VarMoVinator

Ben Busby, Simon Rakov, Faryad Shaneh, Maiwase Tembo • 05/16/23

Overview

Objective:

To create a knowledge graph that we will use to build a model to validate Large Language Models

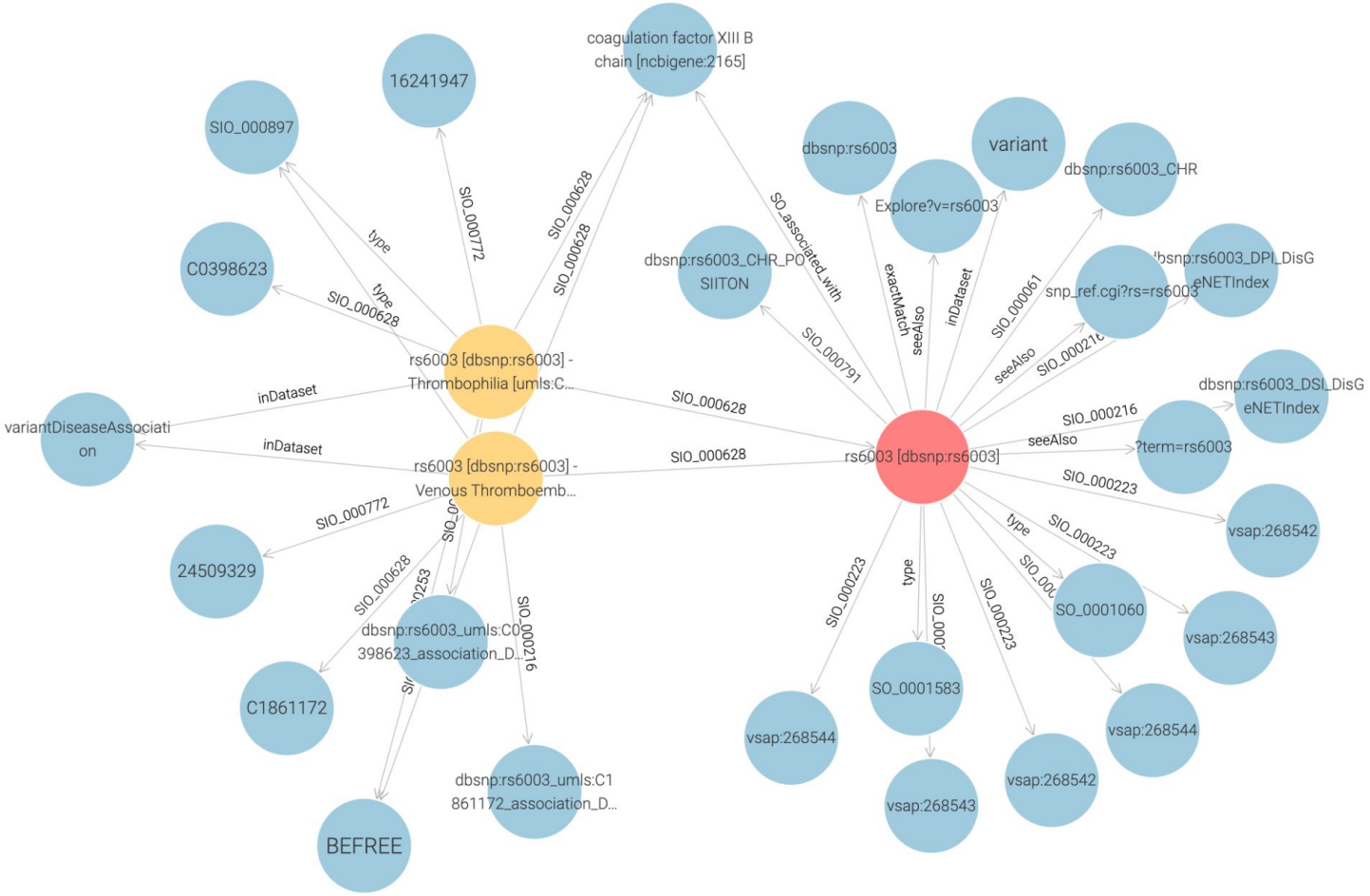
Progress:

- Started pulling the data to create the knowledge graph
- Queried GPTs
- Scored the GPTs against pulled data

Future steps:

Build out the validation model

Knowledge graph



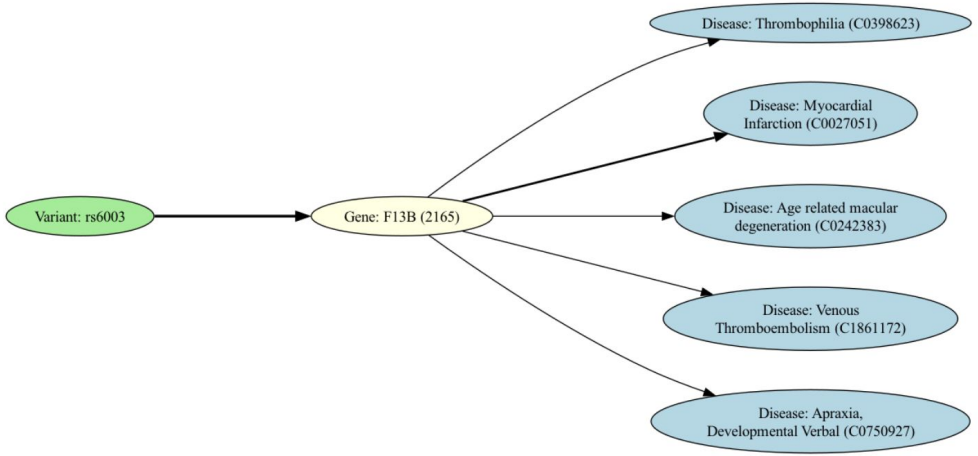
App

DisGeNET Variant Query and AI Analysis

Enter a SNPid (e.g., rs6003):

rs6003

Variant-Gene-Disease Graph



Graph for rs6003

App

LMM Analysis

Without context: rs6003 is a single nucleotide polymorphism (SNP) located in the F5 gene, which encodes the coagulation factor V protein. This SNP is also known as Factor V Leiden mutation.

The Factor V Leiden mutation is associated with an increased risk of developing abnormal blood clots in veins, a condition called venous thromboembolism. This includes deep vein thrombosis (DVT) and pulmonary embolism (PE). Individuals who carry this mutation have a higher risk of developing these conditions, especially if they have additional risk factors such as obesity, immobility, or use of oral contraceptives.

It is important to note that not all individuals with the rs6003 SNP will develop venous thromboembolism, but they have an increased risk compared to those without the mutation.

App

With context: The SNPid rs6003 is related to the F13B gene, which encodes the coagulation factor XIII B chain. This gene has been studied in relation to various diseases and conditions, including thrombophilia, myocardial infarction, and venous thromboembolism.

From the provided citations:

1. Thrombophilia: The study with PMID 16241947 (2005) investigated the His95Arg variant in the Leiden Thrombophilia Study and found a pooled odds ratio (OR) of 1.5 (CI95 1.1-2.0), suggesting a potential association between this variant and thrombophilia.
2. Myocardial Infarction: The study with PMID 17515963 (2007) examined the effects of factor XIIIA-V34L and factor XIIIB-H95R gene variants on survival in myocardial infarction patients. However, the study with PMID 25569091 (2015) found that the p.His95Arg polymorphism did not influence the risk of coronary artery stenosis (CAS) or myocardial infarction (MI).
3. Venous Thromboembolism: The study with PMID 24509329 (2014) investigated the factor XIIIA-V34L and factor XIIIB-H95R variants in venous thromboembolism in central Iran, finding that the factor XIIIA-V34L variant had a protective effect, while the factor XIIIB-H95R variant had a neutral effect.

In summary, the SNPid rs6003 (F13B gene) has been studied in relation to thrombophilia, myocardial infarction, and venous thromboembolism, with varying results.

App

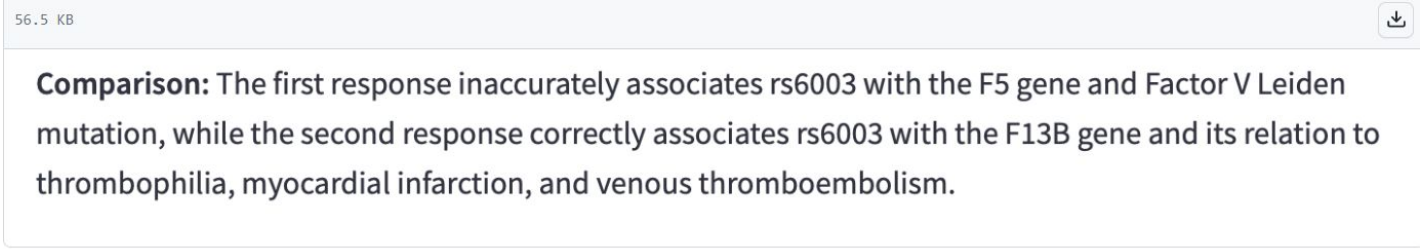
With context: The SNPid rs6003 is related to the F13B gene, which encodes the coagulation factor XIII B chain. This gene has been studied in relation to various diseases and conditions, including thrombophilia, myocardial infarction, and venous thromboembolism.

From the provided citations:

1. Thrombophilia: The study with PMID 16241947 (2005) investigated the His95Arg variant in the Leiden Thrombophilia Study and found a pooled odds ratio (OR) of 1.5 (CI95 1.1-2.0), suggesting a potential association between this variant and thrombophilia.
2. Myocardial Infarction: The study with PMID 17515963 (2007) examined the effects of factor XIIIA-V34L and factor XIIIB-H95R gene variants on survival in myocardial infarction patients. However, the study with PMID 25569091 (2015) found that the p.His95Arg polymorphism did not influence the risk of coronary artery stenosis (CAS) or myocardial infarction (MI).
3. Venous Thromboembolism: The study with PMID 24509329 (2014) investigated the factor XIIIA-V34L and factor XIIIB-H95R variants in venous thromboembolism in central Iran, finding that the factor XIIIA-V34L variant had a protective effect, while the factor XIIIB-H95R variant had a neutral effect.

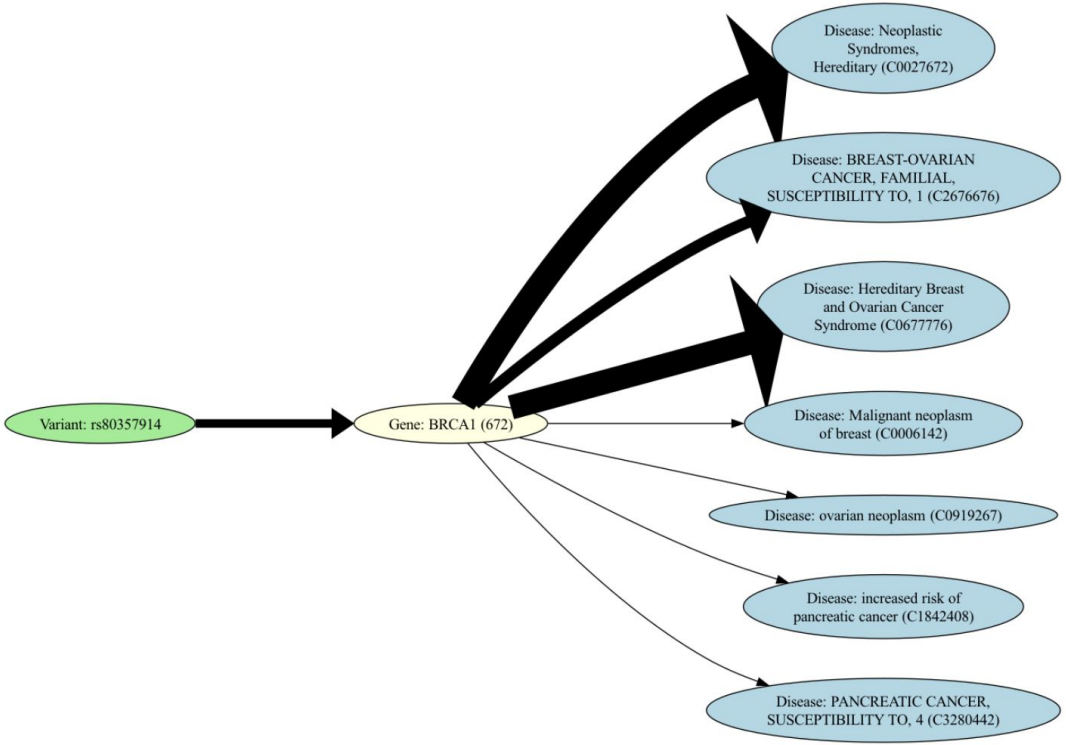
In summary, the SNPid rs6003 (F13B gene) has been studied in relation to thrombophilia, myocardial infarction, and venous thromboembolism, with varying results.

App


56.5 KB
Comparison: The first response inaccurately associates rs6003 with the F5 gene and Factor V Leiden mutation, while the second response correctly associates rs6003 with the F13B gene and its relation to thrombophilia, myocardial infarction, and venous thromboembolism.

App

Variant-Gene-Disease Graph



Graph for rs80357914

Knowledge graph

Knowledge graph disease labels

- Observed ~40 distinct disease labels
- Labels include: Venous thrombosis, inflammation of the large intestine, coagulation factor XIII, hematoma, prolonged bleeding after surgery

Validation scores

- For any given model, taking the number of dissimilar findings divided by similar findings. This allows for a coefficient that can provide a weighted edge in a traditional graph or a weighted class in RDF.

ChatGPT

WA

which diseases is rs6003 associated with?

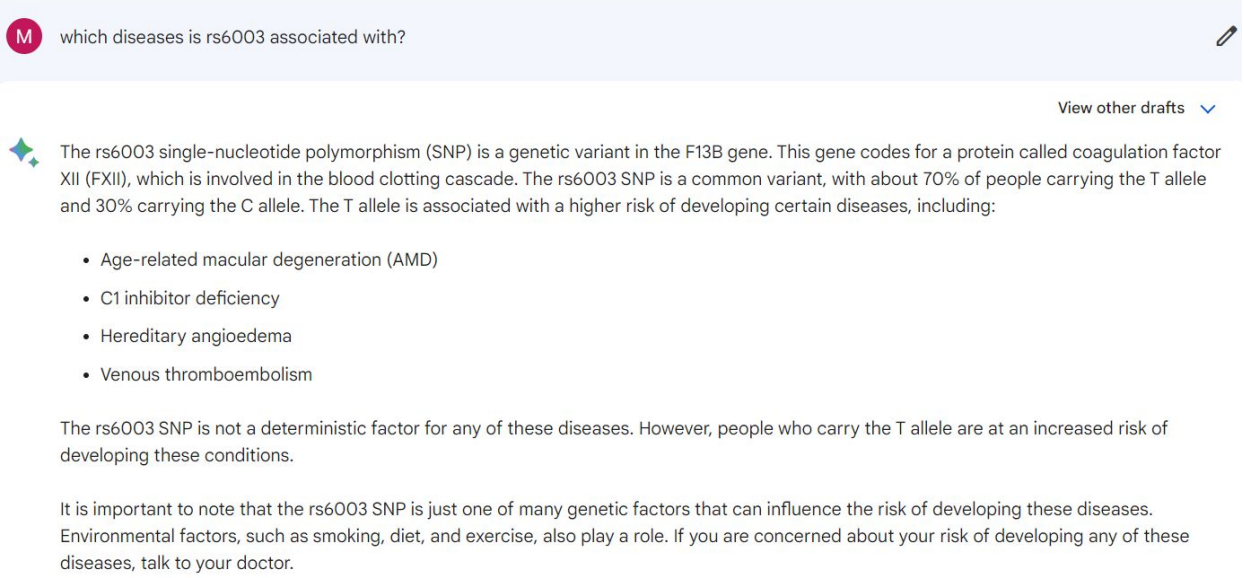
I'm sorry, but I couldn't find any specific information about a disease or condition directly associated with the term "rs6003." It's possible that the term you provided refers to a genetic variant or marker, rather than a disease itself. Genetic variants are often identified using a naming convention that includes the prefix "rs" followed by a number, such as "rs6003." These variants are typically associated with specific genes or regions of the genome.

If you have additional information or context regarding rs6003, please provide more details, and I'll do my best to assist you further.

Overall score: 0

- Common terms: 0
- Uncommon terms: 0

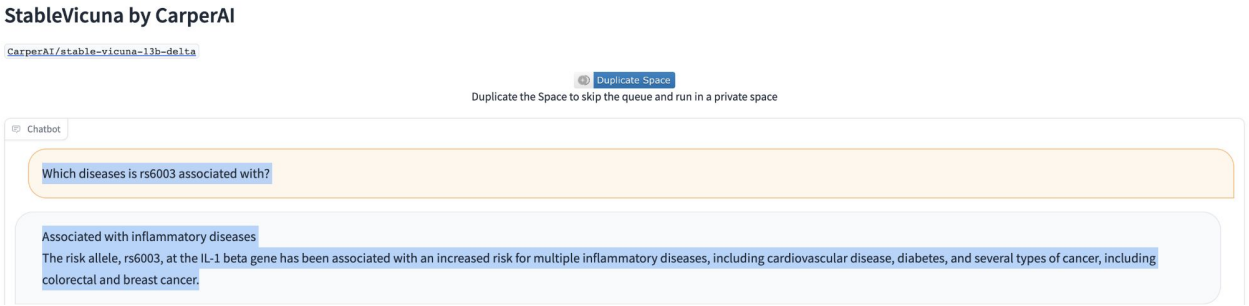
Google Bard



Overall score: 1

- Common terms: 2
  - Coagulation factor FXII
  - Venous thromboembolism
  - 2/40 disease labels
- Uncommon terms: 2
  - Coagulation factor FXII
  - C1 Inhibitor deficiency

StableLM (Hugging face)



Overall score: 1

- Common terms: 2
  - Inflammatory disease
  - Cardiovascular disease
  - **types of cancer
  - 2/40 disease labels
- Uncommon terms: 2
  - Colorectal cancer
  - Breast cancer

Varmovinator risks

Risk 1

- We are designing the validator to attempt to minimize hallucinations. It is not impossible that in some cases we may increase hallucinations by trying to create a balance between novel findings and known ones

Risk 2

- By loading knowledge graphs into the models and saturating local information density, we may end up building an accessible, easy to use NLP interface that does not provide novel hypotheses.

Risk 3 (today)

- Conducted manual scoring and did not use a model. A model could potentially provide better scores

Next steps

Assignment 1

Automate the validation system such that weights are calculated by machine

Assignment 2

Build an NER system such that terms are automatically extracted from model output

Assignment 3

Create a deployable baseline graph to be able to quickly rate new models

Goals for next meeting (CMU October 19-21) biohackathons.github.io

1. Flexible infrastructure for layering assertions with rollback (Apache Kafka perhaps)
2. Flexible tokenization of complex data types
3. Assertion tracking in large models
4. Column headers for data harmonization
5. Validation for disease and drug subtyping with large models