12-4-2020

G2PMineR: A Genome to Phenome Literature Review Approach

John M.A. Wojahn
Boise State University

Stephanie Galla
Boise State University

Sven Buerki
Boise State University

This research is part of the Genome 2 Phenome project.

G2PMineR: A Genome to Phenome Literature Review Approach

Abstract

In this research we aim to develop G2PMineR, a free and open-source R-package literature mining tool that uses C-style string manipulation techniques to facilitate establishing G2P hypotheses for non-model organisms. This approach allows processing tens of thousands of abstracts in a matter of hours. It is highly efficient and unbiased, empowering researchers to draw conclusions linking genes to phenotypes that they otherwise would have missed or not had the time to explore.

Keywords

biological and life sciences, Genome 2 Phenome

Comments

This research is part of the Genome 2 Phenome project.

G2PMineR: A genome to phenome literature review approach

John M. A. Wojahn[1,2], Stephanie J. Galla[1], Sven Buerki[1]
[1]Boise State University, 1910 University Dr, Boise, ID 83725, USA
[2] VIP Genome2Phenome Student

BOISE STATE UNIVERSITY

GEM3 Genes by Environment Modeling · Mechanisms · Mapping

Challenge

Applying G2P research on non-model organisms is challenging due to the lack of genomic resources allowing estimating the role played by genetic processes in adaptation, especially in face of climate change. This means that there is a gap in the conceptual framework linking G2P for these non-model organisms.

To address this, researchers perform ad hoc and post-hoc analyses to understand G2P relationships by curating a list of likely gene candidates, hinging upon other studies already conducted in closely related systems. However, in the post-genomics era, with hundreds of thousands of articles to sift through, this is a cumbersome task, and manual curation of genes of interest may introduce bias into a study..

Aim

We aim to develop G2PMineR, a free and open-source R-package literature mining tool that uses C-style string manipulation techniques to facilitate developing G2P hypotheses for non-model organisms. This approach allows processing tens of thousands of abstracts in a matter of hours. It is highly efficient and unbiased, empowering researchers to draw conclusions linking genes to phenotypes that they otherwise would have missed or not had the time to explore.

General structure of G2PMineR

1. Assessing efficiency of literature search using network analysis
2. Mining abstracts for genes
3. Mining abstracts for taxonomy
4. Mining abstracts for phenotypes
5. Quality control
6. Analyzing genes, taxonomy, and phenotypes data internally
7. Linking G2P within a taxonomical framework

Inputs/Outputs

Inputs: A csv of abstract text and a csv of unique identifiers for each abstract. G2PMineR is agnostic as to the origin of the abstracts (e.g. PubMed, GoogleScholar, manual cut-and-paste) as long as they are one long string (i.e. abstract text) per cell of the csv). *Optionally the user can also add their own lists of genes, phenotypes, and taxa to customize their analyses. The manual details how to do this*

Outputs: The package has numerous intermediate outputs from each module, but the primary outputs are the *bipartite graphs* from module 7 (Gene-Phenotype (see fig 2); Gene-Taxonomy; Phenotype-Taxonomy), the *internal coocurrence networks* from module 6 (one each for Genes, Phenotypes, and Taxonomy), and the *abstracts text clustering network* from module 1. All outputs have associated matrices

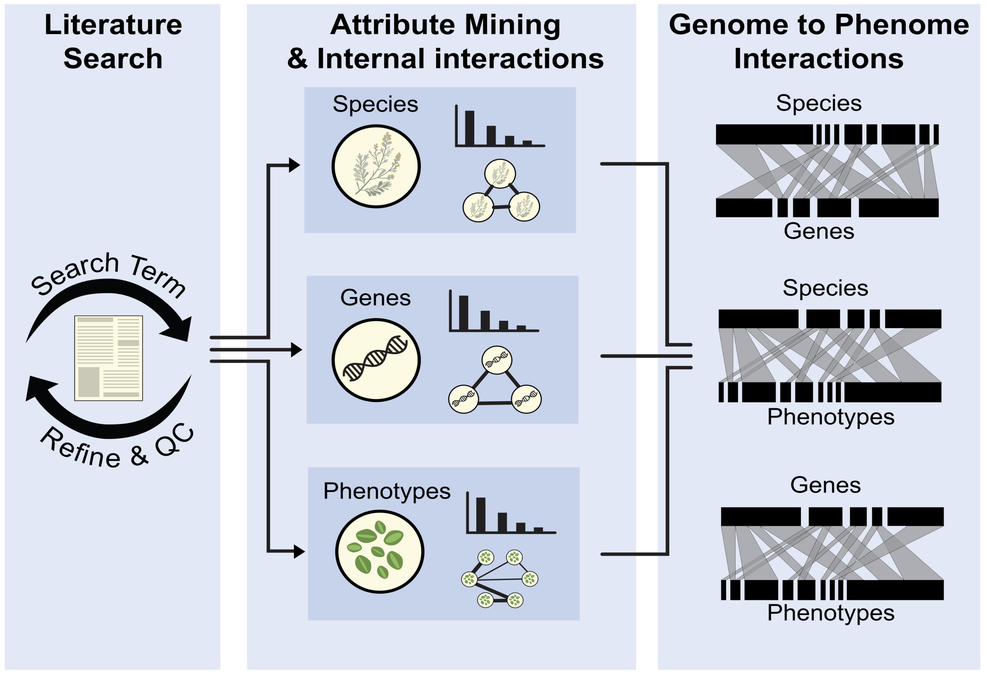

Fig 1: Methodological flowchart

G2PMineR can be used to study plants, animals, and fungi

The package can conduct analyses on data for three kingdoms (Plantae, Animalia, and Fungi), and therefore is useful for a broad range of users and study systems

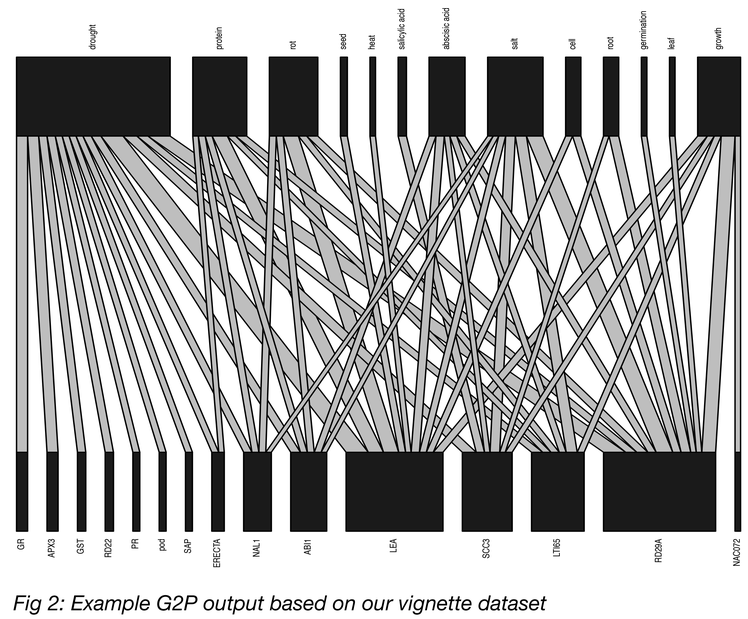

Fig 2: Example G2P output based on our vignette dataset

G2PMineR links G2P in a multifaceted and intersectional way

The gene-phenotype bipartite graph (Fig. 2) allows the user to infer the genes that have the strongest cooccurrence connections with their phenotype(s) of interest (the width of the bars between the upper and lower levels indicates the relative number of abstracts containing them. It also allows the user to see what genes may be expressed together in response to a single (or several) phenotype(s). The other outputs are best interpreted in the context of this output.

G2PMineR is free and open-source software (FOSS)

We have licensed G2PMineR under the GNU Affero General Public License Version 3 (AGPL3) so that it is available for free to anyone. It was also developed to be able to run on a single personal computer so that researchers of all means can have access to it. Nevertheless it can still be run in parallel on a supercomputer if desired. Once publicly released it can be downloaded from GitHub.