Text Mining At Sanofi: Genotype-Phenotype Associations In A Multiple Sclerosis Biomarker Discovery Project

The human leukocyte antigen (HLA) genotype is an important risk factor for multiple sclerosis (MS). As part of a project to discover potential MS biomarkers, Sanofi decided to annotate the association of HLA alleles and haplotypes with diseases and drug hypersensitivity. There are some public resources that associate HLA alleles with over 40 different autoimmune diseases, some cancers, infectious disease and drug hypersensitivities, but none provides systematic annotation of these associations.

Sanofi established a workflow for whole exome sequencing-based HLA typing and analysis that identified more than 400 HLA alleles. They used the Linguamatics NLP platform to analyze and search the literature to annotate the association of the HLA alleles with diseases and drug hypersensitivity. This project resulted in more than double the previous disease associations, and the curated annotations were fed into a searchable knowledge base for broad use within the Sanofi team.

QUICK FACTS

Situation: Sanofi wanted to annotate the association of HLA alleles and haplotypes with diseases and drug hypersensitivity as part of an MS biomarker discovery project; while HLA alleles have been associated with autoimmune diseases, types of cancer, infectious disease and drug adverse events, there are no known resources that systematically annotate the association of HLA alleles and diseases.

Solution: Sanofi needed a comprehensive catalogue of disease annotations to HLA alleles, and used the Linguamatics NLP platform to text mine the scientific literature (25 million PubMed abstracts and 4 million full-text journal articles). An NLP query, equipped with an internal HLA gene ontology, a dictionary of relationship verbs and a disease ontology, text mined the literature sources to identify HLA alleles and their relationships with diseases and drug sensitivity.

Success: The Linguamatics NLP query successfully identified all the 22 previously published autoimmune diseases associated with HLA alleles, and uncovered an additional 33 unpublished disease and drug sensitivity associations. The curated annotations were fed into a searchable knowledge base for broad use within the Sanofi team in its search for novel biomarkers.

Situation

A key requirement for any drug development project—and increasingly in precision/personalized medicine and pharmacogenomics—is a comprehensive understanding of the genetic associations for the disease of interest. The HLA genotype is responsible for some 30% of the risk of MS, and participates in almost every aspect of the disease. The HLA system is a gene complex that encodes the major histocompatibility complex (MHC) proteins in humans, and these cell-surface proteins regulate the immune system. HLA genes are highly polymorphic, which means that they have many different alleles, allowing them to fine-tune the adaptive immune system.

While public databases of genomic variants can provide valuable information, there may be many gaps in the biological knowledge, and as part of a project to identify novel biomarkers for MS, Sanofi needed a comprehensive catalogue of disease and drug sensitivity annotations to HLA alleles and haplotypes.

HLA alleles have been associated with more than 40 different autoimmune diseases, various types of cancer, infectious disease and drug adverse events. However, there are no known resources that systematically annotate the association of HLA alleles and diseases, and Sanofi wanted to address this shortfall.

Solution

The HLA system is the most highly polymorphic region in the human genome, and some of its alleles are known to be associated with a higher risk of MS and other autoimmune disorders. Researchers from Sanofi collected DNA and RNA samples from MS patients for HLA typing, whole exome sequencing, RNA-Seq and genome-wide association studies. Although more than 400 alleles with the potential to serve as candidate biomarkers were identified in the HLA typing study, these were not annotated in any database.

> *"We use [the Linguamatics platform] for gene disease mapping, target ID, toxicity and gene mutation analyses, biomarker discovery, drug repurposing and patent analysis [...] Our results demonstrate that text mining can be applied to identify accurate associations between HLA alleles, haplotypes, and disease, as well as drug hypersensitivity."*
>
> *— Dongyu Liu, Associate Director, Translational Sciences, Sanofi*

Sanofi opted to use the Linguamatics natural language processing (NLP) solution to text mine the published literature for HLA allele- and haplotype-disease and drug sensitivity associations. The source literature content to be mined was 25 million PubMed abstracts and 4 million full-text journal articles. These were

Figure 1: HLA And Disease Associations: Risk Alleles Discovered By NLP Text Mining

Category	Disease	Risk allele
Autoimmune	Ankylosing spondylitis	HLA-B27, HLA-B*14
	Celiac disease	HLA-DQA1*05:01, HLA-DQB1*02:01, HLA-DQA1*03, HLA-DQB1*03:02
	Crohn disease	HLA-DRB1*01:03, HLA-DRB1*1501, HLA-DRB1*1302, HLA-DQB1*0602
	Graves' disease	HLA-DQA1*05:01, HLA-DRB1*03:01, HLA-DQB1*02:01, HLA-DQA1*03:01, HLA-DQB1*03:02 , HLA-DRB1*07
	Hashimoto's thyroiditis	DRB1*04-DQB1*03-DQA1*03, DQA1*0102,DQB1*0602, HLA-DRB1*11:01, HLA-DRB1*14:04
	sy	DQA1, HLA-DRB1, HLA-C*07
	Multiple sclerosis	HLA-DRB1*15:01, DQA1*01:02, DQB1*06:02
	Myasthenia gravis	HLA-C*07:01, HLA-DRB1*15:01, HLA-DQB1*03:03, HLA-DQA1*04:01
	Pemphigus vulgaris	HLA-DQB1*03:01, HLA-DRB1*14, , HLA-DQB1*05:03, HLA-DRB1*14:54
	Psoriasis	HLA-C*06:02, HLA-DRB1*0701, HLA-DRB1*1401, HLA-DQB1*0303
	Rheumatoid arthritis	HLA-DQA1*03:01, HLA-DRB1*04:01, HLA-DRB1*13:01, HLA-DRB1*04:05, DRB1*10:01
	Selective IgA deficiency	HLA-DQB1:02:01
	Systemic lupus erythematosus	HLA-DRB1*03:01,DRB1*15:01, DQB1*06:02
	Type 1 diabetes	DRB1*03-DQA1*05:01-DQB1*02:01, DRB1*04-DQA1*03:01-DQB1*03:02, DQA1*01:02, DQB1*06:02
	Ulcerative colitis	HLA-DRB1*01:03, HLA-DRB1*15:02
Drug reaction	Abacavir hypersensitivity	HLA-B*57:01
	Carbamazepine hypersensitivity	HLA-B*15:02

Previously known alleles for autoimmune or drug hypersensitivity are in blue (right-hand column); NLP text mining found all these associations and an additional 34 (in red).

linguistically processed with Linguamatics text analytics and indexed using an internally developed HLA gene ontology, alongside the NLP platform's dictionary of relationship verbs (e.g. "causes," "leads to," "results in") and Diseases ontology.

An NLP query was constructed to text mine the literature sources to identify HLA alleles and haplotypes, and their relationships with diseases and drug sensitivity. The query identifies HLA alleles, relationship verbs and diseases from sentences in abstracts or full-text articles, and extracts the triples within five words of each other.

The text mining results are then stored in a database that can be searched through a simple web interface for HLA alleles and diseases. The results can be curated by experts, and the curated annotations can be saved back into the knowledge base for wider use by other researchers.

Success

Linking genes and diseases is a key requirement in drug discovery and the development of predictive biomarkers. It is also important in the development and delivery-end of the drug development pipeline—and is closely integrated in the ability to deliver personalized or precision medicine, and for patient stratification using biomarkers in the clinic. Sanofi was interested in probing the association of HLA alleles and haplotypes with diseases and drug sensitivity as part of an MS biomarker project. While some of these associations were known, Sanofi wanted to systematically and comprehensively annotate the links.

The Linguamatics NLP text mining software processed an extensive collection of literature sources and identified all of the 22 previously published autoimmune disease and drug sensitivities associated with HLA alleles and haplotypes, and uncovered an additional 33 novel, unpublished disease and drug sensitivity associations. This provides Sanofi with a broader and more comprehensive knowledge base from which they can now confidently explore potential new biomarkers.

This application of Linguamatics NLP text mining software revealed previously unknown gene-disease associations for further exploration in biomarker identification.

Sanofi is also making effective use of the power of NLP and text analytics in the Linguamatics platform in other areas of R&D, including target identification and prioritization, drug repurposing, interpretation of genes/proteins identified by 'omics experiments and full-patent text mining for new targets. Beyond R&D, Sanofi is also using text mining along the bench-to-bedside pipeline, in areas as diverse as clinical trial site selection and study design, opportunity scouting, pharmacovigilance, competitive intelligence, and "voice of the customer" and social media analysis.

CONTACT US
+44 (0)1223 651 910 (U.K.) | +1 617 674 3256 (U.S.)
nlp@iqvia.com
linguamatics.com