SoyBase
Integrating Genetics and Genomics
to Advance Soybean Research

Source: http://seedgenenetwork.net/media/SoyBaseLogo.jpg

To Principle Investigator,

For our research question on identifying genes in soybean linked to drought tolerance and how these genes can be leveraged to improve drought resistance, I believe we should consult a database such as **SoyBase**. This database appears to be highly promising in terms of the breadth and depth of data it offers and could significantly aid our research.

About the Database

SoyBase holds a variety of data types, including genomic data such as gene annotations, molecular markers, QTLs (Quantitative Trait Loci), and Genome-Wide Association Studies (GWAS). It also includes phenotypic data related to soybean-associated traits, such as drought resistance, disease resistance, growth, and yield. Additionally, SoyBase houses data on metabolic pathways, gene expression, and gene functions, along with geographic and breeding information for different soybean varieties.

This database can be used to answer a variety of biological research questions, such as:

- What are the different soybean genomes, and how can we analyze them across species?
- How can we map loci and genes in soybeans?
- What are the various gene expression pathways, and how can we compare them across breeds?
- How can we identify new genes and map them across different soybean varieties?
- How can traits be improved across multiple cultivars?

Questions like these are vital to soybean research and can be explored using the data available in SoyBase. For our purposes, the database holds a lot of data on genes associated with drought resistance, making it ideal for our research.

Primary and Secondary Data Characteristics

SoyBase contains both primary and secondary data. Primary data includes direct submissions from researchers, such as genome sequences, annotations, QTLs, and GWAS results, which are directly deposited into the database. It also features secondary data, such as curated phenotypic information, gene functions, and maps derived from primary sources.

Data, Storage, and Tools available

Most of the data in SoyBase is contributed by the original researchers. The database sources information from peer-reviewed research articles, sequencing projects, and individual researchers. Initially developed by the USDA-ARS SoyBase and Legume Clade Database group at Iowa State University, SoyBase is now funded and curated by the USDA Agricultural Research Service (ARS). USDA experts validate, organize, and incorporate the data into the database. SoyBase also integrates data from external sources such as

NCBI and GenBank. Additionally, USDA soybean research projects contribute data, which is continuously curated by experts. Researchers and the community can also submit contributions to the database.

The storage of data depends on the type of data being accessed. There are a wide variety of formats present in line with common bioinformatic standards. For genomic data such as genetic sequences the data is stored in a FATSA format, for other genomic data such as gene models, exons, introns and other features data is stored in the GFF3 format. For other types of data such as genetic markers, QTL, phenotypic data, gene expression data formats are in excel and csv forms make it easily accessible and cross compatible across many analysis tools. The mapping and visualization tools data is stored in PNG or SVG formats.

Hyperlink to a gene entry for soybean: Link

SoyBase additionally offers a variety of tools to analyze the data:

- GO Enrichment Analysis and Northern Uniform Soybean Trials provide detailed phenotypic reports from 1989 to the present.
- Basic and advanced search for genes, sequences, and genomes
- Sequence analysis tools such as BLAST and genome annotation tools are available.
- Visualization tools include the Soybean Genome Browser, QTL tools, and gene expression tools.
- Parentage and Ontology tools detailing growth, development, breeding, and pedigrees

The data can also be accessed computationally via APIs and FTP servers for bulk downloads. Many datasets, including genome sequences, QTL data, and phenotypic data, are downloadable in various formats. This availability of data will be a crucial component for our research question as it will allow us to utilize a variety of tools and allows further manipulation of data.

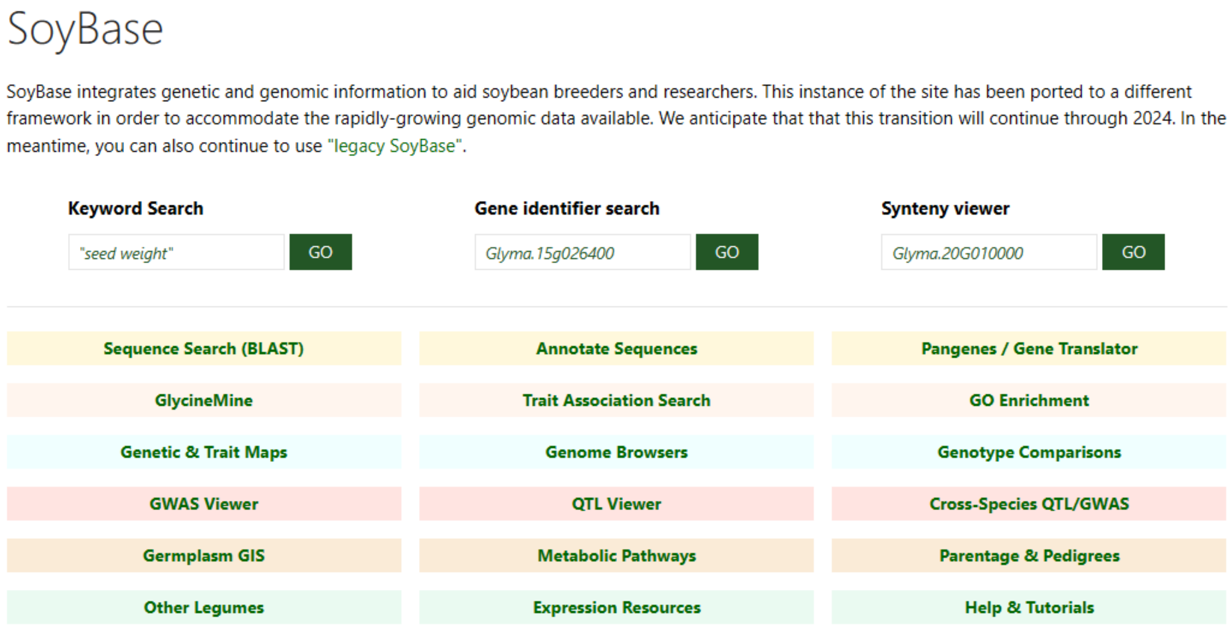
Data Usage and Licensing

Since SoyBase is maintained by the USDA, it is considered a public resource, meaning the data are free to use, as long as they are properly cited. The data have a public domain license, allowing anyone to access, modify, and distribute the data without legal restrictions. This ensures the data are freely available for research and educational purposes. For our purposes this makes sure that we have ease for collaboration and citation for when we conduct our research.

Strengths and Critiques

Positives

SoyBase recently underwent a site migration and redesign, making it more user-friendly compared to its legacy version. It is relatively easy to navigate if users know what they are searching for. The search bar offers both basic and advanced search queries for genes, traits, and sequence data. Tools like the Gene Identifier and Synteny Viewer also aid in navigation. The homepage is well-organized into tabs that simplify finding what users need. An example of the homepage and its organization is given below in Figure 1. The data's quality and comprehensiveness can be trusted, as it comes from peer-reviewed journals and is curated by experts. The database is frequently updated as new data becomes available, making it a reliable and up to date source of information. The comprehensiveness of the data and since most of it is up to date, when looking for different genetic markers for drought resistance we have a good chance of finding the associated genes we are interested in.



SoyBase

SoyBase integrates genetic and genomic information to aid soybean breeders and researchers. This instance of the site has been ported to a different framework in order to accommodate the rapidly-growing genomic data available. We anticipate that that this transition will continue through 2024. In the meantime, you can also continue to use "legacy SoyBase".

Figure 1: Homepage of SoyBase website source: https://www.soybase.org/

Critiques

While SoyBase is well-designed, there are areas that could be improved:

1. Accessibility for First-Time Users: For users unfamiliar with genetic databases, SoyBase can be difficult to navigate. The vast amount of data can be overwhelming, and understanding how it is organized may be challenging for non-experts. A better user interface with clearer explanations for each tab, or comprehensive instructional videos, would help first-time users.
2. Lack of Detailed Documentation: The "About" section is somewhat vague, offering limited information about the database and its tools. Adding more detailed documentation, step-by-step guides, or tutorials would assist users in navigating the database more effectively.
3. General Gene Information: For some genes, only basic information is available (e.g., sequence, length, introns, exons). Integrating SoyBase with other genomic databases, such as TAIR, would provide more detailed annotations and functional insights.

Additionally, a feature that could be added is a 3D protein structure viewer, as structure often correlates with function. A 3D model could help visualize protein domains and ligand interactions. This could be done by integrating with external databases like RCSB PDB. For our purposes for the research question, depending on our researchers we may have to have a walk through of the website as the amount of data we would have to sift through would be immense. By figuring out where the data is stored and is helpful to us, we can create some ease of access and help us concentrate on drought resistance traits.

Conclusion

Overall, **SoyBase** is an extremely comprehensive resource for soybean research, compiling various data types including genomic, phenotypic, and functional data. As a public domain database, it is freely accessible for research without restriction. Its mix of primary and secondary data makes it highly valuable for exploring a wide range of genetic and phenotypic characteristics in soybeans. While SoyBase can be overwhelming at first due to its extensive dataset, improving the user interface and documentation would make it more accessible. For our research question, SoyBase will be an invaluable tool in identifying drought-resistant traits and QTLs, exploring gene expression data, and comparing genes across different soybean varieties. SoyBase offers the characteristics we need from a database, and it will be a very useful resource for our research.