

# The DISEASES database: improved scoring of disease–gene associations mined from the literature

Alexander Junge and Lars Juhl Jensen

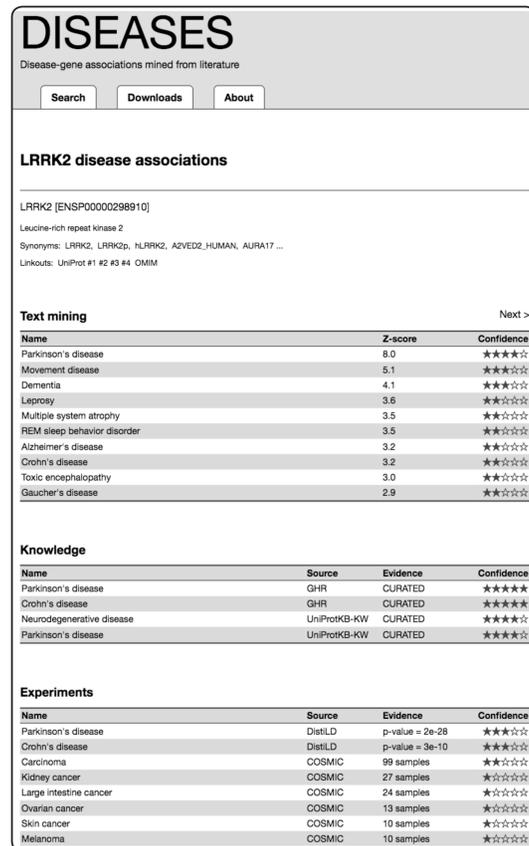
Disease Systems Biology Program, Novo Nordisk Foundation Center for Protein Research, University of Copenhagen, Denmark

Contact: alexander.junge@cpr.ku.dk

## Relevance

Text mining is an established approach to identify associations between biomedical entities. The DISEASES database integrates disease–gene associations found by text mining with expert-curated evidence, cancer mutation data, and genome-wide association studies. Our improved context-aware scoring scheme CoCoScore [3] further improves the text mining pipeline.

DISEASES URL: <https://diseases.jensenlab.org/>



**DISEASES**  
Disease-gene associations mined from literature

Search Downloads About

**LRRK2 disease associations**

LRRK2 [ENSP0000298910]  
Leucine-rich repeat kinase 2  
Synonyms: LRRK2, LRRK2p, HLRRK2, AZVED2\_HUMAN, AURA17, ...  
Links: UniProt #1 #2 #3 #4 OMIM

**Text mining**

Name	Z-score	Confidence
Parkinson's disease	8.0	★★★★☆
Movement disease	5.1	★★★★☆
Dementia	4.1	★★★★☆
Leprosy	3.6	★★★★☆
Multiple system atrophy	3.5	★★★★☆
REM sleep behavior disorder	3.5	★★★★☆
Alzheimer's disease	3.2	★★★★☆
Crohn's disease	3.2	★★★★☆
Toxic encephalopathy	3.0	★★★★☆
Gaucher's disease	2.9	★★★★☆

**Knowledge**

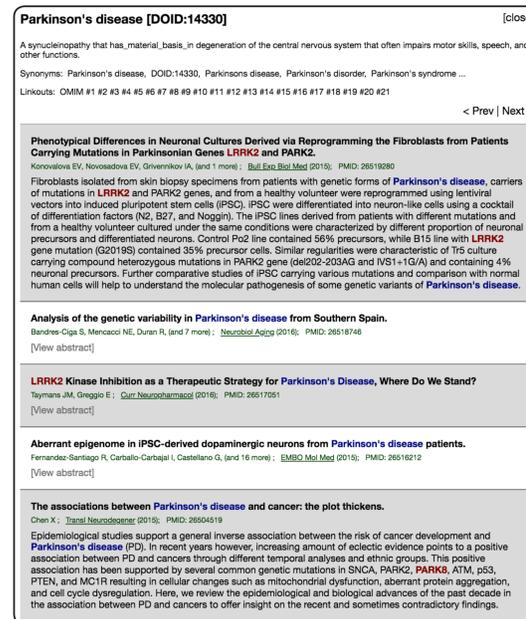
Name	Source	Evidence	Confidence
Parkinson's disease	GHR	CURATED	★★★★☆
Crohn's disease	GHR	CURATED	★★★★☆
Neurodegenerative disease	UniProtKB-KW	CURATED	★★★★☆
Parkinson's disease	UniProtKB-KW	CURATED	★★★★☆

**Experiments**

Name	Source	Evidence	Confidence
Parkinson's disease	DiatLD	p-value = 2e-28	★★★★☆
Crohn's disease	DiatLD	p-value = 3e-10	★★★★☆
Carcinoma	COSMIC	99 samples	★★★★☆
Kidney cancer	COSMIC	27 samples	★★★★☆
Large intestine cancer	COSMIC	24 samples	★★★★☆
Ovarian cancer	COSMIC	13 samples	★★★★☆
Skin cancer	COSMIC	10 samples	★★★★☆
Melanoma	COSMIC	10 samples	★★★★☆

The DISEASES search interface allows to query genes or diseases.

Detailed evidence viewer for associations found by text mining.



**Parkinson's disease [DOID:14330]**

A synucleinopathy that has ,material\_basis, in degeneration of the central nervous system that often impairs motor skills, speech, and other functions.

Synonyms: Parkinson's disease, DOID:14330, Parkinsons disease, Parkinson's disorder, Parkinson's syndrome, ...  
Linkouts: OMIM #1 #2 #3 #4 #5 #6 #7 #8 #9 #10 #11 #12 #13 #14 #15 #16 #17 #18 #19 #20 #21

**Phenotypic Differences in Neuronal Cultures Derived via Reprogramming the Fibroblasts from Patients Carrying Mutations in Parkinsonian Genes LRRK2 and PARK2.**

Kononova EV, Novosadova EV, Givernikov IA, (and 1 more) : *Bull Exp Biol Med* (2015); PMID: 26519280

Fibroblasts isolated from skin biopsy specimens from patients with genetic forms of Parkinson's disease, carriers of mutations in LRRK2 and PARK2 genes, and from a healthy volunteer were reprogrammed using lentiviral vectors into induced pluripotent stem cells (iPSC). iPSC were differentiated into neuron-like cells using a cocktail of differentiation factors (N2, B27, and Noggin). The iPSC lines derived from patients with different mutations and from a healthy volunteer cultured under the same conditions were characterized by different proportion of neuronal precursors and differentiated neurons. Control Po2 line contained 56% precursors, while B15 line with LRRK2 gene mutation (G2019S) contained 35% precursor cells. Similar regularities were characteristic of Tr5 culture carrying compound heterozygous mutations in PARK2 gene (del202-203AG and IVS1+1G/A) and containing 4% neuronal precursors. Further comparative studies of iPSC carrying various mutations and comparison with normal human cells will help to understand the molecular pathogenesis of some genetic variants of Parkinson's disease.

**Analysis of the genetic variability in Parkinson's disease from Southern Spain.**

Bardes-Ciga S, Mencacci NE, Duran R, (and 7 more) : *Neurobiol Aging* (2016); PMID: 26518746

[View abstract]

**LRRK2 Kinase Inhibition as a Therapeutic Strategy for Parkinson's Disease, Where Do We Stand?**

Taymans JM, Greggio E : *Curr Neuropharmacol* (2016); PMID: 26517051

[View abstract]

**Aberrant epigenome in iPSC-derived dopaminergic neurons from Parkinson's disease patients.**

Fernandez-Santiago R, Carballo-Carbajal I, Castellano G, (and 16 more) : *EMBO Mol Med* (2015); PMID: 26516212

[View abstract]

**The associations between Parkinson's disease and cancer: the plot thickens.**

Chen X : *Transl Neurodegener* (2015); PMID: 26504519

Epidemiological studies support a general inverse association between the risk of cancer development and Parkinson's disease (PD). In recent years however, increasing amount of eclectic evidence points to a positive association between PD and cancers through different temporal analyses and ethnic groups. This positive association has been supported by several common genetic mutations in SNCA, PARK2, PARK3, ATM, p53, PTEN, and MCT1R resulting in cellular changes such as mitochondrial dysfunction, aberrant protein aggregation, and cell cycle dysregulation. Here, we review the epidemiological and biological advances of the past decade in the association between PD and cancers to offer insight on the recent and sometimes contradictory findings.

All disease–gene associations, the tagger software, and dictionaries used for text mining are available under open licenses [2]. The tissue expression database TISSUES (<https://tissues.jensenlab.org>) uses the same text mining approach.

Coverage by evidence channel.

Evidence	Genes	Diseases	Associations
Text mining	15631	4598	478407
Knowledge	2001	735	15231
Experiments	10711	423	89073

## Improved text mining via CoCoScore

The DISEASES text mining scheme ignores context when scoring disease–gene pairs. CoCoScore [3] uses textual context to score whether an association is described.

The co-occurrence count  $C(G,D)$  of gene  $G$  and disease  $D$  is defined as:

$$C(G, D) = \sum_{k=1}^n \omega_k(G, D) + w_a \delta_{ak}(G, D)$$

Sum over all Medline abstracts

Best score of sentences co-mentioning D and G

Constant score if D and G appear in the same abstract

$$\omega_k(G, D) = \max_{i \in T(k, G, D)} r(i)$$

$T(k,G,D)$  are all sentences in abstract  $k$  that co-mention  $G$  and  $D$ .  $r(i)$  is the sentence-level score returned by a fastText-based model [4] (see Figure).

The co-occurrence score  $S(G,D)$  is:

$$S(G, D) = C(G, D)^\alpha \left( \frac{C(G, D) C(\cdot, \cdot)}{C(G, \cdot) C(\cdot, D)} \right)^{1-\alpha}$$

Observed-over-expected ratio

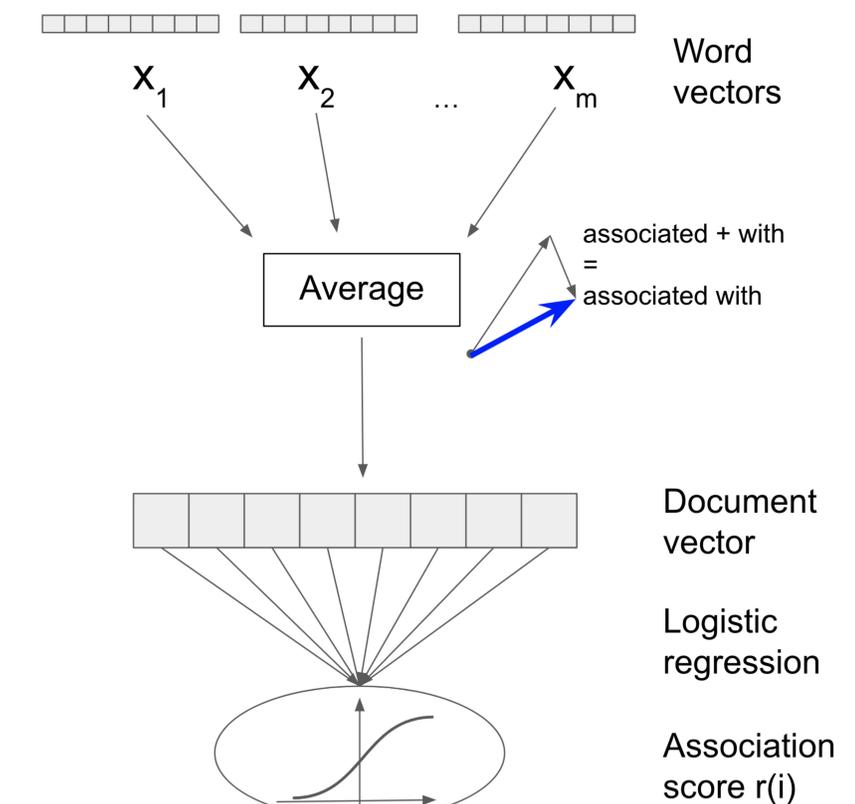
Performance on test dataset derived from curated associations using distant supervision.

	AUROC	AP
CoCoScore	0.97	0.77
DISEASES score	0.96	0.72

AUROC: Area under the ROC curve  
AP: Average Precision

## CoCoScore

LLRK1 is not associated with Parkinson's disease. Sentence  $i$



## Key features

- DISEASES integrates scored disease–gene associations from text mining, curated knowledge, and experiments
- weekly text mining updates and data downloads
- CoCoScore, our novel context-aware scoring scheme, improves the text mining performance

Want to know more?

- [1] Pletscher-Frankild *et al.*, *Methods* (2015)
  - [2] tagger on Bitbucket: <https://goo.gl/hefLqj>
  - [3] CoCoScore on GitHub: <https://goo.gl/xCxdjt>
  - [4] Joulin, Grave *et al.*, *arXiv* (2016)
- Visit <https://diseases.jensenlab.org/>

Funding: Novo Nordisk Foundation, NIH Common Fund