make a list of all the biotypes ordered by prevalence, Gaurab can help with definitions

Orthologs

· Getting the AGR orthology data will be important, but not necessary for phase I, other sources to follow, too

Transcripts

· make a list of all the biotypes ordered by prevalence, Gaurab can help with definitions

Variants

- · keep all the alleles, not just the SNPs
- · for now it's OK to just use the upstream/downstream assessment from Ensembl, this may change when we start adding epigenetic data

Queries

- query based on Ensembl IDs
- filter on variant_effect from the variant table
- user may have another identifier that we would have to have a mapping from that to the Ensembl ID. We will have to figure out how to handle all
 the mappings to/from GW IDs also.
- from human GWAS study, if we roll the variants up to the gene level, then how similar are the human and mouse genes?
 - How would we measure the similarity? Start with Jaccard similarity between the ortholog cluster IDS for the genesets.
- what are the variants between mouse and human for this given gene?

UI

- how to we want to show these data in GW?
- enumerate variants in the orthologous set?
- Extract from Elissa and Erich what they have in mind for the user interface.

Phase II

For each gene, also return a list of orthologs

Update the process for rapidly updating the builds. This should be done twice a year at a minimum.

Phase III

Provide a link / URL to mouse variant allele registry database (MVARD)

Phase IV

Provide a UI on top of the swagger-documented page.

Phase V

Include gene regulation information

This is other ways that variants impact genes. Tim is currently pulling data from the Ensembl regulatory build.

Additional Notes

We will need to update the GW database every 6 months.

Each time we do a new build we will not throw anything away. Each new build will map the new build to existing genesets. We have to map everything, partial updates are no good.

Check out mappers from a relational database to a graph database.

Explore the use of neo4j (or other graph database) for storing the gene/ortholog/variant triples. Look at caps (upper bound) on the number of triples that we can store. Can postgres 11 handle this? We're dubious...

Data Model and Index Choices

The entity-relation diagram for the varolog graph database is shown below. Early in the development of the database we need to make some choices about which fields to index, and how to further normalize the graph DB.

variant_transcript_rel trn_id char(24) -CAUSES trn_id integer integer build integer varchar variants char(64) src_id integer char(24) var_id integer ref allele varcahr score float alt_allele alt_allele_freq float bio id integer char(8) start integer char(24) gene_id char(64) char(8) transcripts char intege end integer integer strand char(8) char uid integer eff_id char char bio_id build integer varchar var id varchar - MODIFIES -- MODIFIES

Graph Database ERD v2

Make sure we also store variants that may not be associated with a gene.

Use bio type to condition on the entities

Generic bipartite representation may not be the best approach for the long term (for all tools).

We can map biotypes on to one another.

Can this capture all the relevant regulatory relationships? TADs? Epigenetic marks? Chromatin structures? Distal enhancers? (Jake talk with Tim about this)

* Yellow highlighted fields are to be indexed

Database Indexes

Indexes in graph databases are a lot like those in relational databases. They do take up storage space, and they also significantly speed up query performance. Since theses relations will hold so much data (~200GB) we prefer to limit the number of indexes defined in the schema. Only the key query fields will be indexed. The tables below show the current values for the fields that we expect to index.

Indexed Fields

Variant_transcript_rel.clinical_significance

transcript	1338078
primary_transcri pt	1313212
ncRNA	517205
mRNA	285680
miRNA	40

Variants.variant_effect

intron_variant	253555
non_coding_tran script_variant	119122
downstream_ge ne_variant	109435
upstream_gene_ variant	103619
NMD_transcript_ variant	15548
non_coding_tran script_exon_vari ant	10755
3_prime_UTR_v ariant	8182
missense_variant	7338
synonymous_va riant	5406
5_prime_UTR_v ariant	1874
frameshift_varia nt	282
coding_sequenc e_variant	248
stop_gained	178
inframe_deletion	141
splice_donor_va riant	124
splice_acceptor_ variant	103
inframe_insertion	57
start_lost	16
stop_lost	13
stop_retained_v ariant	5

mature_miRNA_ variant	4
splice_region_va riant	4
start_retained_v ariant	2
protein_altering_ variant	1

protein_coding	Genes.biotype	
dogene lincRNA 12800 antisense_RNA 8244 unprocessed_ps eudogene TEC 4154 miRNA 4081 snRNA 3283 misc_RNA 2775 snoRNA 2450 processed_trans cript sense_intronic 1214 transcribed_unpr ocessed_pseud ogene rRNA 897 transcribed_proc essed_pseudog ene IG_V_gene 362 IG_V_pseudoge 343 ne TR_V_gene 252 sense_overlappi ng bidirectional_pro moter_IncRNA 17R_J_gene 149 polymorphic_ps eudogene unitary_pseudoge 117 ene scaRNA 100	protein_coding	41781
antisense_RNA 8244 unprocessed_ps eudogene TEC 4154 miRNA 4081 snRNA 3283 misc_RNA 2775 snoRNA 2450 processed_trans cript sense_intronic 1214 transcribed_unpr ocessed_pseud ogene rRNA 897 transcribed_proc essed_pseudog ene IG_V_gene 362 IG_V_gene 362 IG_V_pseudoge ne TR_V_gene 252 sense_overlappi ng bidirectional_pro moter_IncRNA TR_J_gene 149 polymorphic_ps eudogene transcribed_unit ary_pseudogene 123 unitary_pseudoge 117 ene scaRNA 100		19228
unprocessed_ps eudogene TEC 4154 miRNA 4081 snRNA 3283 misc_RNA 2775 snoRNA 2450 processed_trans cript 1320 sense_intronic 1214 transcribed_unpr ocessed_pseud ogene rRNA 897 transcribed_proc essed_pseudog ene IG_V_gene 362 IG_V_pseudoge ne IG_V_gene 252 sense_overlappi ng bidirectional_pro moter_IncRNA TR_J_gene 149 polymorphic_ps eudogene transcribed_unit ary_pseudoge 117 scaRNA 100	lincRNA	12800
eudogene TEC 4154 miRNA 4081 snRNA 3283 misc_RNA 2775 snoRNA 2450 processed_trans cript sense_intronic 1214 transcribed_unpr ocessed_pseud ogene rRNA 897 transcribed_proc essed_pseudog ene IG_V_gene 362 IG_V_pseudoge ne IG_V_gene 252 sense_overlappi ng bidirectional_pro moter_IncRNA TR_J_gene 149 polymorphic_ps eudogene transcribed_unit ary_pseudogene 123 unitary_pseudoge 117 scaRNA 100	antisense_RNA	8244
miRNA 4081 snRNA 3283 misc_RNA 2775 snoRNA 2450 processed_trans cript sense_intronic 1214 transcribed_unpr ocessed_pseud ogene rRNA 897 transcribed_proc essed_pseudog ene IG_V_gene 362 IG_V_pseudoge ne TR_V_gene 252 sense_overlappi ng bidirectional_pro moter_lncRNA TR_J_gene 149 polymorphic_ps eudogene transcribed_unit ary_pseudoge ene 123 unitary_pseudoge ene 117 scaRNA 100		5205
snRNA 3283 misc_RNA 2775 snoRNA 2450 processed_trans 1320 cript sense_intronic 1214 transcribed_unpr ocessed_pseud ogene rRNA 897 transcribed_proc essed_pseudog ene IG_V_gene 362 IG_V_pseudoge ne IR_V_gene 252 sense_overlappi ng bidirectional_pro moter_lncRNA 153 TR_J_gene 149 polymorphic_ps eudogene 123 unitary_pseudoge 117 ene scaRNA 100	TEC	4154
misc_RNA 2775 snoRNA 2450 processed_trans cript sense_intronic 1214 transcribed_unpr ocessed_pseud ogene rRNA 897 transcribed_proc essed_pseudog ene IG_V_gene 362 IG_V_gene 362 IG_V_pseudoge ne TR_V_gene 252 sense_overlappi ng bidirectional_pro moter_lncRNA TR_J_gene 149 polymorphic_ps eudogene transcribed_unit ary_pseudogene unitary_pseudog 117 scaRNA 100	miRNA	4081
snoRNA 2450 processed_trans cript 1320 sense_intronic 1214 transcribed_unpr ocessed_pseud ogene rRNA 897 transcribed_proc essed_pseudog ene 1G_V_gene 362 IG_V_pseudoge ne 252 sense_overlappi ng bidirectional_pro moter_lncRNA 153 TR_J_gene 149 polymorphic_ps eudogene 123 unitary_pseudoge 117 scaRNA 100	snRNA	3283
processed_trans cript sense_intronic 1214 transcribed_unpr ocessed_pseud ogene rRNA 897 transcribed_proc essed_pseudog ene IG_V_gene 362 IG_V_gene 343 TR_V_gene 252 sense_overlappi ng bidirectional_pro moter_lncRNA 153 TR_J_gene 149 polymorphic_ps eudogene 123 unitary_pseudog 117 scaRNA 100	misc_RNA	2775
cript sense_intronic 1214 transcribed_unpr ocessed_pseud ogene rRNA 897 transcribed_proc essed_pseudog ene IG_V_gene 362 IG_V_pseudoge ne TR_V_gene 252 sense_overlappi ng bidirectional_pro moter_lncRNA TR_J_gene 149 polymorphic_ps eudogene transcribed_unit ary_pseudogene unitary_pseudog 117 scaRNA 100	snoRNA	2450
transcribed_unpr ocessed_pseud ogene rRNA 897 transcribed_proc essed_pseudog ene IG_V_gene 362 IG_V_pseudoge ne TR_V_gene 252 sense_overlappi ng bidirectional_pro moter_lncRNA 153 TR_J_gene 149 polymorphic_ps eudogene transcribed_unit ary_pseudogene 117 scaRNA 100		1320
ocessed_pseud ogene rRNA 897 transcribed_proc essed_pseudog ene IG_V_gene 362 IG_V_pseudoge ne 252 sense_overlappi ng bidirectional_pro moter_lncRNA 153 TR_J_gene 149 polymorphic_ps eudogene 123 unitary_pseudog 117 ene scaRNA 100	sense_intronic	1214
transcribed_proc essed_pseudog ene IG_V_gene 362 IG_V_pseudoge 343 TR_V_gene 252 sense_overlappi 217 mg bidirectional_pro moter_lncRNA 153 TR_J_gene 149 polymorphic_ps eudogene transcribed_unit ary_pseudogene unitary_pseudog ene scaRNA 100	ocessed_pseud	1056
essed_pseudog ene IG_V_gene 362 IG_V_pseudoge 343 TR_V_gene 252 sense_overlappi 217 bidirectional_pro moter_IncRNA TR_J_gene 149 polymorphic_ps eudogene transcribed_unit ary_pseudogene unitary_pseudog 117 scaRNA 100	rRNA	897
IG_V_pseudoge ne 343 ne 252 TR_V_gene 252 sense_overlappi 217 ng bidirectional_pro moter_IncRNA 153 TR_J_gene 149 polymorphic_ps eudogene transcribed_unit ary_pseudogene unitary_pseudog 117 ene scaRNA 100	essed_pseudog	725
ne TR_V_gene 252 sense_overlappi ng 217 bidirectional_pro moter_IncRNA 153 TR_J_gene 149 polymorphic_ps eudogene transcribed_unit ary_pseudogene unitary_pseudog ene scaRNA 100	IG_V_gene	362
sense_overlappi ng bidirectional_pro moter_lncRNA TR_J_gene 149 polymorphic_ps eudogene transcribed_unit ary_pseudogene unitary_pseudog scaRNA 100		343
ng bidirectional_pro moter_IncRNA TR_J_gene 149 polymorphic_ps eudogene transcribed_unit ary_pseudogene unitary_pseudog scaRNA 100	TR_V_gene	252
moter_IncRNA TR_J_gene 149 polymorphic_ps eudogene 141 transcribed_unit ary_pseudogene 117 ene 1100		217
polymorphic_ps eudogene transcribed_unit ary_pseudogene unitary_pseudog 117 ene scaRNA 100		153
transcribed_unit ary_pseudogene unitary_pseudog ene scaRNA 100	TR_J_gene	149
ary_pseudogene unitary_pseudog ene scaRNA 100		141
ene scaRNA 100		123
		117
pseudogene 86	scaRNA	100
	pseudogene	86

TR_V_pseudoge ne IG_D_gene 56 Mt_tRNA 44 3prime_overlapp ing_ncRNA IG_J_gene 32 ribozyme 30 IG_C_gene 27 TR_J_pseudoge ne 14 IG_C_pseudoge 10 TR_D_gene 8 sRNA 7 IG_LV_gene 4 Mt_rRNA 4 IG_D_pseudoge 3 IG_pseudogene 3 IG_seudogene 3 IG_seudogene 3 IG_treated 3 IG_treated 4 IG_treated 56 Id_treated 57 Id		
Mt_tRNA 44 3prime_overlapp ing_ncRNA 33 IG_J_gene 32 ribozyme 30 IG_C_gene 27 TR_J_pseudoge 14 IG_C_pseudoge 10 TR_D_gene 8 sRNA 7 IG_LV_gene 4 Mt_rRNA 4 IG_D_pseudoge 3 ne IG_J_pseudoge 3 ne IG_pseudogene 3 non_coding 3 scRNA 2 translated_proce ssed_pseudoge ne macro_IncRNA 2		64
3prime_overlapp ing_ncRNA IG_J_gene 32 ribozyme 30 IG_C_gene 27 TR_J_pseudoge 14 TR_C_gene 14 IG_C_pseudoge 10 ne 8 sRNA 7 IG_LV_gene 4 Mt_rRNA 4 IG_D_pseudoge 3 ne IG_pseudogene 3 non_coding 3 scRNA 2 translated_proce ssed_pseudoge ne macro_IncRNA 2	IG_D_gene	56
ing_ncRNA IG_J_gene 32 ribozyme 30 IG_C_gene 27 TR_J_pseudoge 14 ne 14 IG_C_pseudoge 10 ne 8 sRNA 7 IG_LV_gene 4 Mt_rRNA 4 IG_D_pseudoge 3 ne 1G_pseudogene IG_pseudogene 3 non_coding 3 scRNA 2 translated_proce 2 ssed_pseudoge 2 macro_IncRNA 2	Mt_tRNA	44
ribozyme 30 IG_C_gene 27 TR_J_pseudoge 14 TR_C_gene 14 IG_C_pseudoge 10 TR_D_gene 8 sRNA 7 IG_LV_gene 4 Mt_rRNA 4 IG_D_pseudoge 3 ne IG_J_pseudoge 3 ne IG_pseudogene 3 scRNA 2 translated_proce ssed_pseudoge ne macro_IncRNA 2		33
IG_C_gene 27 TR_J_pseudoge 14 ne 14 IG_C_pseudoge 10 TR_D_gene 8 sRNA 7 IG_LV_gene 4 Mt_rRNA 4 IG_D_pseudoge 3 ne IG_J_pseudoge 3 ne IG_pseudogene 3 non_coding 3 scRNA 2 translated_proce ssed_pseudoge ne macro_IncRNA 2	IG_J_gene	32
TR_J_pseudoge ne 14 TR_C_gene 14 IG_C_pseudoge 10 TR_D_gene 8 sRNA 7 IG_LV_gene 4 Mt_rRNA 4 IG_D_pseudoge 3 ne IG_J_pseudoge 3 ne IG_pseudogene 3 non_coding 3 scRNA 2 translated_proce ssed_pseudoge ne macro_IncRNA 2	ribozyme	30
ne TR_C_gene 14 IG_C_pseudoge 10 ne TR_D_gene 8 sRNA 7 IG_LV_gene 4 Mt_rRNA 4 IG_D_pseudoge 3 ne IG_J_pseudoge 3 ne IG_pseudogene 3 non_coding 3 scRNA 2 translated_proce ssed_pseudoge ne macro_IncRNA 2	IG_C_gene	27
IG_C_pseudoge ne 10 TR_D_gene 8 sRNA 7 IG_LV_gene 4 Mt_rRNA 4 IG_D_pseudoge 3 ne IG_J_pseudoge 3 ne IG_pseudogene 3 non_coding 3 scRNA 2 translated_proce ssed_pseudoge ne macro_IncRNA 2		14
ne TR_D_gene 8 sRNA 7 IG_LV_gene 4 Mt_rRNA 4 IG_D_pseudoge 3 ne IG_J_pseudoge 3 ne IG_pseudogene 3 non_coding 3 scRNA 2 translated_proce ssed_pseudoge ne macro_IncRNA 2	TR_C_gene	14
sRNA 7 IG_LV_gene 4 Mt_rRNA 4 IG_D_pseudoge 3 ne IG_J_pseudoge 3 ne IG_pseudogene 3 non_coding 3 scRNA 2 translated_proce ssed_pseudoge ne macro_lncRNA 2		10
IG_LV_gene 4 Mt_rRNA 4 IG_D_pseudoge 3 ne 3 IG_J_pseudoge 3 ne 3 IG_pseudogene 3 non_coding 3 scRNA 2 translated_proce ssed_pseudoge ne 2 macro_IncRNA 2	TR_D_gene	8
Mt_rRNA 4 IG_D_pseudoge 3 ne 3 IG_J_pseudoge 3 ne 3 IG_pseudogene 3 non_coding 3 scRNA 2 translated_proce ssed_pseudoge ne 2 macro_lncRNA 2	sRNA	7
IG_D_pseudoge a ne a	IG_LV_gene	4
ne IG_J_pseudoge 3 ne IG_pseudogene 3 non_coding 3 scRNA 2 translated_proce ssed_pseudoge ne macro_IncRNA 2	Mt_rRNA	4
ne IG_pseudogene 3 non_coding 3 scRNA 2 translated_proce ssed_pseudoge ne macro_lncRNA 2		3
non_coding 3 scRNA 2 translated_proce ssed_pseudoge ne	,	3
scRNA 2 translated_proce ssed_pseudoge ne	IG_pseudogene	3
translated_proce ssed_pseudoge ne 2 macro_IncRNA 2	non_coding	3
ssed_pseudoge ne macro_IncRNA 2	scRNA	2
_	ssed_pseudoge	2
vaultRNA 1	macro_IncRNA	2
	vaultRNA	1

Transcripts.biotypes

protein_coding	137554
retained_intron	47410
processed_transcript	43390
lincRNA	21425
nonsense_mediated_decay	20693
processed_pseudogene	19232
antisense_RNA	15144
unprocessed_pseudogene	5215
TEC	4330
miRNA	4081
snRNA	3283
misc_RNA	2792
snoRNA	2462

sense_intronic	1301
transcribed_unprocessed_pseudogene	1075
rRNA	897
transcribed_processed_pseudogene	732
IG_V_gene	445
sense_overlapping	392
IG_V_pseudogene	343
TR_V_gene	302
bidirectional_promoter_IncRNA	275
polymorphic_pseudogene	182
TR_J_gene	149
pseudogene	132
transcribed_unitary_pseudogene	125
unitary_pseudogene	117
non_stop_decay	111
scaRNA	100
TR_V_pseudogene	64
IG_D_gene	56
IG_C_gene	44
Mt_tRNA	44
3prime_overlapping_ncRNA	38
IG_J_gene	32
ribozyme	30
TR_C_gene	16
TR_J_pseudogene	14
translated_processed_pseudogene	14
IG_C_pseudogene	10
TR_D_gene	8
sRNA	7
IG_LV_gene	4
Mt_rRNA	4
non_coding	3
macro_IncRNA	3
IG_pseudogene	3
IG_J_pseudogene	3
IG_D_pseudogene	3
scRNA	2
vaultRNA	1

There is a graph database set up for development with the following connection parameters:

"http://gwdev01:7474/", username="neo4j", password="j4cks0nl4b"

