

A

WORKFLOW

Query

Tokenize query, remove stopwords and sort lexicographically

Remove variants and gene names and add to query

Search for matches to phrases for all ordered subsets of tokens

Find best matching phrase
for remaining keywords
using word2vec

EXAMPLE

"endurance capacity for marathon col5a1 rs12722"

capacity	col5a1	endurance	marathon	rs12722
----------	--------	-----------	----------	---------

capacity endurance marathon

- $\{ \dots,$
 - "capacity",
 - "capacity endurance",
 - "endurance",
 - "endurance exercise",
 - "endurance training",
 - $\dots \}$

The diagram illustrates cosine similarity. On the left, a set of phrases is enclosed in curly braces: {"marathon runners", "Marathon race"}. A horizontal line with arrows at both ends connects this set to the word "marathon" on the right. Below the line, the text "cosine similarity" is written.

OUTPUT GENE LIST

col5a1

rs12722

Casp8,Mmp3,Tnc,
Col6a1,Ndufb2,Col12a1,
Col1a1,Col3a1

capacity endurance

Il6,Ins2,Ppara,Cs,
Ppargc1a,Cd59b

marathon

Crp, Il10, Tnf

B

Query

"factor xa"

**“Irx4, Myl2, Xdh, Dlk1, Hyal2, Tmem190,
Cpne4, Cyp1a1 and Irx5”**

**“endurance capacity for marathon
Col5a1 rs12722”**

Cox8a, F3, Serpinc1, F2, F10

Myl2, Hyal2

Col1a1,Col3a1

A word cloud visualization of terms related to thrombosis. The most prominent word is "thrombosis" at the bottom center. Other large words include "thrombus", "thrombotic", "thrombin", "factor viii", "partial thromboplastin time", "thrombin inhibitor", "diseases throm", "platelet activation", "venous thrombosis", "activated protein c", "pathway inhibitor", "appendage", "thrombin time", "clotting time", "splanchnic throm", "disorders thrombotic", "vitamin k subclavian thrombosis vein", "renal thrombosis vein", "vein thrombosis", and "occlusions thrombotic". Smaller words include "serine proteases", "arterial thrombosis", "factor viii", "thrombin receptor", "intravascular coagulation", "rs180133 molecular weight", "plasminogen activator", "complications thrombotic", "promotes vein", "procoagulant activity", "w molecular", "v tumor", "activated partial", "cardiac thrombus", "portal thrombosis vein", "al thrombus", "brand factor blood coagulation", "pts", "artery hepatic th", "activated protein c", "clotting time", "disorders thrombotic", "vitamin k subclavian thrombosis vein", "renal thrombosis vein", "vein thrombosis", and "occlusions thrombotic".

[illegible]

rs35796750 b derived factor growth platelet
rs587779451 congenital imperfecta osteogenesis
imperfecta tarda dentinogenesis imperfecta
rs587776916 runt-related transcription
col1a1-pdgfr fusion rs2075555
imperfecta osteogenesis tarda
i imperfecta osteogenesis
b factor growth platelet-derived
rs2586488
rs127222
s587779707
s587779585
rs387907358
bone brittle disease
rs2228480 pro alpha
rs2412298
blue sclerae
brittle fracture
rs1800472
triple helix
dissecting aortic aneurysm
fragilitas osseum
rs11079346
active tissue disorder
bone mass
rs1800215
rs679620
hepatic fibrosis
ehlers-danlos syndrome
rs12658163 type i collagen
type i procollagen
type iii procollagen
coll1a2 gene
rs1800255
runt-related transcription factor
coll1a2 gene

lethal osteogenesis
femoral neck
first intron
hepatic stellate
brittleness
type iii
pdgfr
type i collagen
blue sclerae
osteogenesis
coll1a1 gene
procollagen
type i procollagen
type iii procollagen
imperfecta osteogenesis
imperfecta osteogenesis recessive

Wordcloud

Highlighted texts

Thrombus Thrombosis

ventricular myosin
rs104894368, rs104894369

Osteogenesis imperfecta (oi)
Osteogenesis