SciBite
an ELSEVIER company

Use case

Using phenotype triangulation to improve disease understanding

SciBite
an ELSEVIER company

Using phenotype triangulation to improve disease understanding

Phenotypic similarity between diseases is an important factor in biomedical research since similar diseases often share similar molecular origins. This forms the basis of an inference-led approach to disease characterisation known as Phenotype Triangulation. Where there is strong overlap in phenotype signatures, we can hypothesize that a disease pair could share an underlying mechanistic relationship. The more phenotypes that two otherwise disconnected diseases have in common, the more likely that there is some sort of relationship between them. For example, in the example illustrated below, diseases 1 and 2 are strongly linked via uniquely shared phenotypes (phenotypes 1, 2 and 3) whereas the relationship between Diseases 2 and 3 is weak because their single shared Phenotype 4 is also shared with a high number of other diseases.

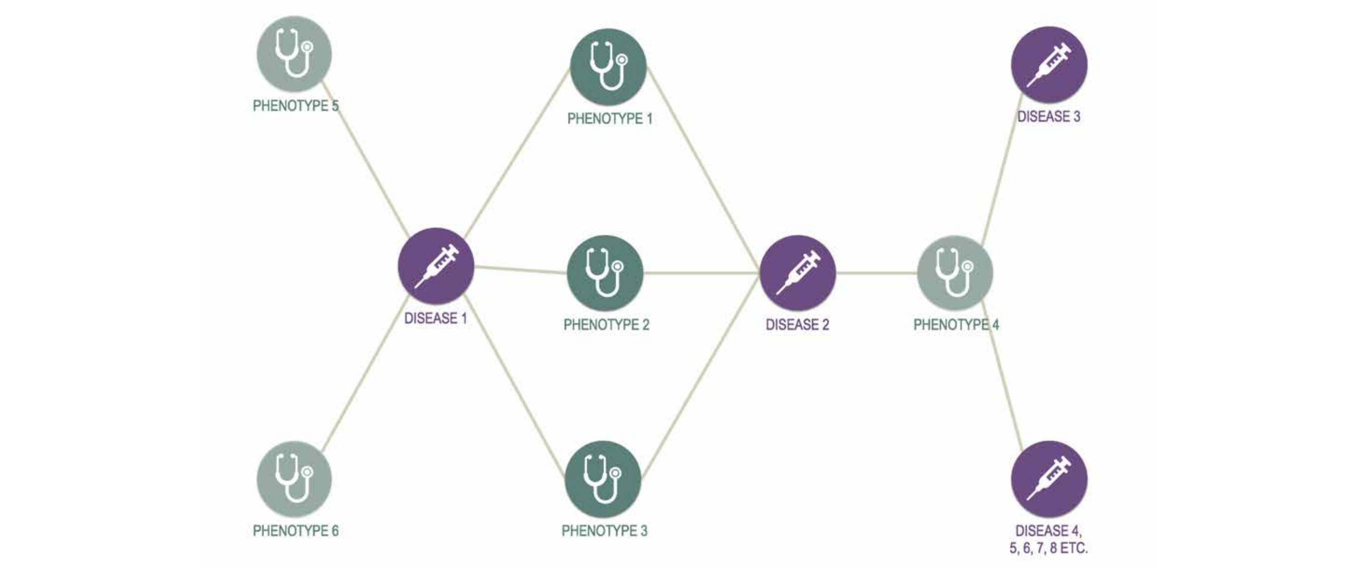


Figure 1: The principle of Phenotype Triangulation

Potentially interesting relationships can be investigated further, for example by exploring known genetic associations.

However, evidence of disease similarity is often hidden within unstructured biomedical literature and often not presented as direct evidence. This necessitates a time consuming and costly review process to stand any chance of identifying relevant linkages.

Here we describe how Phenotype Triangulation can be used to facilitate the characterisation of a disease by revealing and quantifying similarities that are not explicitly stated elsewhere. In particular, we describe the application of Phenotype Triangulation to the characterisation of rare diseases (defined as those that affect less 1:1500 people in the US), for which the amount of research is still at a relatively low volume compared to more common conditions.

SciBite: enabling phenotype triangulation

A plethora of disease-related information can be found in the biomedical literature. However, many of these sources are unstructured and are not designed with searching in mind. SciBite has developed a method for extracting phenotypes by combining the results of Semantic Analytics and Machine Learning to unlock the potential of biomedical literature and quantify disease similarities based on their phenotypic signatures.

As described below, SciBite's approach involves four key steps:

- Mining the scientific literature for potential phenotype similarities
- Ranking potential phenotype similarities
- Creating an integrated knowledgebase
- Visual and computational data exploration function

Mining the scientific literature for potential phenotype similarities

The first stage of Phenotype Triangulation is to identify co-occurring pairs of diseases and phenotype terms from over 24 million MEDLINE abstracts. The approach is an extension of one used by Caniza and colleagues, in which Medical Subject Headings Thesaurus (MeSH) keyword annotations from OMIM were used to quantify similarity between heritable diseases[1].

MeSH annotations are high quality but limited in what they represent since only a fraction of possible phenotypes and other phenomena are recorded. In contrast, SciBite's manually curated, enriched vocabularies, or VOCabs, comprising tens of millions of synonyms, have unrivalled depth and breadth, ensuring comprehensive coverage of relevant terminology and providing the robust foundation necessary for Phenotype Triangulation to be effective.

SciBite's ontology management platform, CENtree, enables organisations to maintain up to date ontologies representing evolving scientific language. Subject matter experts easily contribute to keeping things current and augment VOCabs with their proprietary internal terms vocabularies to create a single, authoritative, universally applied terminology. CENtree also leverages machine learning techniques to suggest potential new terms that are similar to other related words because they are used in a similar context.

We also include rules to handle ambiguity in text, thus ensuring exceptional levels of precision and recall. For example, Figure 2 compares SciBite's INDICATION vocabulary to MeSH for a sample of disease terms.

For example, using the SciBite platform, same sentence co-occurrences of pairs of over 9,000 different indications (enriched with more than 2 million synonyms) and over 11,000 human phenotype concepts (enriched with more than 540,000 synonyms) can be identified automatically.

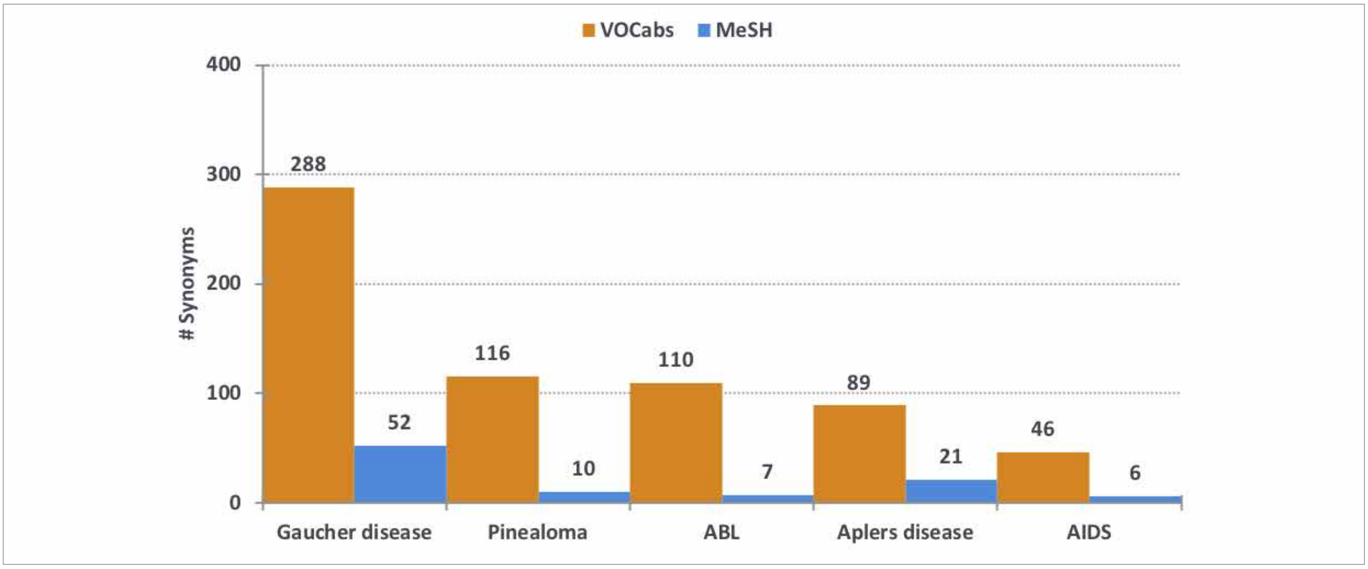
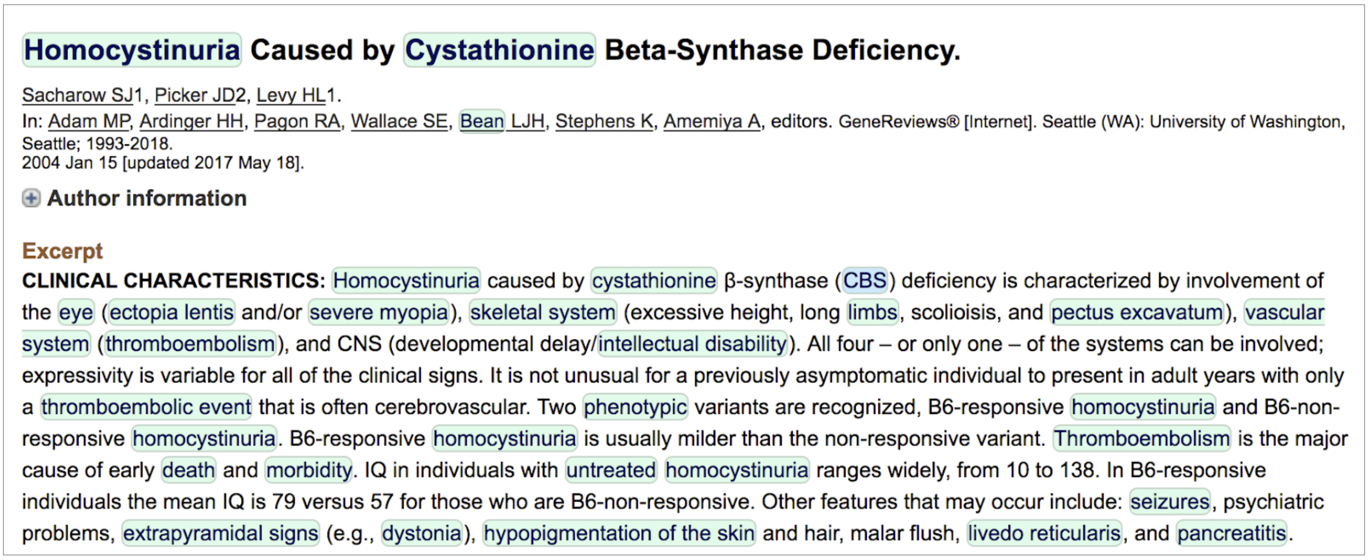


Figure 2: Comparison of SciBite's INDICATION VOCab to the publicly-available Medical Subject Headings Thesaurus for a sample of disease terms

[1] Caniza, H. et al. (2015). A network medicine approach to quantify distance between hereditary disease modules on the interactome. Sci. Rep. 5, 17658; doi: 10.1038/srep17658



Homocystinuria Caused by Cystathionine Beta-Synthase Deficiency.

Sacharow SJ1, Picker JD2, Levy HL1.

In: Adam MP, Ardinger HH, Pagon RA, Wallace SE, Bean LJH, Stephens K, Amemiya A, editors. GeneReviews® [Internet]. Seattle (WA): University of Washington, Seattle; 1993-2018.
2004 Jan 15 [updated 2017 May 18].

⊕ **Author information**

Excerpt

CLINICAL CHARACTERISTICS: Homocystinuria caused by cystathionine β-synthase (CBS) deficiency is characterized by involvement of the eye (ectopia lentis and/or severe myopia), skeletal system (excessive height, long limbs, scolioisis, and pectus excavatum), vascular system (thromboembolism), and CNS (developmental delay/intellectual disability). All four – or only one – of the systems can be involved; expressivity is variable for all of the clinical signs. It is not unusual for a previously asymptomatic individual to present in adult years with only a thromboembolic event that is often cerebrovascular. Two phenotypic variants are recognized, B6-responsive homocystinuria and B6-non-responsive homocystinuria. B6-responsive homocystinuria is usually milder than the non-responsive variant. Thromboembolism is the major cause of early death and morbidity. IQ in individuals with untreated homocystinuria ranges widely, from 10 to 138. In B6-responsive individuals the mean IQ is 79 versus 57 for those who are B6-non-responsive. Other features that may occur include: seizures, psychiatric problems, extrapyramidal signs (e.g., dystonia), hypopigmentation of the skin and hair, malar flush, livedo reticularis, and pancreatitis.

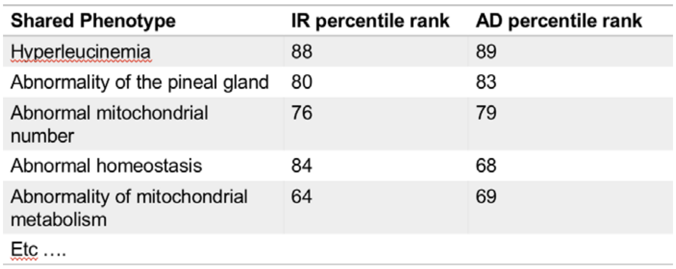
Figure 3: Annotation of disease and phenotype terms, and other concepts of biological importance, in a MEDLINE abstract

Ranking phenotype similarities

Sentence co-occurrence on its own is not enough to determine whether pairs of indications and phenotypes are truly related. Very common phenotypes are observed across multiple diseases, so it is necessary to filter them out from those that are specific to particular disease states. For example, if two diseases share the highly connected phenotype 'nausea', this is scientifically less interesting than if both share a less frequently occurring phenotype.

SciBite has developed Machine Learning algorithms to calculate the strength of each relationship by comparing the same sentence co-occurrence of the terms in each indicationphenotype pair with the frequency of each term appearing individually. The resulting score is then compared against all disease-phenotype co-occurrence scores for an individual disease and given a percentile rank value. This highlights scientifically interesting associations amongst the background noise.

Where there is strong overlap in shared phenotypes, or 'phenotype signatures', it can be hypothesised that a disease pair could share an underlying mechanistic relationship. For example, the table below shows the comparison of phenotype profiles for Insulin Resistance (IR) and Alzheimer's Disease (AD), also sometimes referred to as Type 3 Diabetes[1]. Based on the shared 'phenotype signatures', SciBite's Machine Learning algorithm recognises an association between these two diseases.



Shared Phenotype	IR percentile rank	AD percentile rank
Hyperleucinemia	88	89
Abnormality of the pineal gland	80	83
Abnormal mitochondrial number	76	79
Abnormal homeostasis	84	68
Abnormality of mitochondrial metabolism	64	69
Etc ….		

Table 1: Shared phenotypes with percentile rank scores for Insulin Resistance (IR) vs. Alzheimer's Disease (AD)

This finding is backed up in the literature – resistance to insulin and insulin-like growth factor have been identified as key aspects of Alzheimer's Disease progression[2]. The validation demonstrates how, with no prior knowledge of Insulin Resistance or Alzheimer's Disease, our algorithms can effectively extract themes from the scientific literature without any human intervention.

[1] For example, De la Monte, S. M., & Wands, J. R. (2008). Alzheimer's Disease Is Type 3 Diabetes–Evidence Reviewed. **Journal of Diabetes Science and Technology** (Online), 2(6), 1101–1113.

[2] De la Monte SM, Wands JR. Alzheimer's Disease Is Type 3 Diabetes–Evidence Reviewed. Journal of diabetes science and technology (Online). 2008;2(6):1101-1113.

Creating and exploring an integrated knowledgebase

Following the scoring process, the resulting information was loaded into a Neo4j database to generate a knowledge graph representing the strength of connectivity between diseases based on their phenotype signatures. Use of a graph database enables the data to be queried computationally for downstream analysis. The data can also be explored visually by connecting network visualisation tools, such as Linkurious.

The knowledge graph is enriched by integrating additional data sources, such as gene-disease association information from Open Targets and protein-protein interaction data from iRefIndex. It is also possible to incorporate proprietary sources, such as semantically enriched bioassay and gene expression data[1], which include information such as tissue-specific expression patterns.

This enables subsequent analysis to go beyond phenotypes and classify diseases based on richer signatures comprised of genomic, proteomic and phenotypic information and increase confidence in the relevance of shared phenotypic signatures. It also opens up the possibility to making connections and inferences that might otherwise have remained hidden. For example, once we know that a disease-phenotype relationship mined from MEDLINE refers to the same disease as in a gene-disease association from Open Targets, we can start to build up a picture of the interplay between diseases, phenotypes and genes. Multiplying this across the whole of MEDLINE with cross-references to a range of structured data sources results in a highly interconnected snapshot of current scientific knowledge.

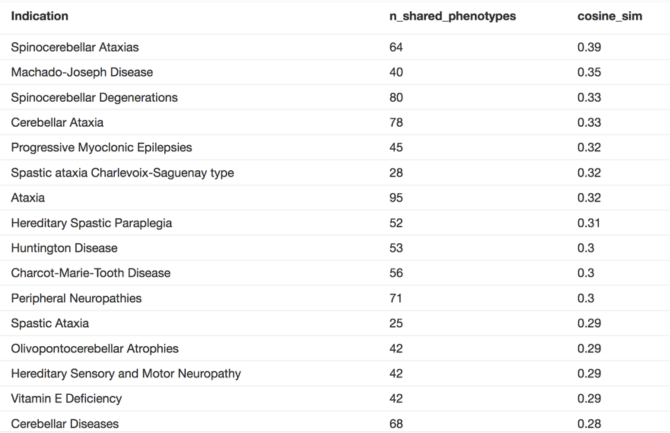
Phenotype triangulation in action

Improved characterisation of rare diseases

The power of the SciBite platform can be illustrated using a real-life rare disease and its related conditions. For example, Friedreich's Ataxia is described as:

> *"...a genetic, progressive, neurodegenerative movement disorder, with a mean age of onset between 10 and 15 years. Initial symptoms may include unsteady posture, frequent falling, and progressive difficulty walking due to impaired ability to coordinate voluntary movements (ataxia)."*

Our aim is to interrogate the knowledgebase described in the previous section and better characterise this rare disease based on its similarities to more widely understood conditions. An initial search reveals that quite a large number and range of phenotypes are known to be associated with Friedreich's Ataxia. As illustrated in the table below, by filtering the network by phenotype relationship score, it is possible to identify conditions that are indirectly linked to Friedreich's Ataxia based on similar phenotype signatures described in the scientific literature.



Indication	n_shared_phenotypes	cosine_sim
Spinocerebellar Ataxias	64	0.39
Machado-Joseph Disease	40	0.35
Spinocerebellar Degenerations	80	0.33
Cerebellar Ataxia	78	0.33
Progressive Myoclonic Epilepsies	45	0.32
Spastic ataxia Charlevoix-Saguenay type	28	0.32
Ataxia	95	0.32
Hereditary Spastic Paraplegia	52	0.31
Huntington Disease	53	0.3
Charcot-Marie-Tooth Disease	56	0.3
Peripheral Neuropathies	71	0.3
Spastic Ataxia	25	0.29
Olivopontocerebellar Atrophies	42	0.29
Hereditary Sensory and Motor Neuropathy	42	0.29
Vitamin E Deficiency	42	0.29
Cerebellar Diseases	68	0.28

Table 1: *Indications with Similar Text-Mined Phenotype Signatures Compared to Friedreich's Ataxia. The table includes the Top 1% of Phenotype Similarity Scores, Sorted by the relationship strength metric (cosine similarity)*

Many of these indications may be well known to experts in the field, but less well-known indications may also be revealed. It is also possible to explore interesting connections in more depth. For example, the image below illustrates the phenotypes shared between Friedreich's Ataxia and Huntington's Disease.

[1] See SciBite's publication 'More than FAIR: Unlocking the Value of Your Bioassay Data', for details

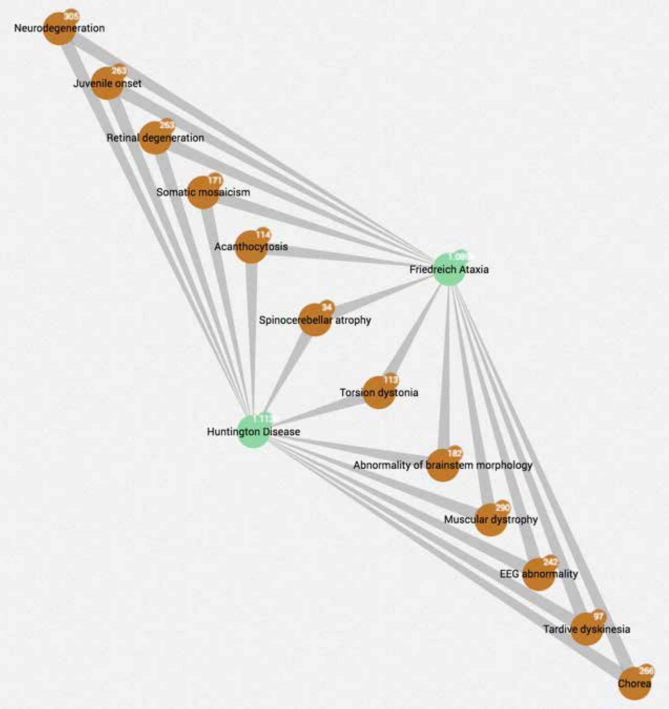


Figure 4: *Major shared phenotypes between Friedreich's Ataxia and Huntington's Disease. Phenotypes ranked using statistical analysis of text mined results*

Now that we have a list of similar conditions, it's helpful to filter it down to a more manageable set. One method is to overlay gene association data from a structured source, such as Open Targets or DisGenNET, to reveal phenotypically similar conditions that are also known to have similarities in their underlying mechanisms. It also helps identify where there are knowledge gaps and highlights potential areas where these gaps might be filled.

As illustrated below, several of the conditions that are phenotypically similar to Friedreich's Ataxia have many known gene associations, partly reflecting the amount of research conducted into different disease areas. In the case of Friedreich's Ataxia, the close-up view shown below illustrates that it is quite isolated from phenotypically similar conditions and that there are clearly gaps in mechanistic understanding. In fact, there is only one direct gene association gene shown – frataxin (FXN).

By incorporating protein-protein interaction data from iRefIndex, some of the gaps are filled – FXN interacts with several genes that are known to be associated with phenotypically related conditions.

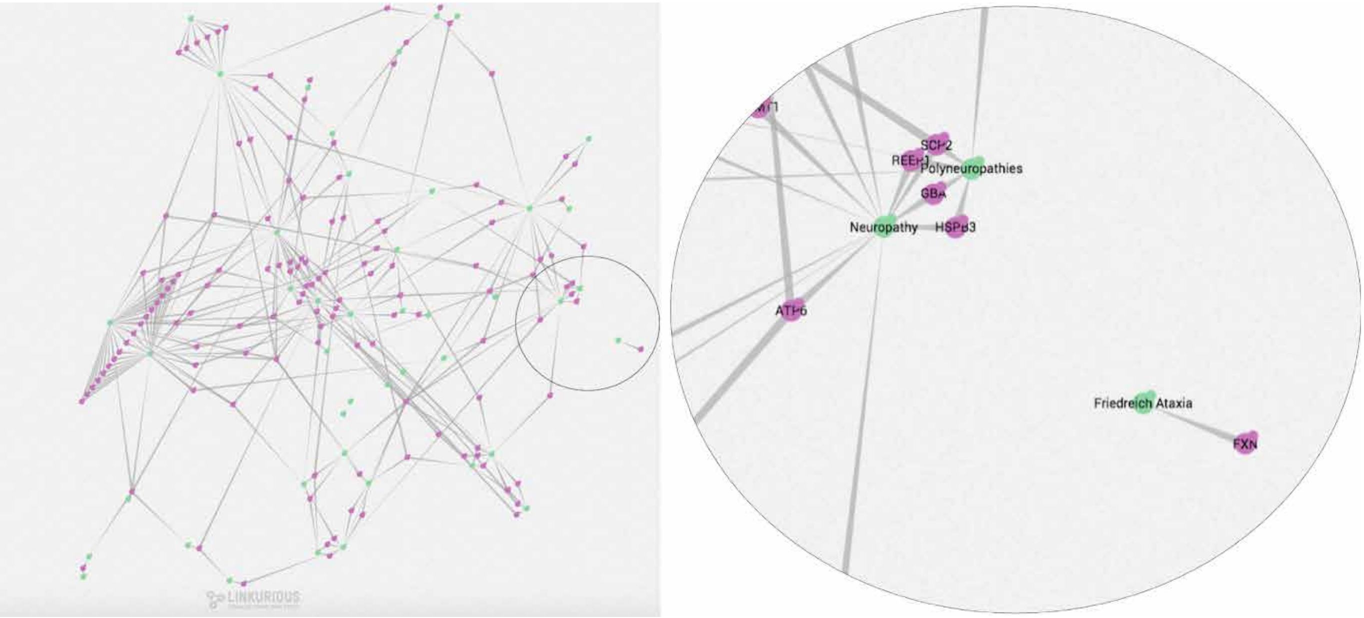


Figure 5: *Gene-Disease Associations for conditions phenotypically similar to Friedreich's Ataxia. The green nodes represent indications and the pink nodes represent associated genes. The close-up view shows Friedreich's Ataxia and FXN in isolation*

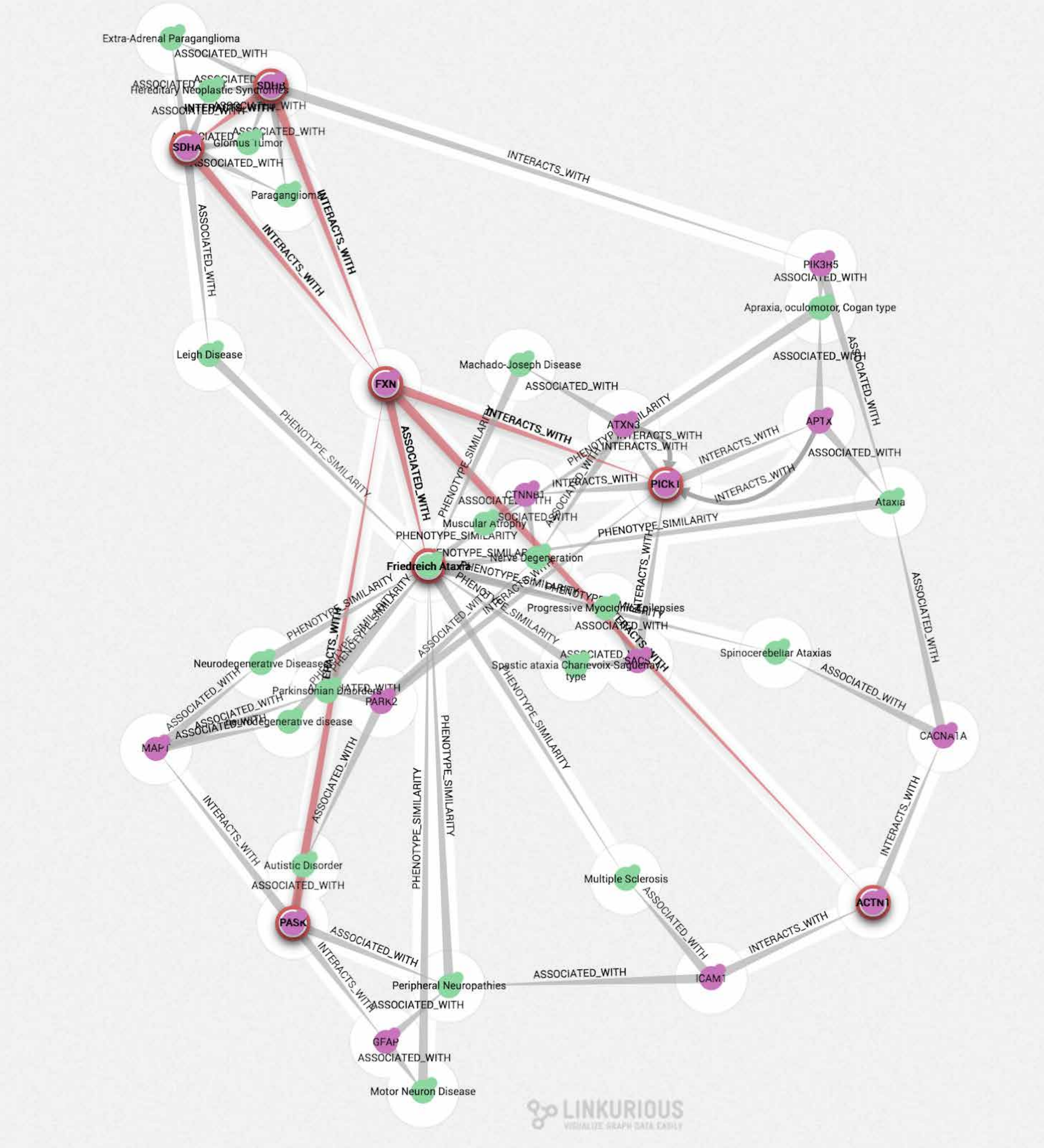


Figure 6: *Integrating genomic, proteomic and phenotypic information*

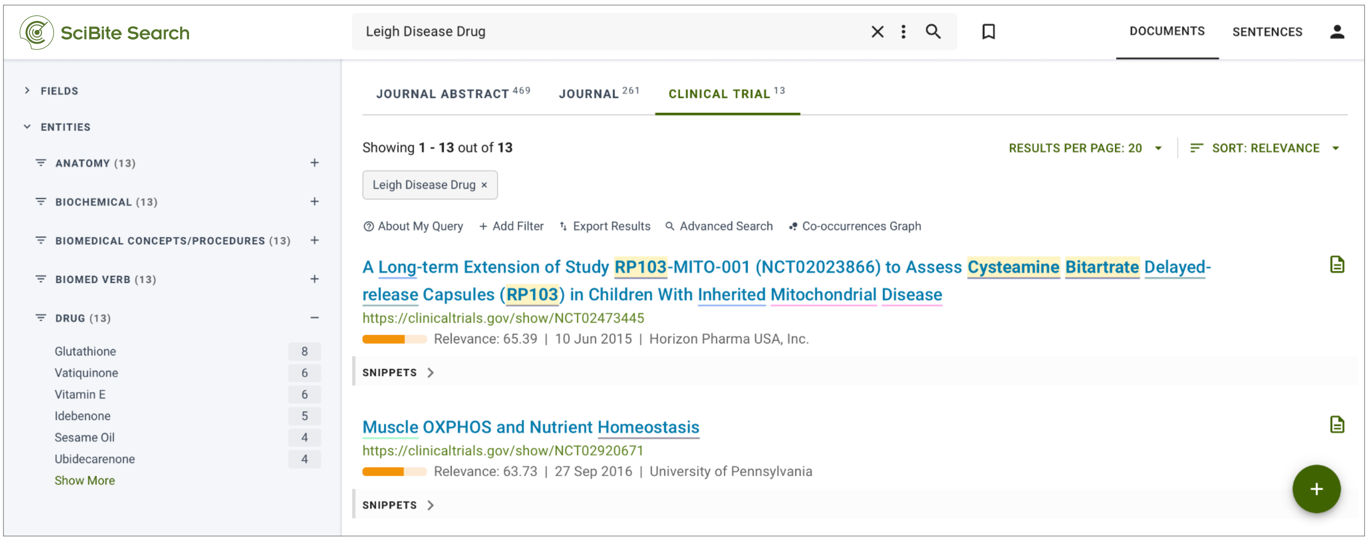
For example, the gene entity PASK (PAS domain containing serine/threonine kinase; shown as a pink node towards the top of the image above), interacts with FXN and is also known to be associated with Peripheral Neuropathies, which was identified earlier as one of the most phenotypically similar conditions to Friedreich's Ataxia. Similarly, SDHA (succinate dehydrogenase complex, subunit A) is linked to a number of phenotypically-related conditions.

With a richer network, it is possible to identify connections that were absent or hidden when relying on phenotypic similarity alone. By incorporating additional sources of data it is possible to hone in on conditions that share both a phenotype signature and have a plausible mechanistic link. For example, Leigh Disease does not have the strongest evidence of phenotypic similarity to Friedreich's Ataxia, but the addition of mechanistic information provides increased confidence in the possible relationship between these two diseases.

Through this analysis, SciBite has quickly revealed potential new areas for research. Importantly, all of these connections were identified by a computer with no prior knowledge of the condition. The same computer can also characterise thousands of other conditions in the same way, without any human intervention.

Informing drug repositioning initiatives

Beyond better characterising rare diseases, semantic analysis can also reveal opportunities for drug repositioning. For example, it is possible to explore the relationship between Friedreich's Ataxia and Leigh Disease in more depth. Using the SciBite platform, a semantic search for Leigh Disease reveals that one of the most frequent drugs mentioned in a clinical trial context is Vatiquinone.



SciBite Search

Leigh Disease Drug

Glutathione	8
Vatiquinone	6
Vitamin E	6
Idebenone	5
Sesame Oil	4
Ubidecarenone	4

Show More

Showing 1 - 13 out of 13

A Long-term Extension of Study RP103-MITO-001 (NCT02023866) to Assess Cysteamine Bitartrate Delayed-release Capsules (RP103) in Children With Inherited Mitochondrial Disease
https://clinicaltrials.gov/show/NCT02473445
Relevance: 65.39 | 10 Jun 2015 | Horizon Pharma USA, Inc.

Muscle OXPHOS and Nutrient Homeostasis
https://clinicaltrials.gov/show/NCT02920671
Relevance: 63.73 | 27 Sep 2016 | University of Pennsylvania

Figure 7: Using the SciBite Search interface to perform a semantic search of data from clinicaltrials.gov involving "Leigh Disease" and any DRUG

A further semantic search of clinicaltrials.gov for 'Vatiquinone' and 'Orphan Disease' reveals a connection with Friedreich's Ataxia. Importantly, without semantic search capabilities, this would not have been found as the references are to a synonym of Vatiquinone (EPI-743) rather than Vatiquinone itself. While this analysis has not identified a brand-new repurposing candidate, it validates the power of this method of hypothesis generation to guide more in-depth research.

SciBite
an ELSEVIER company

Summary

SciBite delivers high-quality, automated text annotation and facilitates data integration to support expert-led research activities. SciBite has developed a method to successfully predict disease relationships without any prior knowledge of the diseases and to infer relationships without any direct evidence, based on the strength of indirect evidence.

By incorporating known genetic associations and protein-protein interactions, SciBite enables end users to seamlessly navigate through previously disconnected datasets and delivers valuable insights into underlying disease mechanisms.

SciBite can drive forward research towards new therapies and treatments by:

- Facilitating deeper understanding of any disease
- Generating hypotheses around candidate drug targets
- Enabling deeper research at a faster rate through making connections between diseases at a mechanistic level
- Discovering relationships between diseases in light of potentially sparse evidence
- Identifying opportunities for drug repositioning
- Aiding patient diagnosis

SciBite quickly opens up new possibilities by presenting data to users in an accessible, easy to understand way, facilitating scientific exploration and giving scientists more time to focus on research.

About SciBite

SciBite is an award-winning semantic software company offering an ontology-led approach to transforming unstructured content into machine-readable clean data. Supporting the top 20 pharma with use cases across life sciences, SciBite empowers customers with a suite of fast, flexible, deployable API technologies, making it a critical component in scientific data-led strategies. Contact us to find out how we can help you get more from your data.

To learn how SciBite can unlock the value of your data, speak to one of our experts today or email us at contact@scibite.com

SciBite
an ELSEVIER company

SciBite's data-first, semantic analytics software is for those who want to innovate and get more from their data. At SciBite we believe data fuels discovery and we are leading the way with our pioneering infrastructure that combines the latest in machine learning with an ontology-led approach to unlock the value of scientific content. Supporting the world's leading scientific organisations with use-cases from discovery through to development, SciBite's suite of fast, flexible, deployable API technologies empower our customers, making it a critical component in scientific, data-led strategies. Contact us to find out how we can help you get more from your data.

To learn how SciBite can unlock the value of your data, speak to one of our experts today or email us at contact@scibite.com

Head Office:
SciBite Limited
BioData Innovation Centre
Wellcome Genome Campus
Hinxton, Cambridge CB10 1DR
United Kingdom

www.scibite.com
contact@scibite.com
LinkedIn: SciBite
Twitter: @SciBite
+44 (0)1223 786 129